Microengineering Aerospace Systems

Henry Helvajian, editor

The Aerospace Press • El Segundo, California

American Institute of Aeronautics and Astronautics, Inc. • Reston, Virginia

The Aerospace Press
2350 E. El Segundo Boulevard
El Segundo, California 90245-4691

American Institute of Aeronautics and Astronautics, Inc.
1801 Alexander Bell Drive
Reston, Virginia 20191-4344

Library of Congress Catalog Card Number 98-074457

Library of Congress Cataloging-in-Publication Data

Microengineering aerospace systems / Henry Helvajian, editor.
p. cm.
Includes bibliographical references and index.
ISBN 1-884989-03-9 (alk. paper)
1. Microelectronics. 2. Space vehicles--Electronic equipment.
I. Helvajian, Henry.
TK7874.M48743 1999
629.47'4--dc21 99-18130
CIP

Preface

This book assembles many of the ideas, fundamental principles, and technology rudiments that can profoundly change the manner by which aerospace systems are designed, engineered, and assembled. The concepts presented do nothing less than advocate the use of microengineering principles to impart "intelligence," "volition," and "motility" to systems on the miniature scale, thereby effecting a change in the paradigm of how aerospace systems should be developed, maintained, and used. These changes in paradigm will initially depend on whether these ideas and the rudiments of the technology can be refined for aerospace applications, but the final result will ultimately depend on the willingness to apply these novel concepts with the conviction that "*very few things [are] indeed really impossible.*" A somewhat incidental but noteworthy fact is that these new possibilities come at the threshold of the new millennium. The next millennium most likely will usher in air and space travel more commensurate with current science fiction lore.

Microengineering is defined here as an interdisciplinary subject made possible by the ability to cofabricate microelectronics with sensors and actuators, using design rules that currently approach the submicron scale but will inevitably reach the nanometer scale in the not too distant future. The subject is guided by the tenet to engineer "intelligent" functionality in the small, either as solitary microsystems or as the concerted intelligent action of innumerable microsystems. As a result of this desire to make things small, the term microengineering can also assume the form of a verb. This interchange in meaning between subject and verb appears without distinction throughout the chapters. The constituent disciplines that buttress microengineering are microelectronics, microelectromechanical systems (MEMS), microsystems, advanced packaging, material processing, micromachining, control systems, information theory, and the basic disciplines (e.g., physics, chemistry, mechanics). In the limit of nanometer or atomic scale fabrication, microengineering borrows heavily from bio and organic chemistry. Concepts such as self replication and assembly become a necessity for practical applications. In the chapters that follow, microengineering refers to systems having dimensions in the micron to centimeter range, with hallmarks being the ability to mass produce the microsystem through batch fabrication techniques and the ability to aggregate microsystems for more complex functionality.

Aerospace systems stand to benefit greatly from the concepts presented in this book, if only because they purport to offer functionality at a reduced size and therefore at a reduced overall mass. It is more likely to be the case that these concepts will bring forth aerospace systems that are more reliable in operation, more "aware" of their inherent health and status, and as a result, less costly to manufacture and operate. The more forward thinking concepts in this book will initiate new aerospace missions that can only be made possible by the mass-producibility aspects in microengineering and the possibility to distribute the sensing, analysis, and action functions of a particular mission. The Aerospace Corporation in El Segundo, California, has been keenly aware of the potential for microengineering to alter the paradigm of space systems development and mission definition. To disseminate this vision, the scientific and engineering staff began in 1993 to publish a series of reports[1] on the subject and also to serve as advocates of microengineering

---

[1]H. Helvajian and E. Y. Robinson, eds, *Micro- and Nanotechnology for Space Systems: An Initial Evaluation*, Monograph 97-01 (The Aerospace Press, El Segundo, CA, 1997); first published as The Aerospace Corp. Report no. ATR-93(8349)-1 (1993). H. Helvajian, ed., *Microengineering Technology for Space Systems*, Monograph 97-02 (The Aerospace Press, El Segundo, CA, 1997); first published as The Aerospace Corp. Report no. ATR-95(8168)-2 (1995).

space systems.[2] The corporation also presented to the aerospace community the concept of the 1-kg-class, mass-producible silicon-nanosatellite.[3]

This book captures the basic microengineering concepts that might be applicable to the advance of aerospace systems. Unlike the earlier Aerospace Corporation reports,[1] this book is written in a tutorial style providing worked examples, key equations, and process sequences. It also provides a snapshot view of the state of the art in a very fast moving subject. This presentation approach, to some extent, was chosen because of the diverse audience expected (e.g. aerospace, microengineering communities) and also to help the reader better assess applicability. To date, there are several books on the subjects of microengineering,[4] future of aerospace systems,[5] and specific applications,[6] but to the best knowledge of this editor, no similar book has yet been published for the aerospace community as a whole. The book comprises 17 chapters and is loosely organized into sections that can be labeled as introduction and state-of-the art overview, materials/mechanics/processing and packaging, microsystem devices, distributed system architectures, and satellite subsystems. As with any digest that claims to encapsulate a particular subject, pertinent matter is always left out. This assemblage fares no better: missing subjects include nanoelectronics, nanoelectromechanical systems (NEMS), bio-MEMS, microfluidics, *lithographie galvanoformung abformung* (LIGA), and microfabrication in polymeric materials. Also, not actually discussed but easily identifiable, are specific applications to hypersonic and transatmospheric vehicles. A pregnant application to all aerospace vehicles, but not presented in detail, is the use of microengineered devices integrated into the vehicle skin that help to identify and monitor microfractures and material corrosion.

---

[2]S. Janson, H. Helvajian, and E. Y. Robinson, "The Concept of 'Nanosatellite' for Revolutionary Low-Cost Space Systems," *Proceedings of the 44th Congress of the International Astronautics Federation* (Graz, Austria, 16–22 October 1993; Proceeedings of the International Conference on Integrated Micro/Nanotechnology for Space Applications, Houston, Texas, November 1995 (The Aerospace Press, El Segundo, CA); E. Y. Robinson, H. Helvajian and S. W. Janson, "Small and Smaller: The world of Micro-and Nanotechnology," *Aerospace America* (September 1996); E. Y. Robinson, H. Helvajian, and S. W. Janson, "Big Benefits from Tiny Technologies," *Aerospace America* (October 1996).

[3]S. W. Janson, "Chemical and Electric Micropropulsion Concepts for Nanosatellites," Paper 94-2998, *Proceedings 30th AIAA Joint Propulsion Conference* (Indianapolis, IN, 27–29 June 1994); S. W. Janson and H. Helvajian, "Batch-Fabricated Microthrusters for Kilogram-class Spacecraft," *Proceedings of GOMAC '98, Micro-Systems and their Applications* (Arlington, VA, 16–19 March 1998).

[4]R. S. Muller, R. T. Howe, S. D. Senturia, R. L. Smith and R. M. White, eds., *Microsensors* (IEEE Press, NY, 1991); N. Taniguchi, ed., *Nanotechnology: Integrated Processing Systems for Ultra-Precision and Ultra-Fine Products* (Oxford University Press, NY, 1996); I. Fujimasa, *Micromachines: A New Era in Mechanical Engineering* (Oxford University Press, NY, 1996); M. Madou, *Fundamentals of Microfabrication*, (CRC Press, Boca, Raton, FL,1997); T. A. Kovacs, *Micromachined Transducers Sourcebook*, (WCB McGraw Hill, NY, 1999); W. S. Trimmer, ed., *Micromechanics and MEMS : Classic and Seminal papers to 1990* (IEEE Press, NY, 1997).

[5]A. K. Noor and S. L. Venneri, eds., *Future Aeronautical and Space Systems, Progress in Astronautics and Aeronautics*, Vol. 172 (American Institute of Aeronautics and Astonautics, Reston, VA, 1997).

[6]Hector De Los Santos, *Introduction to Microelectromechanical (MEM) Microwave systems* (Boston: Artech House, 1999).

It has been more than 40 years[7] since Richard P. Feynman gave his memorable lecture titled "There's Plenty of Room at the Bottom." In the intervening years, the science and technology of microelectronics, MEMS, microsystems, nanotechnology, and microengineered systems in general has progressed beyond most predictions. For the "newcomer" MEMS, market studies now show that for terrestrial applications, global sales will exceed $10 billion (U.S.) by the year 2003.[8] It has also been more than 40 years since Sputnik I was launched,[9] heralding space as a strategic commodity. In the intervening years, the aerospace industry has made enormous strides in the areas of aircraft safety, national security, space investigation, and civilian services. The industry now appears poised at the threshold of a revolution that is partly driven by a global demand for passenger air traffic to more municipalities, "instant" high bandwidth wireless communication services, overnight "Pony Express" services worldwide, and real-time remote sensing and global navigation capabilities. Microengineering is one cornerstone discipline that can make these services prevalent worldwide.

Henry Helvajian
The Aerospace Corporation
31 March 1999

---

[7] *Annual Meeting of the American Physical Society* (California Institute of Technology, 26 December 1959).

[8] *MicroElectroMechanical Systems (MEMS): An SPC Market Study* (System Planning Corporation Press [SPC], January 1999).

[9] Sputnik I launched 4 October 1957.

Acknowledgments

I wish to thank the authors and members of the MEMS, microengineering, and aerospace communities for giving freely of their ideas so that a change in vision might take hold in the aerospace industry; Dr. David J. Evans of The Aerospace Institute for seeing the vision and giving me the opportunity to assemble this book; Professor Sheila Colwell for reminding me that rarely does a change in view globally succeed without a representative symbol and suggesting "Alice" as the symbol of this wonderland of the novel and miniature; Anne-Marie Colwell, my wife, for her patience and support through this ordeal; and finally Dr. Donna J. Born of The Aerospace Press, whose high level of professionalism and astute demand for detail has made this book a reality.

Contents

1

Introduction to MEMS

M. Mehregany* and S. Roy*

1.1 Overview

Interest in the development of microelectromechanical systems (MEMS) has mushroomed during the past decade. In the most general sense, MEMS attempts to exploit and extend the fabrication techniques developed for the integrated circuit (IC) industry to add mechanical elements, such as beams, gears, diaphragms, and springs, to the electrical circuits to make integrated microsystems for perception and control of the physical world. MEMS devices are already being used in a number of commercial applications, including projection displays and the measurement of pressure and acceleration. New applications are emerging as the existing technology is applied to the miniaturization and integration of conventional devices.

This chapter starts with an overview of MEMS technology, followed by fabrication technologies, selected applications, commercial aspects, trends in MEMS technology, and journals and conferences (Table 1.1). The review covers both the potential and the limitations of MEMS technology.

Table 1.1. Evolution of MEMS

Silicon anisotropic etching	pre-1950
Piezoresistive effect in silicon	1953
Semiconductor strain gauges	1957
Silicon pressure sensors	post-1960
Solid state transducers	post-1970
Microactuators	post-1980
Mechanisms and motors	1987–89
Microelectromechanical systems Microsystems Micromachines	post-1988

*Microfabrication Laboratory, Electrical Engineering and Applied Physics, Case Western Reserve University, Cleveland, Ohio.

1.1.1 Historical Background

The transistor[1] was invented at Bell Telephone Laboratories on 23 December 1947. This invention, which led to a Nobel Prize awarded in 1948 to Schockley, Bardeen, and Brattain, initiated a fast-paced microelectronic technology. The transition from the original germanium (Ge) transistors with grown and alloyed junctions to silicon (Si) planar double-diffused devices took about 10 years. The IC concept was conceived by several groups, and included RCA's Monolithic Circuit Technique for hybrid circuits (1955). The first IC, shown in Fig. 1.1 was built by Jack Kilby of Texas Instruments in 1958, using Ge devices (a patent was issued to Kilby in 1959). A few months later, Robert Noyce of Fairchild Semiconductor announced the development of a planar Si IC. The complexity of ICs has doubled every 2 to 3 years since 1970. The minimum dimension of manufactured devices and ICs has decreased from 20 μm to the submicron levels of today. Currently, ultra-large-scale-integration (ULSI) enables the fabrication of more than 10 million transistors and capacitors on a typical chip. ULSI-based microprocessors and microcomputers have revolutionized communication, entertainment, health care, manufacturing, management, and many other aspects of our lives. Low-cost, high-performance electronic systems are now available to the public and have improved the quality of life in many ways. However, control and measurement systems, as well as the actual automation facilities used in IC fabrication, would be "deaf," "dumb," and "blind" without sensors to provide input from the surrounding environment. Similarly, without actuators, control systems would be powerless to carry out the desired functions. While IC technology (and more specifically, microfabrication) has provided high-speed, miniaturized, low-cost signal conditioning and signal-processing capabilities, conventional sensors and actuators (also referred to as transducers) are far behind in performance, size, and cost.

The success of Si as an electronic material in ULSI technology was due partly to its wide availability from silicon dioxide (SiO_2) (sand), resulting in potentially lower material costs relative to other semiconductors. Consequently, a significant effort was put into developing Si processing and characterization tools. Today, some of these tools are being utilized extensively to advance transducer technology. In this area, attention was first focused on microsensor (i.e., microfabricated sensor) development. Si microsensors initially addressed the measurement of physical variables, expanded to the measurement of chemical variables, and then progressed to biomedical applications.The first, and to date, the most successful microsensor, is the Si pressure sensor. The history of Si pressure sensors is representative of the evolution of microsensors.

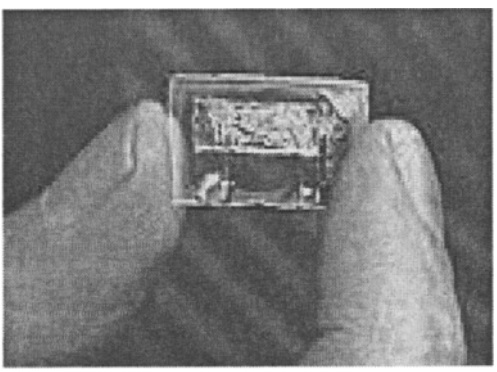


Fig. 1.1. First integrated circuit consisting of one transistor, three resistors, and one capacitor. The IC was implemented on a sliver of germanium that was glued on a glass slide (Texas Instruments, Inc.).

1.1.1.1 Silicon Pressure Sensor Technology

In 1953, Dr. Charles S. Smith of Case Institute of Technology (now part of Case Western Reserve University [CWRU]), during a sabbatical leave at Bell Telephone Laboratories, studied the piezoresistivity of semiconductors and published the first paper on the piezoresistive effect in Ge and Si in 1954.[2] The piezoresistive effect is the change in the resistivity of certain materials due to applied mechanical strain. The measured piezoresistive coefficients indicated that the gauge factor of Ge and Si strain gauges could potentially be 10 to 20 times larger, and therefore much more sensitive than those based on metal films. As a result, discrete Si strain gauges were developed commercially in 1958 by Kulite Semiconductor Products, Honeywell, and Microsystems. Such Si strain gauges were integrated on a thin Si substrate as diffused resistors in 1961 by Kulite. The thin Si substrate was then mounted on a base to act as a diaphragm. In 1966, Honeywell developed a method to fabricate thin Si diaphragms by mechanically milling a cavity into an Si substrate. Isotropic Si etching was used to produce micromachined Si diaphragms in 1970, and anisotropic etching was introduced for this purpose in 1976. Both techniques were introduced by Kulite. Glass frits were introduced to bond the Si wafer (in which the pressure-sensitive diaphragms were fabricated) to a base wafer in the 1970s, allowing wafer-scale fabrication of pressure sensors. The first high-volume pressure sensor was marketed by National Semiconductor in 1974. This sensor included a temperature controller (in a hybrid package) for constant temperature operation. At this point, piezoresistive pressure-sensor technology had become a low-cost, batch-fabricated manufacturing technology. Further improvements of this technology have included the utilization of ion implantation for improved control of the piezoresistor fabrication, etch stops for improved control of the diaphragm thickness after the etch, deep Si-reactive ion etching for increased packing density, anodic bonding (electrostatic bonding), and more recently, Si-to-Si fusion bonding for improved packaging of the pressure sensors. Currently, Si pressure sensors are a billion-dollar industry and growing.[3]

The first monolithic integrated pressure sensor with digital (i.e., frequency) output was designed and tested in 1971 at CWRU,[4] as part of a program addressing biomedical applications. Miniature Si diaphragms with a resistance bridge at the center of the diaphragm and sealed to the base wafer with a gold (Au)-tin (Sn) alloy were developed for implant and indwelling applications. During field evaluation, it was found that the packaging of the sensors determined their performance, and that piezoresistive sensors were very sensitive to interference, such as sideways forces, making them inaccurate for many biomedical applications. To achieve better sensitivity and stability, capacitive pressure sensors were first developed and demonstrated at Stanford University in 1977 and shortly afterward at CWRU. The first integrated monolithic capacitive pressure sensor was reported in 1980.[5] In general, capacitive pressure sensors exhibit superior performance compared to traditional piezoresistive pressure sensors. However, the relatively complex design and implementation of signal-processing circuitry required for electronic readout initially limited the widespread availability of capacitive pressure sensors. During the last 15 years, various processing and transduction techniques have been used to develop new or improved Si pressure sensor designs. While such developments are ongoing, advanced piezoresistive Si pressure sensors still account for almost all of the Si pressure sensor market. During the same period, Si microsensor technology has matured substantially, and a variety of sensors have been developed for measuring position, velocity, acceleration, pressure, force, torque, flow, magnetic field, temperature, gas composition, humidity, pH, solution/body fluid ionic concentration, and biological gas/liquid/molecular concentrations. Some of these sensors have been commercialized.

1.1.1.2 Micromachining

Development of Si microsensors often required the fabrication of micromechanical parts (e.g., a diaphragm in the case of the pressure sensor and a suspension beam for many accelerometers). These micromechanical parts were fabricated by selectively etching areas of the Si substrate away to leave behind the desired geometries. Hence, the term "micromachining" came into use around 1982 to designate the mechanical purpose of the fabrication processes that were used to form these micromechanical parts. Isotropic etching of Si was developed in the early 1960s for transistor fabrication. Anisotropic etching of Si was reported in 1967 by Finne and Klein[6] and in 1973 by Price.[7] Various etch-stop techniques were subsequently developed to provide further process flexibility. Together, these techniques have been used for fashioning micromechanical parts from Si materials, and they also form the basis of the "bulk" micromachining processing techniques. Bulk micromachining designates the point that the bulk of the Si substrate is etched away to leave behind the desired micromechanical elements.

While bulk micromachining has been a powerful technique for the fabrication of micromechanical elements, ever-increasing needs for flexibility in device design and performance improvement have motivated the development of new concepts and techniques for micromachining. For example, the application of the sacrificial layer technique (first demonstrated by Nathanson and Wickstrom in 1965[8]) to micromachining in 1985 gave birth to the concept of "surface" micromachining.[9] Surface micromachining designates the point that the Si substrate is primarily used as a mechanical support upon which the micromechanical elements are fabricated. More recently, the introduction of Si fusion bonding and deep reactive ion etching, as well as high-aspect-ratio lithography and plating processes, have expanded the capabilities of micromachining technology.

Prior to 1987, Si micromachining had been used to fabricate a variety of micromechanical structures, such as thin Si diaphragms, beams, and other suspended structures, in single-crystal Si or in films deposited on an Si substrate. These micromechanical structures were generally limited in motion to small deformations and were physically attached to the substrate. Such elastic components could be used as flexible joints, but their overall usefulness in the design of "mechanisms" was limited. "Mechanism" as used here is a means for transmitting, controlling, or constraining relative movement and refers to a collection of rigid bodies connected by joints. During 1987 to 1988, a turning point in the field was reached when, for the first time, techniques for integrated fabrication of mechanisms on Si were demonstrated.[10,11] It was then possible to fabricate mechanical parts that could execute unrestrained motion in at least one degree of freedom (e.g., gears, gear trains, linkages). Shortly thereafter, this technology enabled the development of electrostatic micromotors[12,13] and motivated the development of other types of microactuators, such as valves, pumps, switches, tweezers, and lateral resonant devices.

1.1.1.3 MEMS

Recent progress in microactuators is transforming the conventional field of solid-state transducers into what has become known as MEMS. The term "MEMS" was coined around 1987, when a series of three workshops on microdynamics and MEMS was held in July 1987 in Salt Lake City, Utah; in November 1987 in Hyannis, Massachusetts; and in January 1988 in Princeton, New Jersey. These workshops ushered in a new era of microdevices. Equivalent terms for MEMS include "microsystems," which is preferred in Europe, and "micromachines," which is favored in Japan. MEMS is application driven and technology limited, and has emerged as an interdisciplinary field that involves many areas of science and engineering.

Miniaturization of mechanical systems promises unique opportunities for new directions in the progress of science and technology. Micromechanical devices and systems are inherently smaller,

lighter, faster, and usually more precise than their macroscopic counterparts. However, the development of micromechanical systems requires appropriate fabrication technologies that enable the following features in general systems:

- Definition of small geometries
- Precise dimensional control
- Design flexibility
- Interfacing with control electronics
- Repeatability, reliability, and high yield
- Low cost per device

1.2 Fabrication Technologies

The three characteristic features of MEMS fabrication technologies are miniaturization, multiplicity, and microelectronics. Miniaturization is clearly an important part of MEMS, since materials and components that are relatively small and light enable compact and quick-response devices. Multiplicity refers to the batch fabrication inherent in semiconductor processing. Consequently, it is feasible to fabricate thousands or millions of components as easily and concurrently as one component, thereby ensuring low unit component cost. Furthermore, multiplicity provides flexibility in solving mechanical problems by enabling the possibility of a distributed approach through use of (coupled) arrays of micromechanical devices. Finally, microelectronics provides the intelligence to MEMS and allows the monolithic merger of sensors, actuators, and logic to build closed-loop feedback components and systems.

Clearly, the successful miniaturization and multiplicity of traditional electronics systems would not have been possible without IC fabrication technology. It is therefore natural that the IC fabrication technology, or microfabrication, has so far been the primary enabling technology for the development of MEMS. Microfabrication provides a powerful tool for batch processing and miniaturization of mechanical systems into a dimensional domain not accessible by conventional (machining) techniques. Furthermore, microfabrication provides an opportunity for integration of mechanical systems with electronics to develop high-performance, closed-loop-controlled MEMS. Integrated fabrication techniques, which are made possible by IC fabrication technology, eliminate the need for discrete component assembly, which is not practical for the fabrication of MEMS. Hence, dimensional control, including component size and intercomponent clearance, is limited only by the processing technology.

Even though miniaturization of mechanical systems is directly compared to that of electronics, two important points should be noted. First, not all mechanical systems will benefit from miniaturization. More likely, microfabricated sensors and actuators that enable performance improvement will be integrated into conventional macroscopic mechanical systems. Miniaturization and the application of microtransducers for monitoring and control is justified when the performance-to-cost ratio is improved. Second, current IC-based fabrication technologies are inherently planar, not allowing full flexibility for three-dimensional (3D) design. A mature technology for micromechanical systems will require complementary fabrication techniques that provide3D design capabilities.

1.2.1 IC Fabrication

Any discussion of MEMS first requires a basic understanding of IC fabrication technology. The major steps in this technology include film growth, doping, lithography, etching, dicing, and packaging (see Fig. 1.2). Devices are usually fabricated on Si substrates, which are grown in boules, sliced into wafers, and polished. Thin films are grown on these substrates and are used to build active components, passive components, and interconnections between circuits. These films

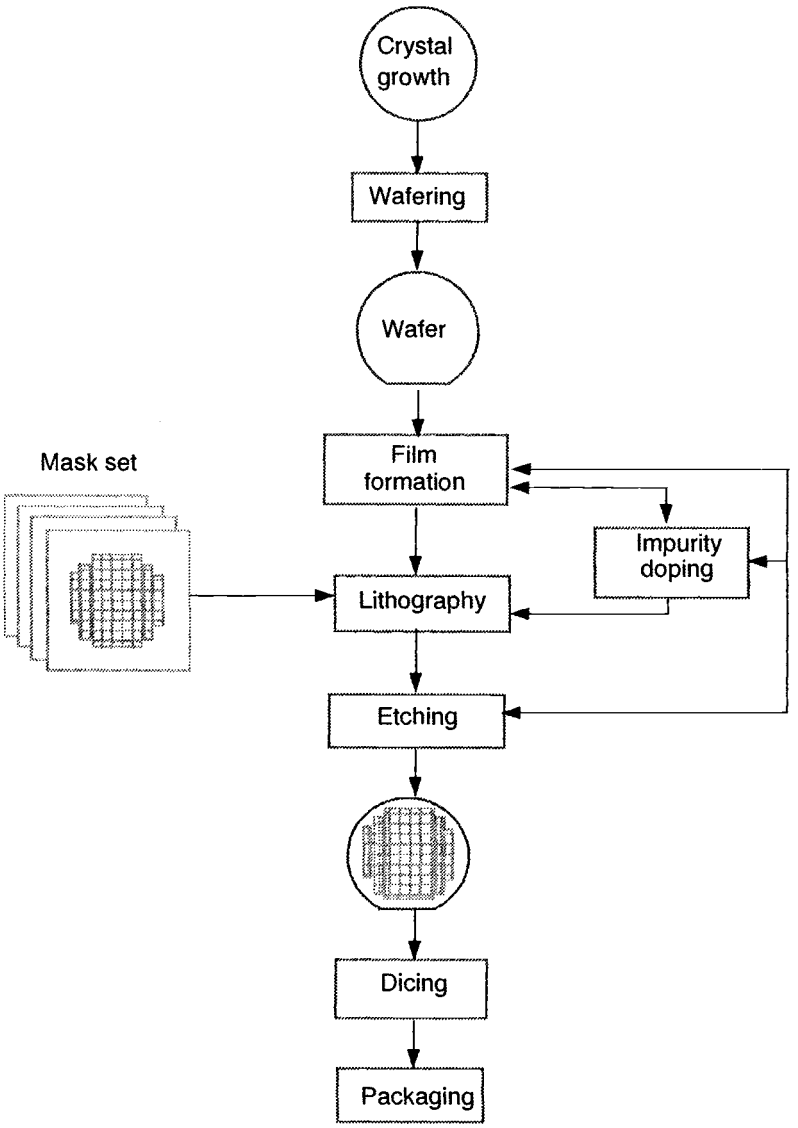


Fig. 1.2. Major processing steps in integrated circuit fabrication.

include: (1) epitaxial Si, (2) SiO_2, (3) silicon nitride (Si_3N_4), (4) polycrystalline Si (polysilicon), and (5) metal films. To modify electrical or mechanical properties, films are doped with impurities by thermal diffusion or ion implantation. Lithography is used to transfer a pattern from a mask to a film via a photosensitive chemical called a photoresist. The film is then selectively etched away, leaving the desired pattern in the film. This cycle is repeated until fabrication is complete. The wafers are then probed for yield, diced into chips, and packaged as final devices. Because there is a market for high-quality, inexpensive Si wafers (namely, microelectronics), most MEMS fabrication facilities focus on thin-film growth, doping, lithography, and etching processes.

1.2.1.1 Film Growth

The growth of SiO_2 by the thermal oxidation of Si is the fundamental film growth process. In IC fabrication, oxidation is used for passivating the Si surface, masking diffusion and ion implantation layers, growing dielectric films, and providing an interface between Si and other materials. In MEMS, SiO_2 films are also used as etch masks and sacrificial layers, which will be discussed later. Although Si exposed to air at room temperature will grow a native oxide (about 20 Å thick), thicker oxide films (0.5–1.5 μm) can be grown at elevated temperatures in a mixture of hydrogen (H_2) and oxygen (O_2) gases. The rate of oxide growth is dependent on the growth temperature, the oxygen partial pressure, and the crystal orientation of the Si substrate. However, for a fixed temperature, oxide thickness increases with time in parabolic fashion.

To deposit SiO_2 films on substrates other than Si, a process known as chemical vapor deposition (CVD) is used. In this process, the chemical components of the film are supplied to the reactor as a mixture of gases. The substrate is heated to a temperature that induces a pyrochemical reaction and film formation. Such depositions are performed at atmospheric pressure (AP) or low pressure (LP). In most instances, the process is ideal for batch coating. Growth rates are much higher in APCVD systems, while films are deposited with excellent thickness uniformity in LPCVD systems. CVD is also used to deposit thick (>1.5 μm) oxide films or when the substrate cannot be simply oxidized thermally. In addition to SiO_2, metals, polysilicon, Si_3N_4, and many other films can be deposited by CVD.

Epitaxial growth is a special class of CVD. Epitaxy is defined as growth of a single crystal film upon a single crystal substrate. If the composition of the film is the same as that of the substrate, the process is called homoepitaxy. However, if the film composition differs from that of the substrate, the process is called heteroepitaxy. Many compound semiconductors, such as gallium arsenide (GaAs) and silicon carbide (SiC), can be grown heteroepitaxially on Si, while doped Si layers are homoepitaxial.

Many films are not thermally stable at temperatures commonly used in conventional CVD. To reduce deposition temperatures so that existing films will not be adversely affected, CVD pyrochemical reactions are combined with a radio-frequency (RF) plasma in a process known as plasma-enhanced CVD (PECVD). PECVD results in films with good step coverage and low pinhole density. However, PECVD films generally suffer from significant hydrogen incorporation and lower mass density, degrading the electrical and mechanical properties of a material.

Metal films can be deposited by vacuum evaporation, sputtering, CVD, and plating, and are most commonly used for interconnections, ohmic contacts, and rectifying metal-semiconductor contacts. Vacuum evaporation is used to deposit single-element conductors, resistors, and dielectrics. Alloys can also be deposited by this method, but the process is complicated by the widely varying evaporation rates of different metals. Resistive and electron beam heating are the two most common heat sources. Compound materials and refractive metals can be deposited by sputtering a cathode target with positive ions from an inert gas discharge. Introduction of noninert gases into the ambient during sputtering is called reactive sputtering and is used to deposit compound films such as titanium nickel (TiNi).

1.2.1.2 Doping

In many instances, it is desirable to modulate the properties of a device layer by introducing a low and controllable level of impurity atoms into the layer. This process is called doping and is accomplished by either thermal diffusion or ion implantation. Thermal diffusion is performed by heating the wafers in a high-temperature furnace and passing a dopant-containing carrier gas across the wafer. The diffusion process occurs in two stages: predeposition and drive-in. During

predeposition, dopant atoms are transported from the source onto the wafer surface and are diffused into the near-surface region. The sources can be gaseous (e.g., diborane [B_2H_6]) or solid (e.g., boron nitride [BN]), depending on the dopant. During drive-in, the temperature is increased, and the dopant diffuses into the wafer to the desired depth and concentration. Ion implantation introduces dopants below the wafer surface by bombardment with an energetic beam of dopant ions. Because the energy loss of these ions in Si is well known, precise control of the dose and depth of dopants is possible. The crystal lattice is damaged during this process, but the damage can often be reduced by subjecting the wafer to a high-temperature, postimplant anneal.

1.2.1.3 Lithography

Lithography is the technique by which the pattern on a mask is transferred to a film or substrate surface via a radiation-sensitive material. The radiation may be optical, x-ray, electron beam, or ion beam. For optical exposure, the radiation-sensitive material is more commonly called "photoresist", and the process is called "photolithography." Photolithography consists of two key steps: (1) pattern generation and (2) pattern transfer. Pattern generation begins with mask design and layout using computer-aided design (CAD) software, from which a mask set is manufactured. A typical mask consists of a glass plate coated with a patterned chromium (Cr) film. Pattern transfer involves: (1) dehydration and priming of the surface, (2) photoresist coating of the wafer, (3) "soft bake" of the photoresist, (4) exposure of the photoresist through the mask, (5) chemical development of the photoresist, (6) wafer inspection, and (7) postdevelopment bake or "hard bake." After hard bake, the mask pattern has been completely transferred to the photoresist.

1.2.1.4 Etching

Following hard bake, the desired pattern is transferred from the photoresist to the underlying film or wafer by a process known as etching. Etching is defined as the selective removal of unwanted regions of a film or substrate and is used to delineate patterns, remove surface damage, clean the surface, and fabricate 3D structures. Semiconductors, metals, and insulators can all be etched with the appropriate etchant. The two main categories of etching are wet-chemical and dry-etching. As the name implies, wet-chemical etching involves the use of liquid reactants to etch the desired material. However, tighter governmental regulations on safety and waste, coupled with the trend toward smaller device features, have led to an increasing emphasis on dry etching. There are various types of dry-etch processes, ranging from physical sputtering and ion-beam milling to chemical-plasma etching. Reactive ion etching, the most common dry-etch technique, uses a plasma of reactant gases to etch the wafer, and thus is performed at low pressure in a vacuum chamber. Well-characterized wet-chemical and dry-etch recipes for most semiconductor processing materials can be found in the literature and will not be detailed here.

In order to fabricate structures, etching is used in conjunction with photolithographically patterned etch masks. The effectiveness of an etchant depends on its selectivity, that is, its ability to effectively etch the exposed layer without significantly etching the masking layer. Since most etch masks are not completely impervious to etchants, mask thicknesses depend on the selectivity of the etchant and the total etch time. Suitable etch-mask materials for many dry- and wet-chemical etchants include SiO_2, Si_3N_4, and hard-baked photoresist.

1.2.2 Bulk Micromachining and Wafer Bonding

Bulk micromachining was developed between 1970 and 1980, as an extension of IC technology, for fabrication of 3D structures.[14] Bulk micromachining of Si uses wet- and dry-etching techniques in conjunction with etch masks and etch stops to sculpt micromechanical devices from the

Si substrate. There are two key capabilities that make bulk micromachining of Si a viable technology. First, anisotropic etchants of Si, such as ethylene-diamine and pyrocatechol (EDP), potassium hydroxide (KOH), and hydrazine (N_2H_4), are available that preferentially etch single crystal Si along given crystal planes. Second, etch masks and etch-stop techniques are available that can be used in conjunction with Si anisotropic etchants to selectively prevent regions of Si from being etched. As a result, it is possible to fabricate microstructures in an Si substrate by appropriately combining etch masks and etch-stop patterns with anisotropic etchants.

Good etch masks for typical anisotropic etchants are provided by SiO_2, Si_3N_4, and some metallic thin films such as Cr and Au. These etch masks protect areas of Si from etching and define the initial geometry of the region to be etched. Alternatively, etch stops can be used to define the microstructure thickness. Two techniques for etch stopping have been widely used in conjunction with anisotropic etching in Si. One technique that uses heavily boron (B)-doped Si, called "p+ etch stop," is effective in practically stopping the etch. Another technique, called "pn junction," stops the etch when one side of a reverse-biased junction diode is etched away.

Anisotropic wet etchants of Si, such as KOH, are able to etch Si <100> and <110> crystal planes significantly faster than the <111> crystal planes. In a <100> Si substrate, etching proceeds along the (100) planes but is practically stopped along the <111> planes. Since the <111> crystal planes make a 54.7-deg angle with the <100> planes, slanted walls result, as shown in Fig. 1.3. Because of the slanted <111> planes, the size of the etch-mask opening determines the final size of the etched hole or cavity. If the etch mask openings are rectangular and the sides are aligned with the [110] direction, practically no undercutting of the etch-mask feature takes place. However, significant undercutting below the mask may occur in convex corners (corners with angles greater than 180 deg), where the etch masks are misaligned with the [110] direction, or where there are curved edges in the etch-mask openings. Under these circumstances, the undercutting continues until it is limited by the <111> planes. Undercutting can be used to fabricate suspended microstructures. Figure 1.4 shows a bulk micromachined Si cantilever fabricated by undercutting the beam's convex corners–defined by an etch stop–from the front side of the wafer.

A drawback of wet anisotropic etching is that the microstructure geometry is defined by the internal crystalline structure of the substrate. Consequently, fabricating multiple, interconnected micromechanical structures of free-form geometry is often difficult or impossible. Two additional processing technologies have extended the range of traditional bulk micromachining technology: deep anisotropic dry etching and wafer bonding. Deep anisotropic dry etching of Si can be achieved using reactive gas plasmas, which will etch exposed Si vertically. Recent improvements in this technology allow the patterning and etching of high-aspect-ratio (e.g., 20:1), anisotropic, randomly shaped features into a single crystal Si wafer, with only photoresist as an etch mask.[15] As shown in Fig. 1.5, etch depths of a few hundred microns into an Si wafer are possible while maintaining smooth, vertical sidewall profiles. The other technology, wafer bonding, permits an Si substrate to be attached to another substrate, typically Si or glass. Electrostatic (or anodic) bonding of Si to glass substrates is performed under application of pressure and high voltage (400–1000 V), while Si fusion bonding (SFB) is the bonding of two Si wafers at high temperatures (near 1000°C), in an O_2 or N_2 ambient. By combining anisotropic etching and wafer bonding techniques, bulk micromachining technology can be used to construct 3D complex microstructures such as microvalves and micropumps. Figure 1.6 presents a microvalve that is fabricated by anisotropic etching and bonding of four Si wafers. In addition to dry etching and wafer bonding, the capabilities of bulk micromachining are further enhanced by laser processing techniques (Chapter 5) applied to microstructures up to 1-mm thick with 20:1 aspect ratios.

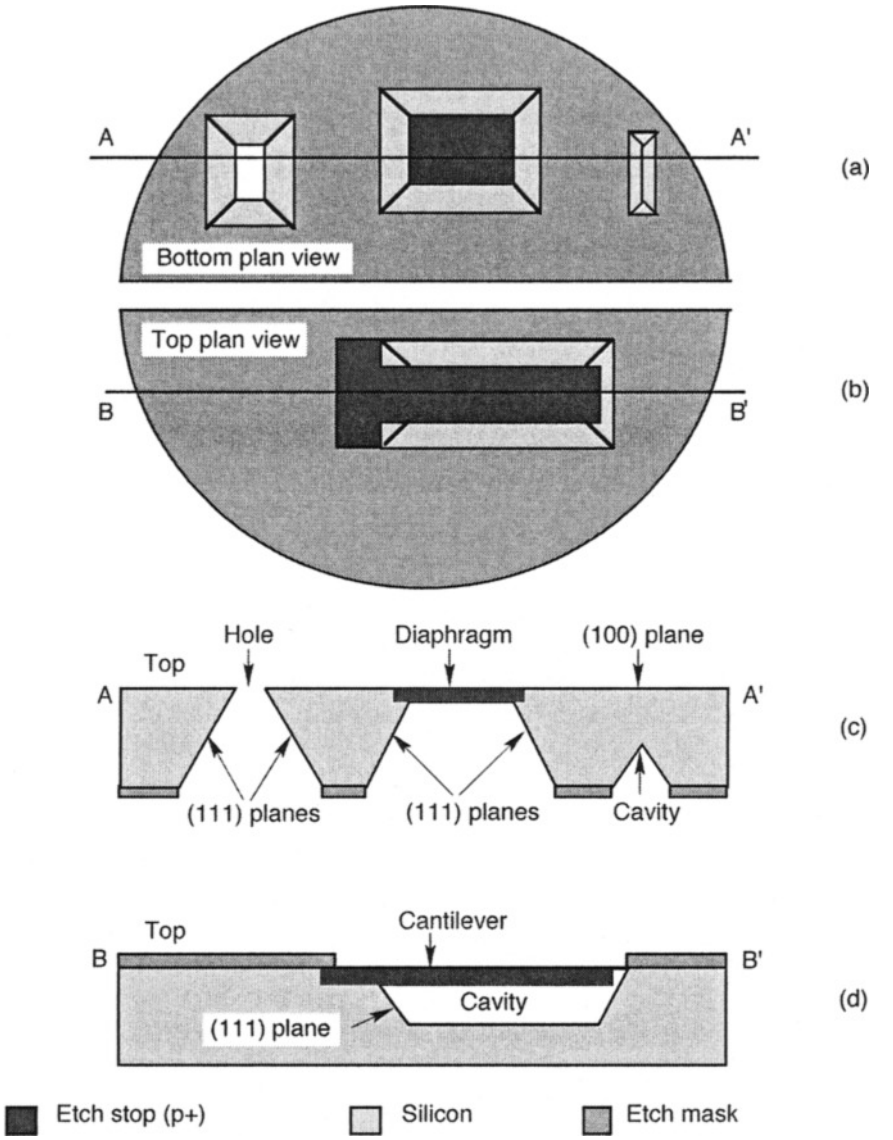


Fig. 1.3. Bulk micromachined features realized by anisotropic etching of silicon. (a) Bottom plan view of etched wafer with cavities, diaphragms, and holes; (b) top plan view of an anisotropically etched wafer showing the fabrication of a cantilever beam using etch stop layer; (c) cross section, AA', showing the hole, diaphragm, and cavity of (a); and (d) cross section, BB', showing the cantilever beam of (b).

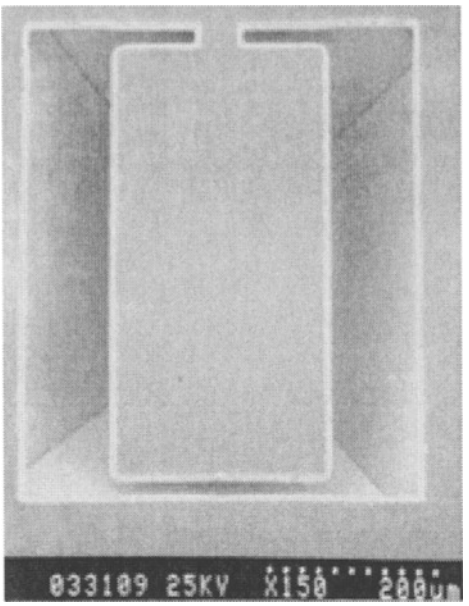


Fig. 1.4.Bulk micromachined cantilever fabricated by p+ etch stop and anisotropic etching.

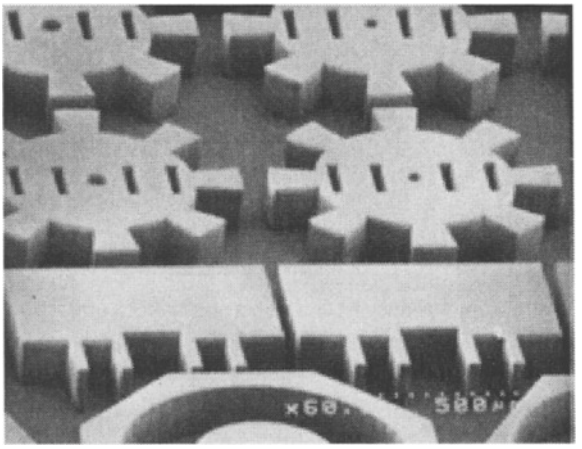


Fig. 1.5. Complex shapes patterned using deep reactive ion etching (DRIE).

1.2.3 Surface Micromachining

Surface micromachining relies on encasing specific structural parts of a device in layers of a sacrificial material during the fabrication process. The sacrificial material is then dissolved in a chemical etchant that does not attack the structural parts. In surface micromachining, the substrate wafer is used primarily as a mechanical support on which multiple, alternating layers of structural and sacrificial material are deposited and patterned to realize micromechanical structures. Surface micromachining enables the fabrication of complex, multicomponent, integrated micromechanical structures that would be impossible with traditional bulk micromachining.

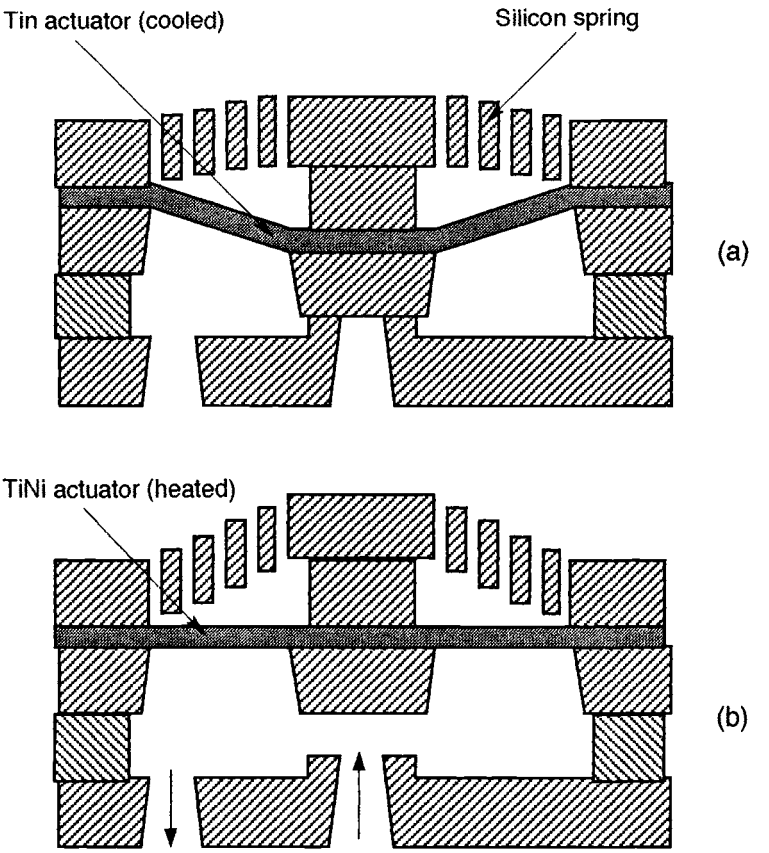


Fig. 1.6. Schematic cross section of a microvalve fabricated by bulk micromachining and wafer bonding. The TiNi shape memory film is thermally actuated to open and close the microvalve (H. Kahn, Case Western Reserve University).

A typical surface micromachining process, shown in Fig. 1.7, begins with the deposition of a sacrificial layer, which is then patterned to create openings to the underlying substrate. Next, the structural layer is deposited and patterned into the desired geometry. Finally, the structural components are released by removal of the underlying and surrounding sacrificial material. The structural components are attached to the underlying substrate at the anchor regions.

Surface micromachining is a versatile technology for three key reasons. First, the patterning of the structural and sacrificial layers is typically accomplished by etching processes that are insensitive to the crystalline structure of the films, thereby providing flexibility for planar free-form designs. Second, surface micromachining enables integrated multilevel structures using multiple layers of structural and sacrificial material. Third, there is no express restriction on the structural-sacrificial material system, as long as the compatibility between the structural and sacrificial materials is maintained. Therefore, different application-specific structural layers can be used in conjunction with suitable sacrificial layers.

Polysilicon surface micromachining using polysilicon as the structural material and SiO_2 as the sacrificial material has been the most widely used surface micromachining technique. When electrical isolation of the substrate and/or the structural components is required, Si_3N_4 is used as

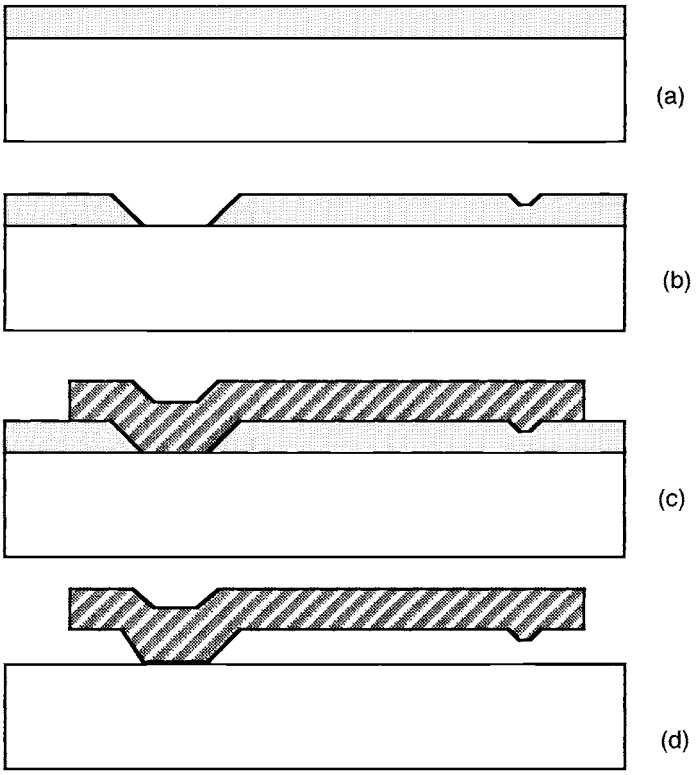


Fig. 1.7. Cross-sectional schematic demonstration of surface micromachining. (a) Sacrificial layer deposition, (b) definition of the anchor and bushing regions, (c) structural layer patterning, and (d) free-standing microstructure after release.

an insulator. In this process, hydrofluoric acid (HF) is used to dissolve the sacrificial oxide during release. Figure 1.8 presents a surface-micromachined shear-stress microsensor fabricated using a single structural layer of polysilicon. Another commercially used surface micromachining technique utilizes aluminum (Al) and photoresist as the structural and sacrificial layers, respectively. In this case, the release of the structural Al layer is accomplished by removing the sacrificial photoresist using a plasma etch. A number of other material systems have also been investigated as structural/sacrificial layers for surface micromachining: Al/polyimide, Si_3N_4/polysilicon, and Si_3N_4/SiO_2. The maximum thickness of structural layers in traditional surface micromachining is limited to 10 μm or less because of residual stresses in films. Excessive residual stress can lead to mechanical failure during fabrication. Furthermore, there are process limitations due to slow film deposition rates in traditional methods such as CVD, sputtering, and evaporation. Faster deposition rates can be realized for films that can be grown using pulsed laser deposition (PLD) or plating techniques.

1.2.4 Micromolding

Micromolding refers to fabrication of microstructures using molds to define the deposition of the structural layer. After the structural layer deposition, the final microfabricated components are realized when the mold is dissolved in a chemical etchant that does not attack the structural material. Micromolding is an additive process, in that the structural material is deposited only in those

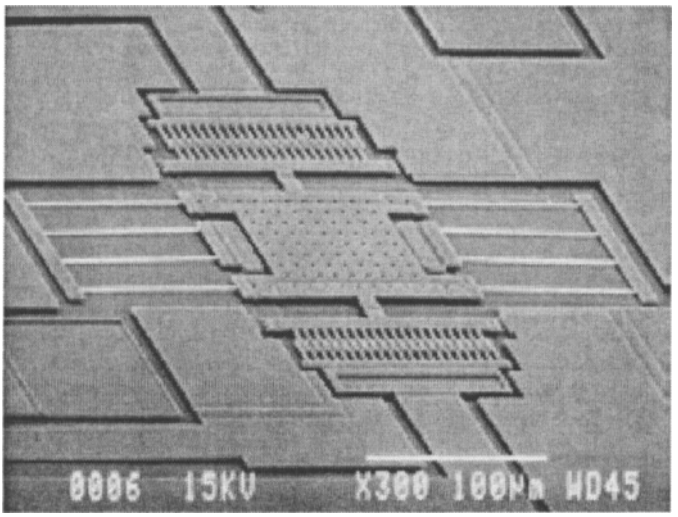


Fig. 1.8. SEM of a shear-stress microsensor fabricated by surface micromachining using a single structural layer of polysilicon.

areas constituting the microdevice structure. In contrast, bulk and surface micromachining are examples of subtractive micromachining processes, which feature blanket deposition of the structural material followed by etching to realize the final device geometry.

A widely known micromolding process is *lithographie, galvanoformung, und abformung* (LIGA). This German acronym means lithography, electroplating, and molding. The process can be used for the manufacture of high-aspect-ratio, 3D microstructures in a wide variety of materials (e.g., metals, polymers, ceramics, and glasses).[16,17] As shown in Fig. 1.9, high-intensity, low-divergence, hard x rays are used as the exposure source for the lithography. These x rays are usually produced by a synchrotron radiation source. Polymethylmethacrylate (PMMA) is used as the x-ray resist. Thicknesses of several hundreds of microns and aspect ratios of more than 100 have been achieved. A characteristic x-ray wavelength of 0.2 nm allows the transfer of a pattern from a high-contrast x-ray mask into a resist layer with a thickness of up to 1000 µm so that a resist relief may be generated with an extremely high depth-to-width ratio. The openings in the patterned resist can be preferentially plated with metal, yielding a highly accurate complementary replica of the original resist pattern. The mold is then dissolved away to leave behind plated structures with sidewalls that are vertical and smooth. It is also possible to use the plated metal structures as an injection mold for plastic resins. After curing, the metallic mold is removed, leaving behind microreplicas of the original pattern. By combining LIGA with the use of a sacrificial layer, it is also possible to realize free-standing micromechanical components.[18]

A chief drawback of the LIGA process is the need for a short-wavelength collimated x-ray source like a synchrotron. Consequently, LIGA-like processes using conventional exposure sources are being developed. Photoresists with high transparency and high viscosity can be used to achieve a single-coating mold thickness in the range of 15 to 500 µm.[19–21] Thicker photoresist layers may be realized by multiple coatings. In such photoresist layers, standard ultraviolet (UV) photolithography is used to achieve mold features with aspect ratios as high as 11:1.

Photosensitive polyimides are also used for fabricating plating molds.[22] The photolithography process is similar to conventional photolithography, except that polyimide works as a negative resist. In this process, about 10-µm-wide lines can be delineated in several tens of microns-thick

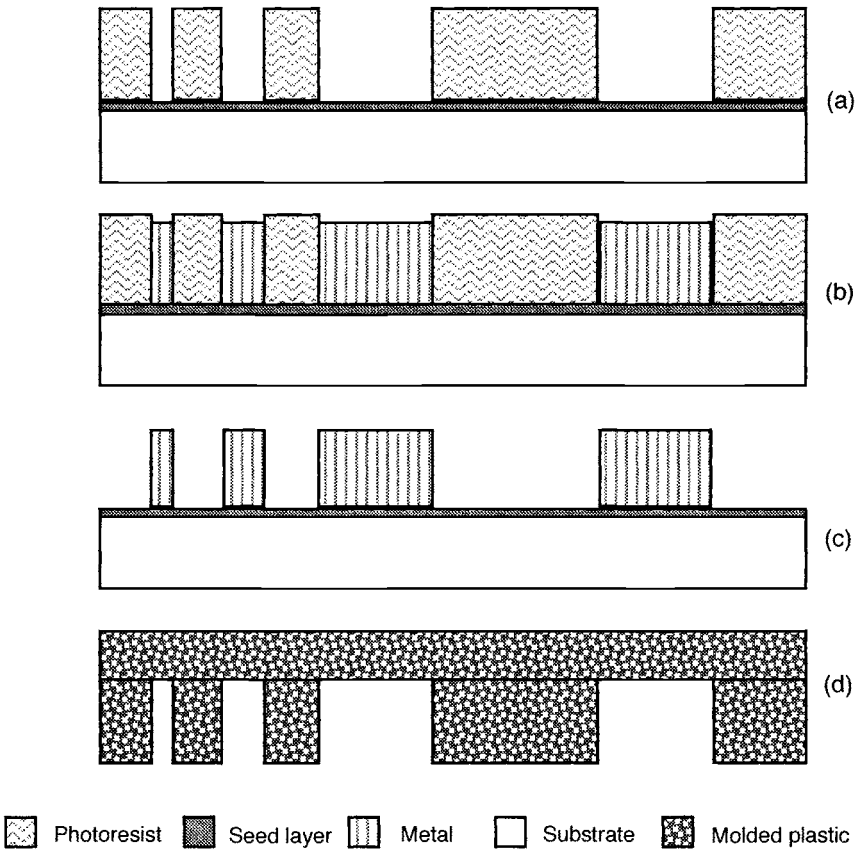


Fig. 1.9. Outline of the micromolding process using LIGA technology. (a) Photoresist patterning, (b) electroplating of metal, (c) resist removal, and (d) molded plastic components.

resist. A maximum aspect ratio of 8:1 can be achieved, but depends on the geometry of the mask layout. Polyimide is a very stable material and does not have to be cured to act as a plating mold, but it is also limited in terms of the thickness and the aspect ratio.

All methods stated above make use of lithography techniques to make a mold, but dry etching of polyimides to form high-aspect-ratio molds has also been reported.[23] In these methods, some modifications of traditional reactive ion etching (RIE) systems are necessary to achieve high-aspect ratios. For example, dry etching of fluorinated polyimides with a titanium (Ti) mask has been used for deep etching with high-aspect ratios, excellent mask selectivity, and smooth sidewalls.[23,24]

Using micromolding processes, it is possible to realize high-aspect-ratio metallic microstructures, which are especially attractive for certain applications, including reflective surfaces for optical components, low resistivity contacts for relays, magnetic metals for electromagnetic actuators/sensors, and microfabricated coils. Additionally, the larger thickness of high-aspect-ratio structures provides for greater stiffness perpendicular to the substrate, as well as for increased force/torque in electrostatic actuators. Plated nickel (Ni), copper (Cu), or alloys that contain at least one of these metals are the structural metals commonly used; Cr, SiO_2, polyimide, photoresist, and Ti have been often used as the sacrificial material.

1.3 MEMS Components

The miniaturization, multiplicity, and microelectronics characteristics of MEMS technology make it especially attractive to realize small-size, low-cost, high-performance systems integrated on one chip. Microfabricated pressure sensors have dominated the MEMS application market for the last two decades. With advances in IC technology and corresponding progress in MEMS fabrication processes in the last decade, additional integrated microsensor and microactuator systems are now being commercialized, and even more applications are expected to benefit. In this section, we present examples of some commercially available MEMS components selected on the basis of fabrication technique and system complexity. First, pressure sensors are presented as an example of a MEMS device fabricated using bulk micromachining, followed by integrated accelerometers that are fabricated by surface micromachining. Next, the suitability of MEMS technology in complex, array-type application systems is demonstrated using the example of a digital micromirror device (DMD). Finally, the potential of MEMS components in aerospace applications is discussed, and some promising devices are listed.

1.3.1 Pressure Sensors

MEMS technology has been utilized to realize a wide variety of differential, gauge, and absolute pressure microsensors based on different transduction principles. Typically, the sensing element consists of a flexible diaphragm that deforms due to a pressure differential across it. The extent of the diaphragm deformation is converted to a representative electrical signal, which appears at the sensor output.

Figure 1.10 shows a manifold absolute pressure (MAP) sensor for automotive engine control, designed to sense absolute air pressure within the intake manifold (manufactured by Motorola, Schaumburg, Illinois). This measurement can be used to compute the amount of fuel required for each cylinder in the engine. The microfabricated sensor integrates on-chip, bipolar op-amp circuitry and thin-film resistor networks to provide a high output signal and temperature compensation.

The sensor die/chip consists of a thin Si diaphragm fabricated by bulk micromachining. Prior to the micromachining, piezoresistors are patterned across the edges of the diaphragm region using standard IC processing techniques. After etching of the substrate to create the diaphragm, the sensor die is bonded to a glass substrate to realize a sealed vacuum cavity underneath the diaphragm. Finally, the die is mounted on a package such that the top side of the diaphragm is exposed to the environment through a port. A gel coat isolates the sensor die from the environment while allowing the pressure signal to be transmitted to the Si diaphragm. The ambient pressure forces the diaphragm to deform downward, resulting in a change of resistance of the piezoresistors. This resistance change is measured using on-chip electronics; a corresponding voltage signal appears at the output pin of the sensor package.

1.3.2 Accelerometers

Acceleration sensors are relatively newer applications of MEMS technology. Typically, the sensing element consists of an inertial mass suspended by compliant springs. Under acceleration, a force acts on the inertial mass, causing it to deviate from its zero-acceleration position, until the restoring force from the springs balances the acceleration force. The magnitude of the inertial-mass deflection is converted to a representative electrical signal, which appears at the sensor output.

Figure 1.11 shows a monolithic accelerometer (manufactured by Analog Devices, Inc., Norwood, Massachusetts), the ADXL-50, fabricated by surface micromachining and BiCMOS (a

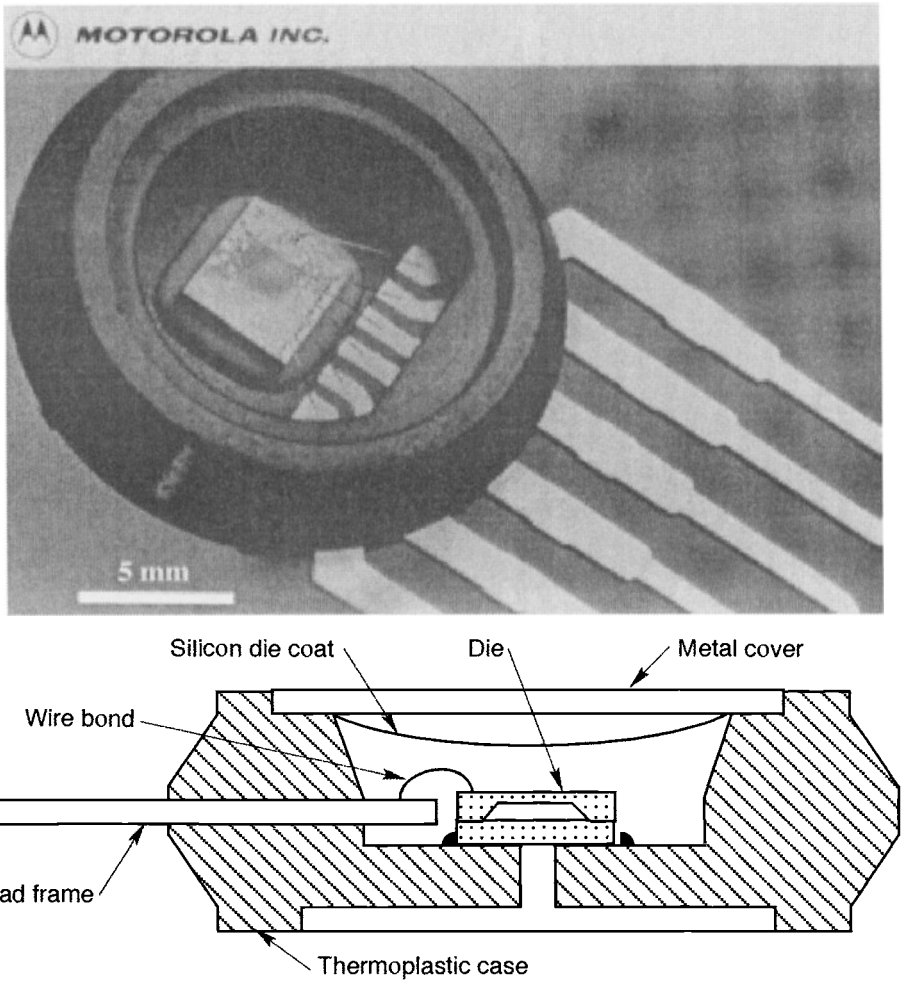


Fig. 1.10. Commercially available absolute pressure sensor. (a) Sensor package; (b) cross-sectional schematic (Motorola, Inc.).[25]

combination of bipolar junction transistor [BJT] and complementary metal-oxide semiconductor [CMOS]) processes. The inertial mass consists of a series of 150-μm-long fingerlike beams connected to a central trunk beam, all suspended 2 μm above the substrate by tether beams. The ADXL-50 uses a capacitive measurement method: the deflection of the inertial mass changes the capacitance between the finger beams and the adjacent cantilever beams. The sensor structure is surrounded by supporting electronics, which transduce the capacitance changes due to acceleration into a voltage, with appropriate signal conditioning.

The analog output voltage is directly proportional to acceleration, and is fully scaled, referenced, and temperature compensated, resulting in high accuracy and linearity over a wide temperature range. Internal circuitry implements a forced-balance control loop that improves linearity and bandwidth. Internal self-test circuitry can electrostatically deflect the sensor beam upon demand, to verify device functionality.

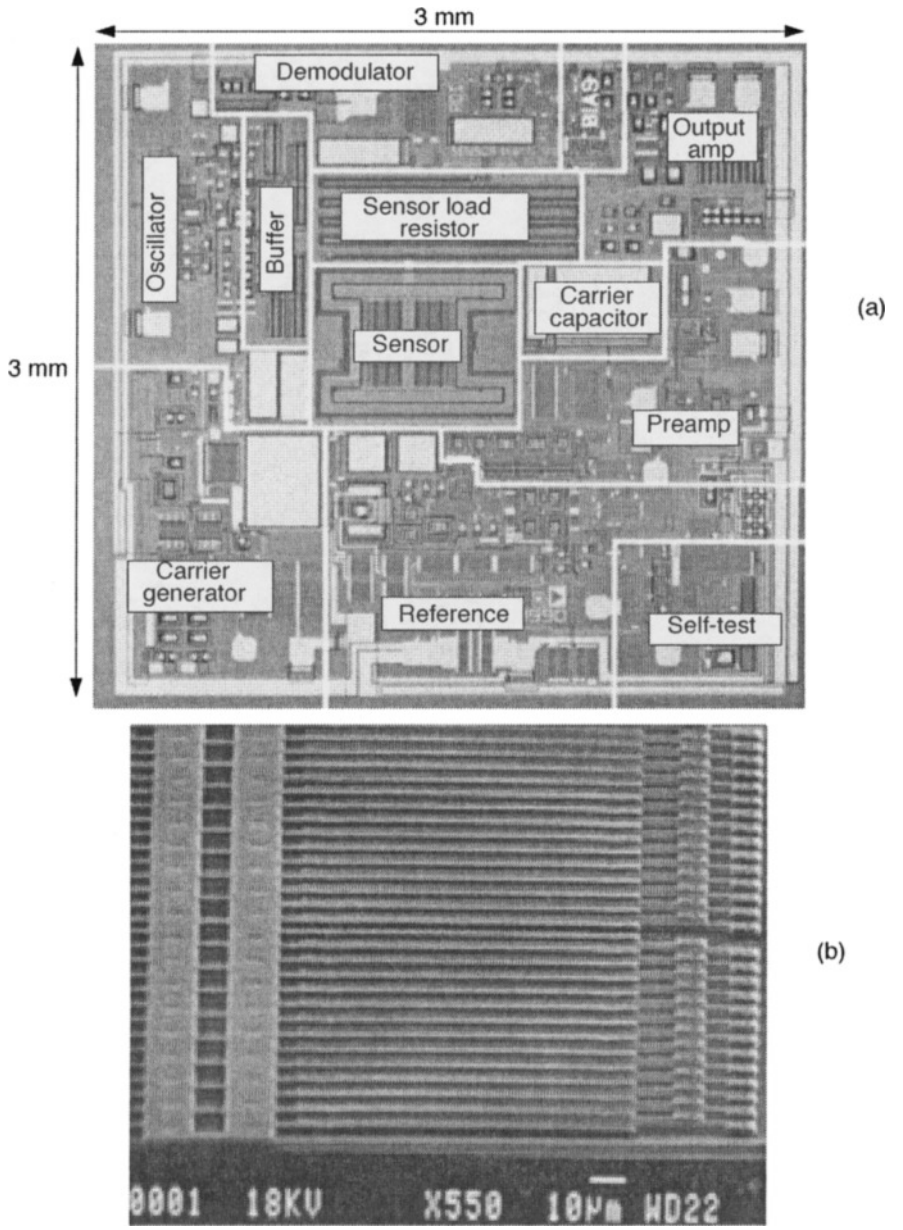


Fig. 1.11. Surface micromachined integrated accelerometer. (a) Chip overview, (b) close-up of sensor structure showing central trunk beam and fingers (Analog Devices, Inc.).

1.3.3 DMDs

The DMD (manufactured by Texas Instruments, Inc., Dallas, Texas) is a microchip consisting of a superstructure array of Al micromirrors functionally located over CMOS memory cells. The DMD digital light switch moves between the "on" and "off" states to create and reflect digital gray-scale images from its surface when light is applied. These digitally created images are transferred through appropriate optics and filters to create projected and/or digitally printed images.

The Al micromirror superstructure is realized by surface micromachining, while the underlying memory cells are fabricated using standard CMOS processes. The mirrors are hermetically sealed beneath nonreflecting glass to prevent contamination-induced failure. Figure 1.12 shows the details of the DMD microchip. Each mirror is 16 μm square with a 1-μm space between mirrors on all sides. The number of mirrors in use on a single chip can range from 307,200 to 1.3+ million (with one mirror per pixel).

To achieve digital operation, the DMD micromirrors are designed to be bistable. In the "on" mode, the mirrors deflect +10 deg, while in the "off" position, the mirrors rest at −10 deg. When a given CMOS memory cell is loaded with a digital 1, electrostatic forces switch the corresponding mirror "on" to reflect light into the aperture of an imaging lens. Memory cells loaded with a digital 0 cause the mirror to switch "off," and to direct incident light away from the imaging lens. In conjunction with appropriate optics, a color wheel, and electronic control circuitry, the DMD can be used to display high-quality projection images.

1.3.4 Sensors and Actuators in Aerospace Applications

Sensors are required in a variety of aerospace instrumentation, including fuel measurement and monitoring, landing gear, ice protection, and navigation. In small, private aircraft, the instrumentation is simple and may consist only of an altimeter to register height, an indicator to register airspeed, and a compass. The most modern airplanes and manned spacecraft, in contrast, have fully automated "glass cockpits," in which a tremendous array of sensor information is continually presented on the aircraft's height, attitude, heading, speed, cabin pressure and temperature, route, fuel quantity and consumption, and on the condition of the engines and the hydraulic, electrical, and electronic systems. Aerospace vehicles are also provided with inertial guidance systems for automatic navigation from point to point, with continuous updating for changing weather conditions, beneficial winds, or other situations. This array of instrumentation is supplemented by vastly improved meteorological forecasts, which reduce the hazard from weather, including such difficult-to-predict elements as wind shear and microburst.

Attitude and direction of aerospace vehicles are handled by flight controls that actuate elevators, ailerons, and rudders through a system of cables or rods. In sophisticated modern aircraft, there is no direct mechanical linkage between the attitude and direction devices used by the pilot and the actual controls used to achieve the changes in attitude and direction; instead, these controls are actuated by electric motors. The catch phrase for this arrangement is "fly by wire." In addition, in some large and fast aircraft, controls are boosted by hydraulically or electrically actuated systems. In both the fly-by-wire and boosted controls, the feel of the control reaction is fed back to the pilot by simulated means.

The use of MEMS devices in aerospace systems is expected to be highly application specific and would typically aim to reduce size, weight, and power consumption at the component level. Changes in both commercial and military markets for fixed-wing and rotor-wing aircraft demand increased performance with less weight. The cost advantage and electronic integration capabilities of MEMS enables the feasibility of distributed measurement and actuation. These features would be based on flexible location of smart transducers and decreased reliance on pneumatics, which would, in turn, lead to more accurate measurements, reduced vulnerability through redundancy, fewer moisture drain traps, and considerable weight savings.

In addition to conventional aircraft, evolution of MEMS technology should lead to the development of micro, unmanned aerial vehicles (μUAVs). These small flight vehicles would perform as aerial robots whose mobility could be used to deploy micropayloads to a remote site or to otherwise hazardous locations.

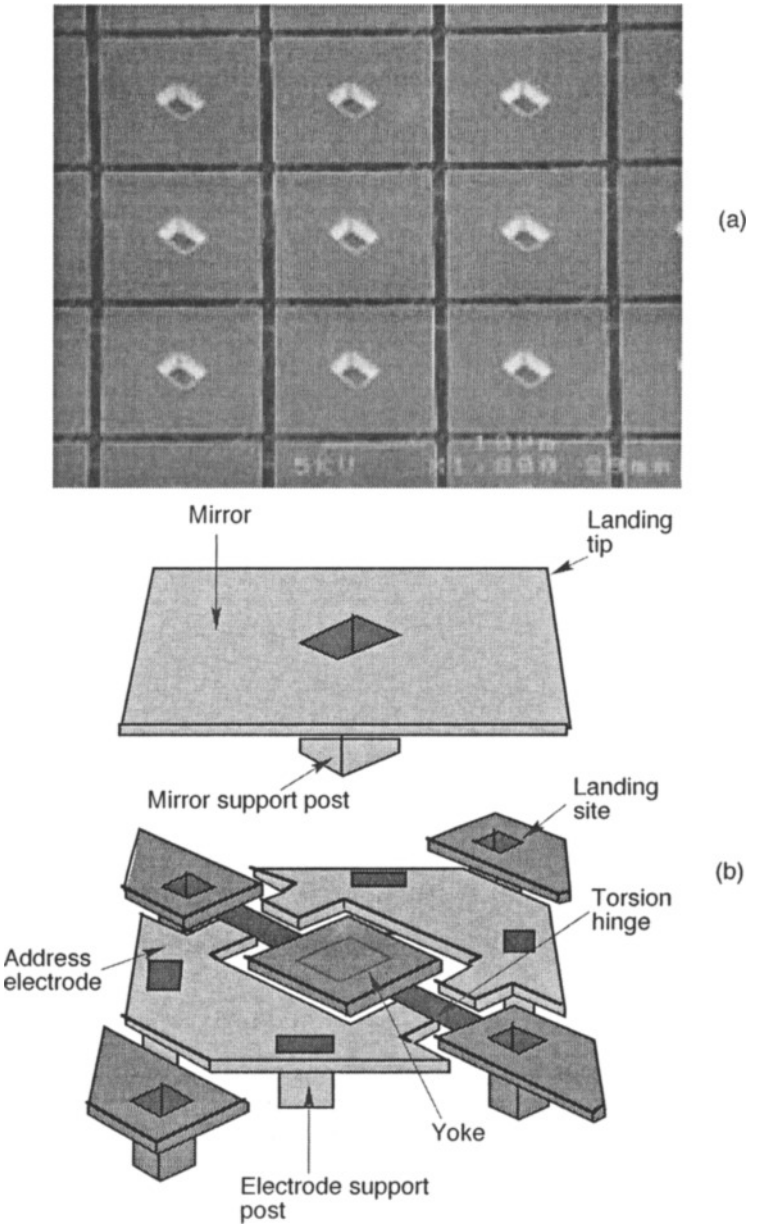


Fig. 1.12. DMD microchip. (a) Portion of the micromirror array, (b) exploded view of a single (16-μm-edge length) micromirror element (Texas Instruments, Inc.).

Figure 1.13 shows a commercially available pressure measurement instrument, the micro, air data transducer (manufactured by BF Goodrich Company, Richfield, Ohio), which measures static and total pressures using micromachined Si-based sensors. This instrument is only 25% of the size and weight of its conventional non-MEMS-based counterparts, and exhibits a 0.02% full-scale pressure accuracy. Applications include primary accuracy air data for flight control, cockpit display, navigation, and fire control.

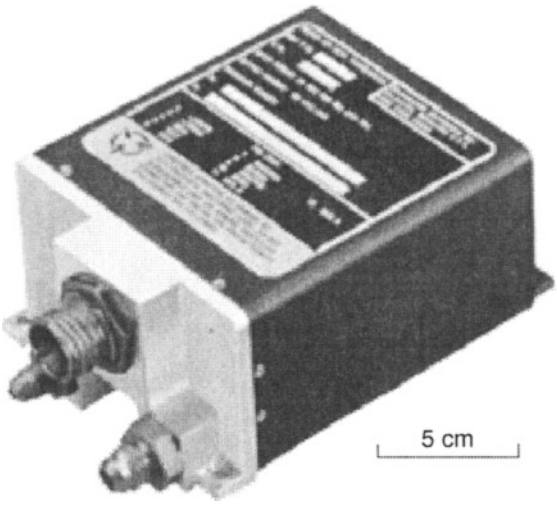


Fig. 1.13. Photograph of the Micro-Air Data Transducer, a commercially available MEMS-based sensor for use in aircraft (BF Goodrich Co.).

A number of other MEMS devices have been developed for aerospace applications. Although most of these devices are still at the research stage, their eventual integration into aerospace systems will revolutionize flight safety and performance. One of the devices that has been realized, the miniaturized ice detector, uses bulk micromachining and wafer bonding techniques. The detector is shown in Fig. 1.14.[26] The sensing element is 2 mm square and can detect ice films as low as 0.1 mm thick. Table 1.2 presents examples of MEMS devices with potential aerospace applications.

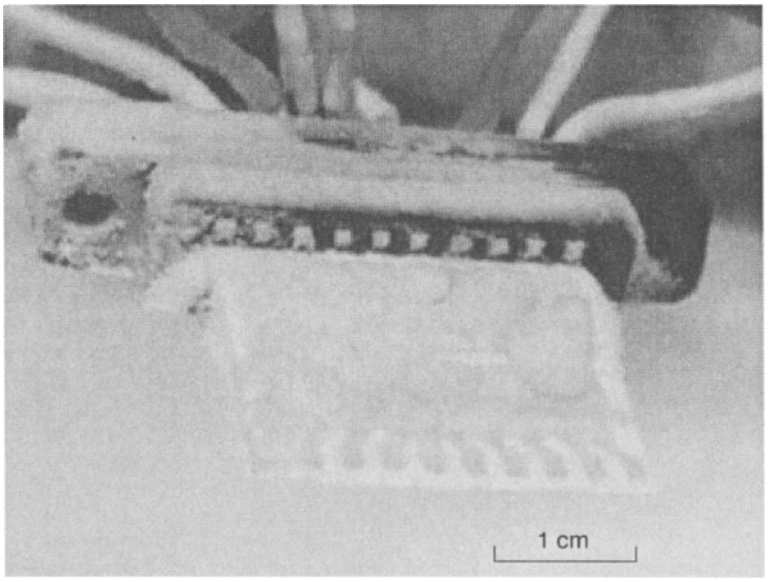


Fig. 1.14.Microfabricated ice detection sensor.

Table 1.2. Examples of MEMS with Potential in Aerospace Applications

Device Application	Fabrication Method	Transduction Principle	Organization
Shear Stress Sensor	Surface micromachining	Capacitive detection	Case Western Reserve University, Cleveland, OH
	Bulk micromachining	Optical detection	Massachusetts Institute of Technology (MIT), Cambridge, MA
Accelerometer	Integrated surface micromachining	Capacitive detection	Analog Devices, Norwood, MA
	Surface micromachining	Capacitive detection	Motorola, Phoenix, AZ
	Bulk micromachining	Piezoresistive detection	Endevco, Capistrano, CA
Pressure Sensor	Bulk micromachining	Piezoresistive detection	Lucas Novasensor, Sunnyvale, CA
	Bulk micromachining	Piezoresistive detection	Motorola, Phoenix, AZ
Angular Rate Gyroscope	Surface micromachining, micromolding	Capacitive detection	Draper Labs, Cambridge, MA
	Surface micromachining	Capacitive sensor	University of California, Berkeley
Drag Reduction	Bulk micromachining, micromolding	Magnetic flap actuator	University of California, Los Angeles
Fuel Atomization	Bulk micromachining	Precision nozzle	CWRU
Screech Control	Bulk micromachining	Electrostatic microactuator	University of Michigan, Ann Arbor
Communication Filters and Oscillators	Surface micromachining	Electrostatic resonators	University of Michigan
Microrelays	Surface micromachining, micromolding	Electrostatic actuator	CWRU
	Surface micromachining	Electrostatic actuator	Hughes Research Labs, Malibu, CA
	Surface micromachining, micromolding	Magnetic actuator	Georgia Institute of Technology, Atlanta
Optical Scanners	Surface micromachining	Electrostatic micromotor	CWRU
	Surface micromachining	Electrostatic resonator	UC, Berkeley

1.4 Commercial Applications

The potential of MEMS technology promises to revolutionize our present-day life-styles as much as the computer has. In addition to completely new applications enabled by MEMS technology, existing applications will likely be replaced by miniaturized, low-cost, high-performance, "smart" MEMS technology. The potential for cost-effective and high-performance systems has attracted attention from both government and industry alike. The substantial up-front investment often required for successful, large-volume commercialization of MEMS is likely to limit the initial involvement to larger companies in the IC industry. These companies can leverage their existing capital investment in semiconductor processing equipment toward the development of MEMS components for large-volume applications.

1.4.1 MEMS Market

Currently, MEMS markets and demand are overwhelmingly in the commercial sector, with the automobile industry being the main consumer for micromachined pressure sensors and accelerometers. Market studies predict that the value of MEMS products will increase to between $12 and $14 billion by the year 2000 (see Fig. 1.15) and that no one product and/or application area will dominate the MEMS industry for the foreseeable future.[27]

The MEMS market for sensors will continue to grow, particularly for sensors with integrated signal processing, self-calibration, and self-test. However, a substantial portion of the MEMS market will be in non-sensing, actuator-enabled applications, such as scanners, fuel-injection systems, and mass data storage devices. Furthermore, because MEMS products will be embedded in larger, non-MEMS systems (e.g., printers, automobiles, biomedical diagnostics), the products will enable new and improved systems, with projected market value approaching $100 billion in the year 2000.[27]

1.4.2 MEMS Industry Structure

A number of companies are already marketing MEMS devices and systems for commercial use. These companies include a broad range of manufacturers of sensors, industrial and residential control systems, electronic components, automotive and aerospace electronics, analytic instruments, and biomedical products. Examples of such companies include Goodyear, Honeywell, Lucas Novasensor, Motorola, Hewlett-Packard, Analog Devices, Texas Instruments, Siemens, and Hitachi. In addition, many small, emerging businesses have also been formed to commercialize MEMS components.

With the advent of commercialization of MEMS, many technology requirements being identified are capabilities beneficial to the industry as a whole, but too costly to develop by any one company. MEMS manufacturing is heavily dependent on microelectronics manufacturing, and at the moment, there is no MEMS-equipment and material-supplier infrastructure separate from that of microelectronics equipment and material industry. While advanced MEMS device designs, systems concepts, and fabrication processes will continue to be important, advances in MEMS manufacturing resources will pace future development, commercialization, and use of MEMS.

Unlike microelectronics, where a few types of fabrication processes satisfy most microelectronics manufacturing requirements, MEMS, given their intimate and varied interaction with the physical world, exhibit a greater variety of device designs and associated manufacturing resources. For example, the thin-film structures created using surface micromachining techniques, while well-suited for the relatively small force encountered in inertial measurement devices, are not adequate for MEMS fluid valves and regulators. Similarly, the thicker structures created using a combination of wafer etching and bonding, while well-suited to the higher forces and motions

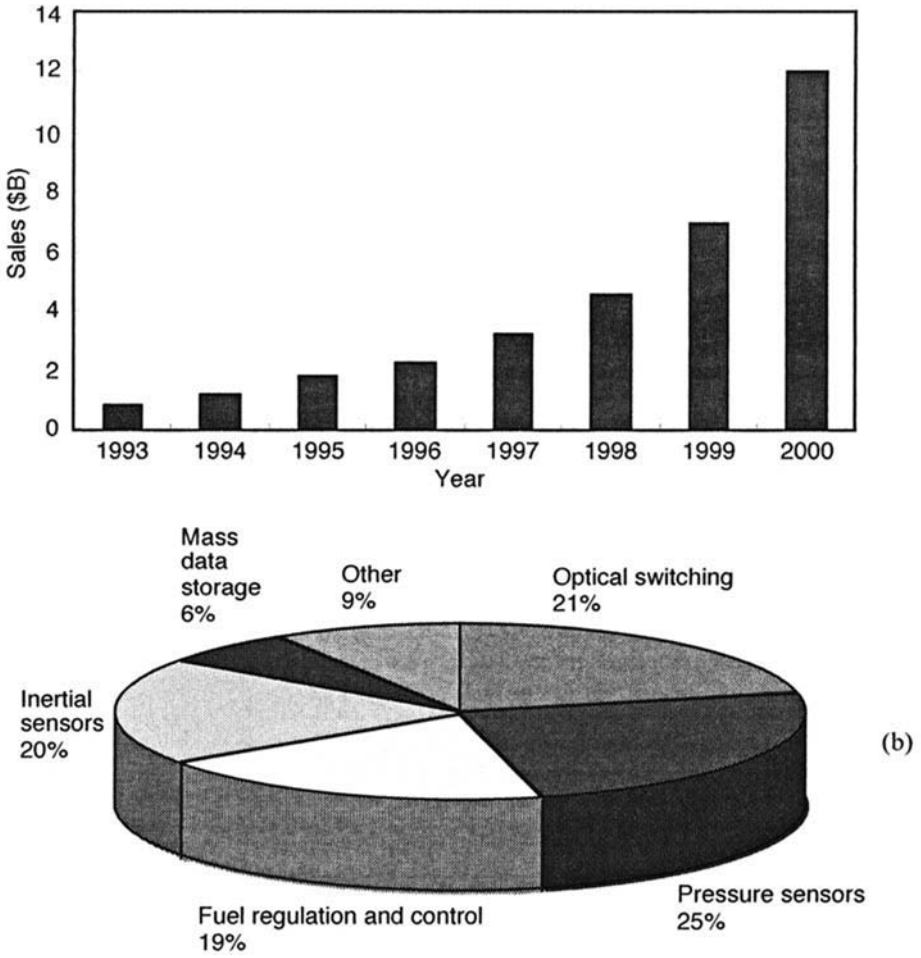


Fig. 1.15. Projected worldwide MEMS market through the year 2000.[27] (a) Growth of worldwide market in MEMS components and devices: (b) non-sensor market segments in the year 2000, which will constitute at least 50% of the market for MEMS products.

in fluid valves and regulators, consume too much power to be used for the fabrication of microoptomechanical aligners and displays. There is not likely to be a MEMS equivalent of a CMOS process like that in microelectronics that will satisfy the majority of MEMS device fabrication needs.[27] MEMS design is strongly coupled to packaging requirements, which are dictated by the application environment.

The different MEMS fabrication processes and equipment will often be developed by larger firms with a particular and large commercial market as the target. Typically, the firm developing the manufacturing resources needs to focus on the production of products for those one or two markets driving applications. But in most cases, once the manufacturing resource is developed, numerous products for smaller markets could be addressed with the same manufacturing resources. No single one of these smaller markets would have justified the development of the fabrication process. For the firms that have developed the manufacturing resource, addressing small and fragmented markets is not currently economically justifiable, given the market diversity and

the embryonic state of electronic design aids. Most of these specialized markets will only be attractive and economically justifiable to smaller businesses that, however, do not have (nor would they want to duplicate) the manufacturing resources.

1.5 Trends in MEMS Technology

MEMS technology is extending and increasing the ability to both perceive and control the environment by merging the capabilities of sensors and actuators with information systems. Future MEMS applications will be driven by processes that enable greater functionality through higher levels of electronic-mechanical integration and greater numbers of mechanical components working either alone or together to enable a complex action. These process developments, in turn, will be paced by investments in the development of new materials, device and systems design, fabrication techniques, packaging/assembly methods, and test and characterization tools.

1.5.1 Design and Simulation

MEMS is more demanding of design aids than microelectronics production. Most industrial designs of physical sensors today are based on detailed finite-element modeling of the mechanical microstructures using software available for conventional mechanics.

MEMS requires new drawing and layout tools to generate the patterns that will be used to add or remove material during processing. In addition, MEMS requires a number of different modeling tools, including simulators for mechanical deformation, electrostatic fields, mechanical forces, electromagnetic fields, material properties, and electronic devices. MEMS also needs the connective algorithms to reconcile and blend results from all the different simulators.

As devices become more complex and multiple simulators are involved, the complexity of both the simulations and the coupling increases considerably. Traditional modeling techniques become impractical and may even fail. Radically new approaches to modeling and simulation for the many physical effects and different MEMS functions have to be developed.

1.5.2 Materials Issues

An extensive, well-documented materials database that meets the requirements of MEMS development is essential for continuing progress in the field. Many of the new material property simulators will need new models and data to relate process parameters to material properties relevant for MEMS design.

The accuracy of the existing microelectronic device simulators is built on historic and huge amounts of material and device measurements, coupled to carefully controlled process conditions. By knowing the relationship between processing conditions and the resulting material parameters, microelectronics manufacturers can control material properties, and hence, device yields. Circuit designers are typically interested in those material properties that relate to the electronic function of the devices, such as doping levels and dielectric constants.

The material needs of the MEMS field are well recognized but are at a preliminary stage. In addition to single-crystal Si, polysilicon, Si_3N_4, and SiO_2, other materials are being explored for MEMS. Interesting examples include SiC, shape memory alloy (SMA) metals, permalloy, and high-temperature superconductive materials. All these materials possess certain unique properties that, when combined with MEMS technology, make them attractive for certain applications.

A thorough understanding of the material properties of existing MEMS materials is just as important as the development of new materials for MEMS. There are very few reliable measurements of material properties (for example, modulus, residual stress, or reflectivity) relevant to the production of MEMS. The goal of studying the material properties in MEMS, and in thin films generally, is to develop models that relate process parameters to the film microstructure, as well

as to the corresponding mechanical, electrical, optical, and thermal properties. Chapter 3 elaborates on the material properties and the required tests to enable a valid database.

1.5.3 Integration with Microelectronics

Future MEMS products will demand yet higher levels of electrical-mechanical integration and more intimate interaction with the physical world. The full potential of MEMS technology will only be realized when microelectronics is merged with the electromechanical components. Integrated microelectronics provides the intelligence to MEMS and allows closed-loop feedback systems, localized signal conditioning, and control of massively parallel actuator arrays.[27]

Although MEMS fabrication uses many of the materials and processes of semiconductor fabrication, there are important distinctions between the two technologies. The most significant distinctions are in the process recipes (the number, sequence, and type of deposition, removal, and patterning steps used to fabricate devices) and in the end stages of production (bonding of wafers, freeing of parts designed to move, packaging, and test). The fundamental challenge of using semiconductor processes for MEMS fabrication is not so much in the type of processes and materials used, but more in the way those processes and materials are used.

MEMS will need the development of operating conditions on standard semiconductor equipment suited and optimized to the requirements of MEMS. For other processing steps unique to MEMS, the development of new manufacturing equipment and associated processes will be required. Table 1.3 lists some of the specialized process equipment that is required to enhance the manufacturability of MEMS.

1.6 Journals and Conferences

Before 1980, the literature on solid-state sensors and actuators was scattered in various application fields such as electronic devices, automobiles, instrumentation, materials, physics, and analytical chemistry. The journal *Sensors and Actuators* was first published in 1980 to provide a forum for publication of papers in the field. Another journal, *Sensors and Materials*, issued by MYU Japan, began publication in 1989. Rapid advances in device design, fabrication, materials,

Table 1.3. Examples of Process Equipment Specific to MEMS

Fabrication Technology	Process Equipment
Surface micromachining	Release and drying systems to realize free-standing microstructures
Bulk micromachining	Dry etching systems to produce deep, 2D free-form geometries with vertical sidewalls in substrates
	Anisotropic wet etching systems with protection for wafer front sides during etching
	Bonding and aligning systems to join wafers and perform photolithography on the stacked substrates
Micromolding	Batch-plating systems to create metal molds in LIGA process
	Plastic injection molding systems to create components from metal molds

testing, packaging, and applications, as well as the rapid expansion of the field, have brought about many journals, symposia, and conferences to report on the progress being made. There are now many publications dealing with sensors and actuators, including trade journals and regional publications. In the MEMS area, the *Journal of Microelectromechanical Systems* is a quarterly (started in 1992) published jointly by the Institute of Electrical and Electronics Engineers (IEEE) and the American Society of Mechanical Engineers (ASME), the *Journal of Micromechanics and Microengineering* is a quarterly (started in 1991) published by the American Institute of Physics, while *Microsystem Technology* is a quarterly (started in 1995) published by Springer-Verlag.

The International Conference on Solid State Sensors and Actuators, also referred to as the Transducers Conference, was established in 1981. The conference sponsors biannual meetings (during odd years), rotating between the United States, Japan, and Europe. The latest meeting concluded in Chicago in June 1997 (called Transducers '97). The conference also publishes a technical digest. In addition, some of the papers presented at the conference are published in special issues of *Sensors and Actuators*.

Regional conferences on sensors, actuators, and MEMS are held in both Japan and Europe. In the United States, a workshop on solid-state sensors and actuators has been held at Hilton Head, North Carolina, during even years. Technical digests from these workshops are also published. Another series of international conferences, entitled IEEE Workshop on Micro Electro Mechanical Systems, started in 1987 and has met annually between 1989 and 1998. Each of the conferences has published a proceedings volume. Finally, a number of conferences in other fields (e.g., International Electron Device Meeting, Device Research Conference/Materials Research Conference, The Electrochemical Society Meeting, and many SPIE conferences) hold sessions on microsensors, microactuators, and MEMS.

1.7 References

1. J. Bardeen and W. H. Brattain, "The Transistors, a Semiconductor Triode," *Phys. Rev.* **74**, 230 (1948); and W. Shockley, J. Bardeen, and W. H. Brattain, "Electronic Theory of the Transistor," *Science* **108**, 678–679 (1948).
2. C. S. Smith, "Piezoresistive Effect in Germanium and Silicon," *Phys. Rev.* **94** (April 1954).
3. *Micromachine Devices* **1**(2) (1996).
4. E. M. Blaser, W. H. Ko, and E. T. Yon, "A Miniature Digital Pressure Transducer," *24th Annual Conference on Engineering in Medicine and Biology* (Las Vegas, Nevada, November 1971), p. 211.
5. W. H. Ko, M. H. Bao, and Y. D. Hong, "A High Sensitivity Integrated Circuit Capacitive Pressure Transducer," *IEEE Trans. Elect. Dev.* **ED-29**, 48–56 (1982).
6. R. M. Finne and D. L. Klein, "A Water-Amine-Complexing Agent System for Etching Silicon," *J. Electrochem. Soc.* **114**, 965 (1967).
7. J. B. Price, "Anisotropic Etching of Silicon with KOH-H2O-isopropyl Alcohol," in *Semiconductor Silicon, Second International Symposium on Silicon Materials Science and Technology* (Electrochemical Society, Princeton, NJ,13–18 May 1973), pp. 339–353.
8. H. C. Nathanson and R. A. Wickstrom, "A Resonant-Gate Silicon Surface Transistor with High Q Bandpass Properties" *Appl. Phys. Lett.* 7, 84 (1965).
9. R. T. Howe, "Surface Micromachining for Microsensors and Microactuators," *J. Vac. Sci. Technol.* **16**, 1809–1813 (1988).
10. M. Mehregany, K. J. Gabriel, and W. S. N. Trimmer, "Integrated Fabrication of Polysilicon Mechanisms," *IEEE Trans. on Electron Devices* **ED-35**, 719–723 (June, 1988).
11. L. S. Fan, Y. C. Tai, and R. S. Muller, "Integrated Movable Micromechanical Structures for Sensors and Actuators," *IEEE Trans. on Electron Devices* **ED-35**, 724–730 (June 1988).
12. L. S. Fan, Y. C. Tai, and R.S. Muller, "IC-Processed Electrostatic Micro-motors," in *Technical Digest, IEEE International Electron Devices Meeting* (San Francisco, CA, December 1988), pp. 666–669.

13. M. Mehregany, S. F. Bart, L. S. Tavrow, J. H. Lang, S. D. Senturia, and M. F. Schlecht, "A Study of Three Microfabricated Variable-Capacitance Motors," *Sensors and Actuators* **A21-23**, 173–179 (February–April 1990).
14. K. E. Petersen, "Silicon as a Mechanical Material," *IEEE Proc.* **70**, 420–457 (1982).
15. J. Bhardwaj, H. Ashraf, and A. McQuarrie, "Dry Silicon Etching for MEMS," *Proceedings of the Symposium on Microstructures and Microfabrication,* Spring Meeting of The Electrochemical Society (Montreal, Quebec, Canada, May 1997).
16. W. Ehrfeld *et al.*, *Proceedings of the IEEE Micro Robots and Teleoperators Workshop* (Hyannis, MA, November 1987).
17. H. Guckel, T. R. Christenson, K. J. Skrobis, D. D. Denton, B. Choi, E. G. Lovell, J. W. Lee, S. S. Bajikar, and T. W. Chapman, "Deep X-ray and UV Lithographies for Micromechanics," in *Technical Digest, 1990 IEEE Solid-State Sensor and Actuator Workshop* (Hilton Head, SC, June 1990), pp. 118–122.
18. T. R. Ohnstein *et al.*, "Tunable IR Filters Using Flexible Metallic Microstructures," *Proceedings of the IEEE Micro Electro Mechanical Systems Workshop (*Amsterdam, Netherlands, January 1995), pp. 170–174.
19. H. Miyajima and M. Mehregany, "High-Aspect-Ratio Photolithography for MEMS Applications," *IEEE Journal of Microelectromechanical Systems* **4**(4), 220–229 (December 1995).
20. B. Lochel, A. Maciossek, H. J. Quenzer, and B. Wagner, "UV Depth Lithography and Galvanoforming for Micromachining," *Proceedings of Second International Symposium on Electrochemical Microfabrication,* 186th Meeting of The Electrochemical Society (Miami Beach, FL, October 1994).
21. M. Despont, H. Lorenz, N. Fahrni, J. Brugger, P. Renaud, and P. Vettiger, "High-Aspect-Ratio, Ultrathick, Negative-Tone Near-UV Photoresist for MEMS Applications," *Proceedings of the IEEE Micro Electro Mechanical Systems Workshop: MEMS '97* (Nagoya, Japan, January 1997), pp. 518–522.
22. C. H. Ahn and M. G. Allen, "Fully Integrated Micromagnetic Actuator with a Multilevel Meander Magnetic Core," *Technical Digest–IEEE Solid-State Sensor and Actuator Workshop* (1992), pp. 16–18.
23. A. Furuya, F. Shimokawa, T. Matsuura, and R. Sawada, "Micro-Grid Fabrication of Fluorinated Polyimide by Using Magnetically Controlled Reactive ion Etchine," *Proceedings of the IEEE Micro Electro Mechanical Systems Workshop: MEMS '93* (Ft. Lauderdale, FL, February 1993), pp. 59-64.
24. F. Shimokura, A. Furuya, and S. Matsui, "Fast and Extremely Selective Polyimide Etching with a Magnetically Controlled Reactive Ion Etching System," *Proceedings of the IEEE Micro Electro Mechanical Systems Workshop* (Nara, Japan, February 1992), pp. 192–197.
25. L. Ristic, *Sensor Technology and Devices* (Artech House, Boston, 1994).
26. R. G. DeAnna, M. Mehregany, and S. Roy, "Microfabricated Ice-Detection Sensor," *Proceedings of the Conference on Smart Structures and MEMS,* SPIE Symposium on Smart Structures and Materials (San Diego, CA, March 1997).
27. *Microelectromechanical Systems–A DOD Dual Use Technology Industrial Assessment,* Defense Advanced Research Projects Agency, U. S. Department of Defense (1995).

2

Microengineering Space Systems

H. Helvajian* and S. W. Janson*

2.1 Introduction

2.1.1 MEMS for Space Systems, an Overview

The evolution of microelectromechanical systems (MEMS) from laboratory curiosities to commercial off-the-shelf components (COTS) is being driven by government investments and strong market forces. These drivers have historically focused on terrestrial applications, which currently dominate MEMS development and usage, and will continue to do so well into the next century. MEMS offers a capability for mass-producing small, reliable, intelligent instruments at reduced cost by reducing the number of piece-parts, eliminating manual-assembly steps, and controlling material variability. These features, together with reduced mass and power requirements, are what space-system designers dream about. To gain an initial understanding of the problems facing spacecraft engineers, consider what your personal computer or other consumer product would look like if:

- The delivery charge to take the product home was between $10,000 and $100,000 per kg.
- The delivery truck had no springs or shock absorbers and traveled over really bumpy roads.
- The product kicked itself off the delivery truck and configured itself to receive your commands while the truck drove itself to the junkyard.
- The product ran off of solar cells by day and rechargeable batteries by night.
- The product had to withstand radiation levels that are 4 to 7 orders of magnitude greater than what you face.
- The product had to operate for 3 to 15 yr without maintenance or mechanical upgrades.

No wonder commercial Earth-orbiting spacecraft cost on the order of $50,000 per kg, not to mention similar costs for the one-shot delivery truck (the launch vehicle). We believe that market forces alone will stimulate the use of MEMS in space systems and foster further development of MEMS specifically for space application. If a commercial space systems provider can use MEMS to save a few kilograms on a spacecraft, perhaps the provider can then justify a low-volume custom foundry run specialized for space-microengineered systems. The saving of tens of kilograms in mass while maintaining capability can result in substantial profit for a provider.

Space systems comprise more than just spacecraft; they include launch vehicles and the ground-based systems used for tracking, command and control, and data dissemination (pictures of the Earth, telephone calls, movies via satellite, etc.). MEMS technology would allow development of new commercial uses of space in the proliferation of miniaturized ground transmitters with on-board sensors. MEMS, coupled with the current generation of digital electronics and telecommunication circuits, can be used for distributed remote-sensing applications. For example, transmitters the size of a fist and smaller can send environmental information, such as local atmospheric pressure, temperature, and humidity, directly to satellites. With Motorola's Iridium or other satellite telephone systems, real-time field data from remote locations can be just a phone call away. Similar remote sensors mounted on launch vehicles can simultaneously monitor

*Center for Microtechnology, The Aerospace Corporation, El Segundo, California.

vibration and acceleration at different locations on the vehicle. MEMS enables mass-production of partially or fully integrated remote sensors inexpensive enough to be "throw-away" add-ons for many aerospace applications. Microengineered devices or application-specific integrated microinstruments (ASIMs) will begin to infiltrate space systems by the end of this decade and become rapidly assimilated during the next. The Aerospace Corporation (Aerospace) coined the term ASIM to define a microengineered instrument designed to fulfill a specific need or solve a specific problem. The ASIM is a conceptual tool enabling the aerospace engineer to solve complex problems by reducing the problem into piecemeal ASIM solution components. The derived benefits of this technology for space systems are quite clear: dramatic cost reductions in manufacturing and operation.[1]

MEMS will also enable a radically new way of building and using spacecraft. Silicon, for example, can be used as a multifunctional material: as structure, electronic substrate, MEMS substrate, radiation shield, thermal control system, and optical material. With proper spacecraft design, it can provide these functions simultaneously. A "silicon" satellite composed of bonded thick wafers could be manufactured by one or more semiconductor foundries. Batch-fabrication would allow mass-production of spacecraft from one-hundred to several-thousand unit lots, which would enable dispersed satellite architectures and spacecraft designs meant for "single-function" and disposable missions. Alternatively, thinned MEMS electronics die, thin-film solar cells, and patterned metal antennas on 10- to 100-μm-thick Kapton™ sheets could permit the development of very large planar-sheet spacecraft for missions where large aperture (power and antenna gain) and low mass are a necessity. A synthetic aperture radar "sail" satellite, proposed by the French Centre National d'Etudes Spatiales (CNES), approaches this goal by allowing the radar sail to provide most of the spacecraft resources, for example, power, payload, and gravity-gradient stabilization.[2]

The chapter is organized into five sections:

- Introduction—Overview of MEMS for space systems
- Applications in which MEMS technology would be useful to space systems
- The silicon satellite concept
- Manufacturing future space systems
- Conclusions

2.1.2 Near-Term Applications of Microengineered Systems in Space

The U.S. National Aeronautics and Space Administration (NASA) and Department of Defense (DOD) have studied the use of microengineered systems for space applications. NASA's Jet Propulsion Laboratory has adopted the technology for producing "smaller, faster, cheaper" space systems to accomplish more interplanetary missions per year in the face of flat or declining budgets.[3] The theory is that many less-ambitious missions spread risk and prevent major program failures; the loss of a single spacecraft is tolerable if several others are in place or soon will be. The Mars Pathfinder mission successfully demonstrated what could be accomplished with "micro" spacecraft that used many COTS components. The "New Millennium"[4] and "X-2000"[5] efforts will utilize more MEMS technology ultimately to produce 10-kg-class interplanetary spacecraft. Most MEMS and ASIM examples discussed in this chapter were developed under some form of U.S. Government funding (e.g., DOD, NASA). In Europe, there is a similar desire to incorporate microengineered systems in space. Space systems development within the European Union (EU) is directed by the European Space Agency (ESA), which in 1997, identified micro/nanotechnology (MNT) as a relevant theme for its future programs. The incorporation of MNT within the ESA program framework is a direct result of two Round-Table meetings hosted by EU industry and

ESTEC representatives in Noordwijk, The Netherlands.[6] These Round-Table sessions established a forum to showcase EU technologies that might be applicable to space systems and to outline the ESA perspective. ESA representatives identified several issues to be resolved before MNT is routinely incorporated into ESA space systems. These same issues have also emerged in the U.S. space community and include the following requisites:

- New ways for designing systems
- Cultural change within the aerospace community to fully exploit the advantages of microsystems technology
- Characterization of materials, systems, and new phenomena at the micron scale
- Customization and subsequent fabrication of space-applicable devices in low-production batches, perhaps necessitating the development of low-cost design, production, and test technologies
- Packaging and interconnection schemes for other than electrical inputs
- Customization of embedded software for space applications
- Development of a new product-assurance philosophy

Examples of microsystem insertion opportunities not readily apparent to the aerospace community but identified by ESA include:

- Low-cost transmission lines, power dividers, phase shifters, and feed horns for frequencies above 100 GHz
- Integration of antennas with solar cells and necessary electronics
- Integration of micro accelerometers, gyroscopes, and charge-coupled devices (CCDs) with small optical apertures

ESA faces the same problem as the U.S. space community: it cannot compete with the blossoming microsystem terrestrial market but must find a way to leverage this technology. As an example, ESA will leverage technology developed under the Brite-Euram projects, which were initially designed to increase the competitiveness of the European industry through targeted research and development on priority industrial objectives. Some of these projects have applications to space systems. As in the United States and Japan, there appears to be considerable industrial interest in the development of micropumps and microvalves with actuation schemes incorporating electrostatics, thermo-pneumatics, shape memory alloys, or the large magnetorestriction effect. The latter approach is somewhat new in that it incorporates the magnetostrictive material (rare earth transition metal thin film) in the microactuating membrane.[7] The major advantages of using these materials over thermal-actuation schemes are the fast response time and noncontact operation. There also appears to be large interest in the development of functionalized materials that can be deposited over large areas. These materials will be useful in future aerospace systems because they enable multifunctionality, thereby reducing the "parts-count." Examples of these novel materials include:

- Thick-film ferroelectric actuators made by sintering ultrafine powder piezoelectric particles that are suspended in a colloid[8]
- Well-established ferrofluids (colloidal suspensions of single domain magnetic nanoparticles, typically 10 nm, in a liquid medium)[9] applied to microinstrumentation
- Low-cost electrochemical deposition (ECD) processes for the production of microstructured permanent magnet devices (CoWPt type)[10]
- Silicates and phosphate glass layers having thicknesses in the 3 to15 μm range for 1.55 μm wavelength waveguiding applications[11]

To accelerate the development of microinstruments and rapid prototyping of designs, the Brite-Euram program has also initiated a commensurate program on developing a versatile electronics package intended to drive/control the microinstruments. Work is being done to fabricate a set of programmable analog cells and routing resources on a chip. The cells include analog interfaces, digital-to-analog and analog-to-digital circuits, amplifiers, and filters. The chip integrates the analog functions of approximately 5-k programmable digital gates, an 8051-based microprocessor core, and program/control memory. A function library also comes with this device to enable a bridge to commercial standard-cell libraries and the migration of programmed prototype segments to hard-wire fabrication using classical application-specific integrated circuit (ASIC) solutions.[12] Examples of space-specific applications sponsored by EU member-state space agencies include the development of a compact ultraviolet (UV)/Vis-spectrometer by the Institute of Microtechnology, Mainz (IMM), Germany. This Hadamard transform optical spectrometer combines the advantages of a diode spectrometer (compactness) with that of a multislit spectrometer (higher signal-to-noise, increased resolution).[13] Other examples include a micro sun-sensor[14] and an infrared (IR) static Earth sensor.[15] There is also ample technology crossover from terrestrial to space applications in the development of compact high-resolution vision systems,[16] microrefrigerators, and microcalorimeters.[17] A specific space application area that cannot readily transfer technology from terrestrial applications is micropropulsion. As a satellite subsystem, propulsion systems are often designed to fit a particular satellite mission and are thus customized. Given that satellites are currently not mass produced in lots greater than 100, there is little incentive to develop standard micropropulsion platforms. However, a group is evaluating the required micro-machined components for developing a miniaturized propulsion platform.[18]

Microsystem technology has also had an impact on determining the method for developing future space life science experiments. Numerous groups are developing biochemical analysis systems specifically for space life science applications. These microsystems are true complex instruments that incorporate reagants, micropumps (self-priming/bubble tolerant[19]), microvalves, microreactors, microsensors, and microflow control schemes with semiautonomous data acquisition electronics. The sensing techniques include capillary electrophoresis,[20] multisensor array chips (O_2, CO_2, pH, ion-concentrations of for example sodium, potassium, calcium),[21] heterogeneous immunoassay,[22] and measurement of calibrated conductivity.[23] These micro total-analysis systems have a large terrestrial application base as a disposable field instrument, but the deployment of the International Space Station (ISS) in the next few years will stimulate interest in space life sciences, promoting further research and development, and technology cross-over.[24]

An increasing amount of research is focusing on developing components/devices/systems for space missions based on nanotechnology rather than microtechnology.[25] Understanding and exploiting design principles found in nature is key to this technology. Biomimetics is a specialized field devoted to understanding and applying natural principles to develop biolike systems composed of nanostructured macrosystems. In Europe these ideas have been distilled as a mandate for a recently formed independent organization called the International Nanobiological Testbed (INT). The INT conducts policy and specialized research on nanobiological concepts,[26] and as part of its defined mission has investigated a conceptual Mars biophysical station that incorporates the existing and projected developments in micro/nanosystems technology.

Japan also has made strong efforts in nanotechnology, with a probable immediate application in microtechnology. Although not directed toward space applications, Japan's research in this area will have use in future space systems. The Japanese Government has identified micromachine technology as a cornerstone technology for the 21st century. The technology does not ignore semiconductor-based processing technology, but emphasizes the miniaturization of

conventional manufacturing techniques like machining, grinding, and electroplating to fabricate micromachines from a wider variety of materials. As an example, the Toyota/Nippon Denso microcar is as tiny as a long-grain rice (7 mm). It is a replica (at 0.001 of the size) of the Toyota Motor Corporation's first automobile, the 1936 Model AA sedan. The minuscule vehicle has 24 parts, including tires, wheels, axles, headlights and taillights, and hubcaps that carry the company name inscribed in microscopic letters. The electromagnetic motor, which is itself made of five parts, is only 1 mm in diameter and can propel the car at speeds of up to 10 cm/s.[27] This micromachined automobile could not easily be fabricated completely of silicon, but with other material microfabrication techniques, it becomes very possible. For space systems, we see applications for micromachines, micromotors (e.g., 1–4-mm-flagella motors [28]), and microbots (e.g., micro-conveyance systems [29]). The Japanese group has designed a variant of the microcar for a small pipe-inspection machine. This application is also of interest to space systems. Examples of other application areas identified for micromachines are transportation safety systems, microsurgery, aircraft engine maintenance, and miniature information devices for appliances.[28]

The authors in collaboration with other scientists from Aerospace have reviewed a number of possibilities for the insertion of MEMS and ASIM components into space hardware. Following is our best estimate of MEMS and ASIM technology that will be inserted in the near term (less than 10 yr). Supporting details can be found in Janson[30] and Robinson.[31]

- Command and Control Systems
  - "MEMtronics" for ultraradiation hard and temperature-insensitive digital logic
  - On-chip thermal switches for latchup isolation and reset
- Inertial Guidance Systems
  - Microgyros (rate sensors)
  - Microaccelerometers
  - Micromirrors and microoptics for FOGs (fiber-optic gyros)
- Attitude determination and control systems
  - Micromachined sun and Earth sensors
  - Micromachined magnetometers
  - Microthrusters
- Power systems
  - MEMtronic blocking diodes
  - MEMtronic switches for active solar cell array reconfiguration
  - Microthermoelectric generators
- Propulsion systems
  - Micromachined pressure sensors
  - Micromachined chemical sensors (leak detection)
  - Arrays of single-shot thrusters ("digital propulsion")
  - Continuous microthrusters (cold gas, combustible solid, resistojet, and ion engine)
  - Pulsed microthrusters (charged droplet, water electrolysis, and pulsed plasma)
- Thermal control systems
  - Micro heat pipes
  - Microradiators
  - Thermal switches
- Communications and radar systems
  - Very high-bandwidth, low-power, low-resistance radio frequency (RF) switches

  - Micromirrors and micro-optics for laser communications
  - Micromechanical variable capacitors, inductors, and oscillators
- Space environment sensors
  - Micromachined magnetometers
  - Gravity-gradient monitors (nano-g accelerometers)
- Distributed semiautonomous sensors
  - Multiparameter-sensor ASIM with accelerometers and chemical sensors
- Interconnects and packaging
  - Interconnects and packaging designed for ease of reparability (e.g., active "Velcro")
  - Field programmable interconnect structures
  - "Smart" interconnects for positive-feedback

2.1.3 Initial Applications to Space Systems—First Steps

The promise of dramatic cost reductions in manufacturing and operating space systems, although appealing, is not a sufficient justification for the wholesale acceptance of MEMS and ASIM technology by the space-systems community. Cost is clearly a driving factor in today's economic environment, but reliability of space systems is the paramount concern and is especially true for both military and civil (i.e., NASA and ESA) space systems. Inserting microengineering technology into current systems can provide better monitoring of system status and health, which can help resolve potential operational anomalies and permit increased functionality with almost negligible weight or power impacts. The latter benefit can enable secondary missions and alternative operational modes to compensate for potential on-orbit failures in spacecraft systems. MEMS and ASIM technology also stands to improve performance and reliability during the other phases of a space-systems life-cycle, specifically, production and logistics (including long-term storage), launch-base facilities and logistic operations, specific prelaunch operations, launch and ascent flight, on-orbit operations involving ground segments, and decommissioning and/or deorbit.[31]

Many aspects of space systems operation actually occur on the ground. Insertion of MEMS and ASIMs into these phases (production, ground operations, logistics) should be somewhat easier as it does not require space-survivable designs. ASIMs that might be important in production and ground operation segments include multiparameter sensors integrated with data loggers and/or wireless (optical or RF) communications. These can be relatively low-bandwidth devices with peak-sensing capabilities to sense, for example, transportation or handling stress variables (e.g., pressure, temperature, humidity, shock, displacement stress, strain, and harmful chemicals) and to ensure that the maximum limits have not been exceeded during production, transportation, and storage operations. MEMS sensors for these parameters already exist and offer mass-production capability for inexpensive and unobtrusive environmental monitoring packs. Typical spacecraft costs range from $1 million for a microsatellite to well over $200 million for a one-of-a-kind geostationary communication satellite; knowing what, when, where, how, and by whom a limit was exceeded is serious business and of importance to setting insurance premiums and liabilities.

MEMS and ASIM technologies can also be used to instrument the launch vehicle. Current launch vehicles such as the Titan IV are often instrumented to measure the lift-off and ascent flight environments. However, these vehicles often have a limited number of channels (<100) to characterize both the dynamic acoustic and vibration environments. By proliferating ASIM units on the launch vehicle, a better characterization of the environment is possible. Similarly, there is a need to instrument the launch site. Ground-based measurement of rocket ignition overpressure and toxic chemical release (e.g., HCl from a solid booster) are needed in conjunction with the launch-vehicle monitoring system to dramatically increase the "awareness" of vehicle status and

the launch site environment. MEMS sensors (accelerometers, chemical sensors, etc.) coupled to data transceivers can be used in a wireless network system onboard the vehicle and on the launch site. The telemetry data can channel real-time or near-real-time information to a ground-based data storage system for postlaunch review.

An important role for ASIMs in both satellite and on-orbit manned vehicle operations is enabling a condition-based maintenance (CBM) status and health-monitoring system. These systems could save future costs by fault detection, isolation, and enabling automated self-test and repair actions. The CBM protocol enables safer operations as well as increased system availability compared with a failure-based maintenance protocol scheme. Reusable launch vehicles are prime candidates for CBM. One type of malfunction not uncommon in spacecraft is the faltering of a high-speed bearing, reaction wheel (momentum wheel), or gyro bearing. Bearing degradation can often be anticipated by monitoring vibration signatures, an excellent application for a micromachined accelerometer coupled to digital signal processor or microprocessor in an ASIM. Corrective action could consist of the metered release of lubricant via a fluidic ASIM. Smart bearings, smart structures, and multifunction structures are already being considered by space engineers. These ideas have also been considered in the aerospace community[32] for developing adaptive structures that have imbedded sensors—actuators, controllers and processors.

Within a few decades, thousands of low Earth-orbiting (LEO) satellites will be designed for global communications and general Earth-monitoring missions. Unlike their ancestor satellites, which were primarily used for mass-media communications and national missions, these satellite constellations will enable two-way communication for even simple pedestrian tasks, such as automatically measuring house utility meters; direct tracking of container ships, cargo, and small packages; and monitoring of natural resources and manufacturing facilities that affect the environment. These capabilities become possible with the current development of miniaturized low-power transceivers, which can be integrated with microsensors and data loggers. In the simplest configuration, the satellite constellation operates as a "store-and-forward" communication mailbox; as they orbit Earth, satellites query and gather data from many microtransmitters and forward them to a central ground station. With increasing sophistication, an orbiting satellite can beam down additional data or reprogram a ground unit altogether; for example, it could be sending your utility bill and reprogramming the amount of communication-bandwidth a particular resident address should have (e.g., for pagers, telephone, cellular, television, and other home digital services). A LEO satellite constellation could also provide automatic air-vehicle targeting[33] for special visual or other sensor reconnaissance (civilian applications might include quantifying factory emissions or following news events in near real-time, true global coverage).

Aerospace has been studying the use of micro untethered aerial vehicles (UAV) with a LEO satellite constellation system. The UAVs could provide a local "search and gather" capability that could be initiated from a remote location. For example, a micro UAV transmitting with 0.1 W of RF power using a low-gain monopole or patch antenna could send a 128- × 128-bit visible or IR image to a LEO satellite once per second. If each pixel contained 8 bits of intensity information, and a 15:1 data compression scheme were used, each image would require about 8740 bits. Assume that the following channels were added: 3 channels of 12-bit acceleration, 3 channels of 12-bit angular rate, 1 channel of 12-bit airspeed, and 10 channels of 8-bit health and status data (e.g., voltages, currents, temperatures), all at a rate of 10 Hz. If in addition to these channels, other information such as 264 bits of Global Positioning System (GPS) data (instantaneous true position and velocity) were added, as well as packet and protocol overhead, the total data rate would be about 9600 bits/s. This rate could be accommodated by a 15-kHz bandwidth at 1 GHz that could be received by a 1-m-diam antenna on a LEO micro/mini satellite at a range of 1200 km.

A less complex application using LEO satellites and fixed ground-based tags is the concept being developed by the National Semiconductor Corporation of Santa Clara, California, and Space Quest Ltd. of Fairfax, Virginia. The application is a vehicle tracking system using micro/nanotechnology satellites and wireless "tags."[34]

2.1.4 Spacecraft Orbits, Use, and Basic Design

Satellites are robots that collect and process energy and information. They exploit the strategic position of space to provide communications relay, Earth-observation, and space environment monitoring. LEO satellites orbit 300–2000 km above the Earth in roughly circular orbits that are typically highly inclined with respect to the plane of the equator. LEO orbits are used primarily by high-resolution meteorological and Earth observation satellites, store-and-forward communication satellites, and emerging personal communication satellites. Medium Earth orbit (MEO) satellites orbit 15,000–25,000 km above the Earth where they can "see" about 40% of the Earth's surface at any given instant. The U.S. GPS satellites, the Russian Glonass global positioning satellites, and the upcoming ICO communication satellites are MEO inhabitants. Geosynchronous Earth orbit (GEO) satellites orbit at an altitude of 35,786 km, where the orbit period matches the Earth's rotation rate. Geostationary satellites, which include almost all commercial communication satellites, orbit in the equatorial plane so that they appear to remain almost motionless in the sky; ground-based high-gain antennas for the uplink and downlink can then be fixed in place.

Satellites require a power supply, electronics, sensors, and in most cases, mechanical actuators. Figure 2.1 shows a block diagram of standard spacecraft functions. The white blocks represent functions required for any satellite, while the shaded blocks represent functions required for more advanced satellites. Spacecraft systems are traditionally constructed in individual housings and are electrically connected by a wiring harness. Note that in this packaging approach, the structure and thermal control systems can be separate entities, adding to the parts count, rather than an integrated unit. Within the context of Fig. 2.1, a simple satellite used as a communication relay or

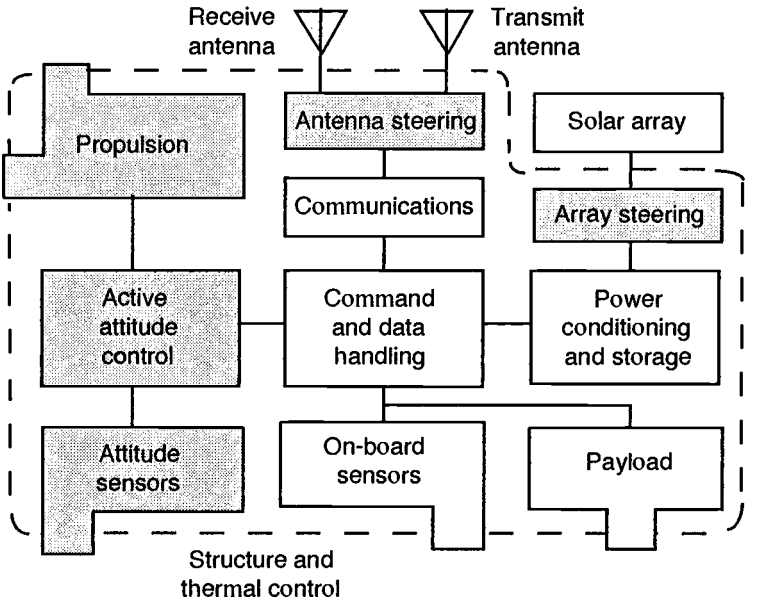


Fig. 2.1. Basic satellite functions.

as a space environment monitor could use fixed on-board antennas for communications and a fixed solar cell array, thus requiring no active attitude control. Examples of this basic spacecraft configuration include the 1960's vintage Courier,[35] the more recent Air Force MACSATs,[36] spacecraft by the University of Surrey, [37] and AMSAT microsats.[38] More advanced satellites require attitude sensing for solar array steering and active attitude control for sensor and antenna pointing. Examples of these spacecraft include GEO commercial and military communication satellites (e.g., Telstar-IV[39] and DSCS-III[40]), weather satellites (e.g., NOAA-14,[41] METEOSAT[42]), radar satellites (ERS-1),[43] and Earth-observation satellites (e.g., Landsat[44] and SPOT[45]). Payloads can be optical or IR imagers, RF imagers (radar), space science experiments, or communication systems while on-board sensors monitor temperature, voltage, current, etc. While not immediately obvious, MEMS can be used in all spacecraft systems shown in Fig. 2.1.

2.1.5 Getting to Orbit

Spacecraft are placed into orbit using a variety of launch vehicles that currently cost anywhere from $10 to $500 million per launch. Getting to LEO costs between $10,000 and $30,000 per kg; it is a function of launch vehicle, launch site location, orbit inclination, and orbit altitude. Putting a satellite into GEO, MEO, and LLO (low lunar orbit) costs about $50,000 per kg. The cost for putting a satellite into orbit around another planet ranges from about $60,000 per kg for Mars and Venus (with aerobraking) to over $300,000 per kg for Pluto (no planetary gravity assist, no aerobraking).

To get to orbit, the payload must survive various mechanical stresses produced by launch-vehicle acceleration, vibration, and shock from explosively driven stage-separation events. Table 2.1 lists worst-case values within the payload bay for selected launch systems. The Shuttle gives the mildest accelerations to accommodate human occupants; while the Pegasus, at least for this table, gives the highest. Both vehicles have wings that can generate substantial transverse accelerations, and the Shuttle has an additional transverse landing load that can reach 4.2 g.

Launch vehicles have very loud acoustic signatures. At lift-off, large vehicles, such as the Saturn-V, Space Shuttle, and the Titan IV, can generate sound levels up to about 200 dB on the ground (1 million times stronger than what causes pain to an average human, about 140 dB) with an acoustic power of about 10 MW. Payload fairings, located at the top of a launch vehicle, typically incorporate acoustic blankets to protect the payload. Table 2.1 lists the acoustic or sound pressure level in decibels in the payload bays of various launch vehicles. Maximum levels occur

Table 2.1. Worst-Case Payload Ascent Environment for Representative Launch Vehicles[a]

Launcher	Axial loads (g)	Transverse loads (g)	Acoustic level (dB)	Shock (g)
Pegasus	13	± 6	133.5	800 from 1000 to 10,000 Hz
Delta 7925	6	± 2.0	144.5	4100 at 1500 Hz
Atlas IIAS	6	± 2.0	138.4	2000 at 1500 Hz
Ariane AR44L	4.5	± 0.2	142	2000 from 1500 to 4000 Hz
U.S. Shuttle	3.2	± 2.5	140	5500 at 4000 Hz

[a]Data from Isakowitz.[46]

at launch and later during transition through maximum dynamic pressure (Max q). The corresponding vibration environment is also fairly high; power spectral densities between 0.01 and 0.1 g^2/Hz can occur over a 30–3000 Hz range. Finally, pyrotechnic actuators are typically used to separate stages, the payload fairing, and the satellite from the launch vehicle. These devices produce mechanical shocks with maximum magnitudes of 1000 to 10,000 g. Fortunately, these are transient events, and the highest accelerations occur between 1000 and 10,000 Hz. To guarantee that spacecraft will survive these abuses during launch, the spacecraft and major components are often ground-tested on "shaker tables" during the flight qualification stage. Microelectromechanical devices and ASIMs for spacecraft and boosters have to be designed not only to survive these short-term levels of abuse, but also to operate over the entire mission life (e.g., 7 yr).

2.1.6 Surviving on Orbit

Surviving on orbit requires attention to the packaging of both MEMS and electronics. While the human-occupied portions of the Space Shuttle, the Russian MIR space station, and the ISS are relatively benign, the outside space environment is much harsher; it significantly exceeds the design parameters for most terrestrial consumer/MEMS products. For example, local vacuum precludes the use of ambient convection cooling schemes; a modern fan-cooled microprocessor would quickly expire of "heat stroke." Spacecraft thermal management requires conductive heat transfer through circuit boards, structure, etc.; convective heat transfer through sealed "heat pipes"; and radiative heat transfer to and from the Earth, sun, and deep space. In general, internal spacecraft temperatures typically range from –10°C to +40°C, with the exception of RF power amplifiers like traveling wave tubes (TWTs) that can reach 70°C. External temperatures on the other hand (i.e., on a deployed solar array) can range from –50°C to +100°C.

Other drawbacks of operating in vacuum are outgasing of high-vapor-pressure materials, such as oils, plastics, and rubbers, and the lack of aerodynamic damping of moving components. Many MEMS devices such as accelerometers are designed to operate at atmospheric or reduced pressure. The surrounding gas provides mechanical damping and decreases the effective Q-factor (Q = frequency of resonance/bandwidth of resonance) to make simple electronic feedback control possible. Any MEMS device that requires gas-dynamic damping or includes high-vapor-pressure materials has to be hermetically sealed.

While the LEO environment is considered a "hard" vacuum by terrestrial standards, it is not a perfect vacuum. At altitudes between 200 and 650 km, the local pressure ranges from 10^{-6} to 10^{-10} mbar, and atomic oxygen is the dominant species. Figure 2.2 shows the atomic oxygen density as a function of altitude for "quiet" and "active" solar conditions. The ultraviolet output of the sun follows the 11-yr sunspot cycle, and during active times the rarefied upper atmosphere heats and expands further into space. Note that local O-atom densities differ by orders of magnitude between quiet and active solar conditions for altitudes above 400 km. At the present time, solar activity is increasing, and we expect to enter the next active phase by the year 2001.

Spacecraft in low Earth orbits are bombarded by atomic oxygen since the atmosphere is fixed with respect to the Earth. This results in aerodynamic drag forces, orbital decay, and surface erosion for orbiting structures. In the "ram" (along the velocity vector) direction, atomic oxygen impacts spacecraft surfaces with kinetic energies up to 5 eV. At these energies, chemical reactions with organic materials, composite structures, and metallic films are possible. This leads to surface modification and erosion, which can destroy micron-thick coatings and structures on the exterior of spacecraft. The most rapid erosion mechanism is surface oxidation into volatile byproducts. Kapton™, a DuPont Corporation polyimide that is used on many spacecraft, erodes at 3 μm per atomic impingement fluence of 10^{20} atoms/cm^2, which translates to an average erosion rate of

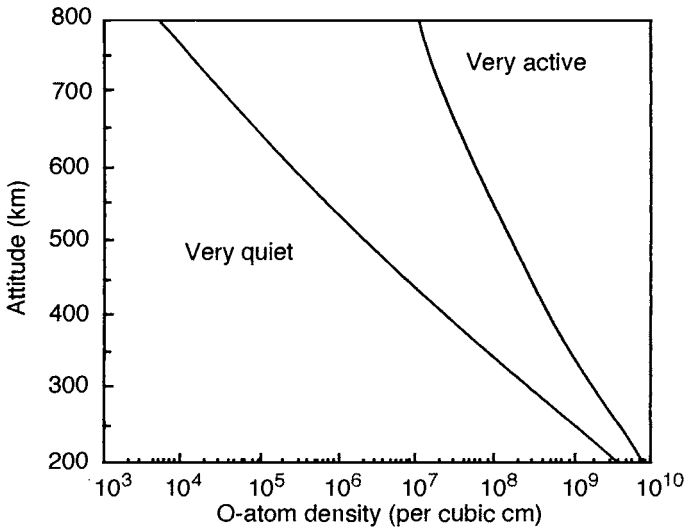


Fig. 2.2. Atomic oxygen density as a function of altitude under very quiet and very active solar conditions.

90 μm per year at a 400-km altitude for average solar activity. Figure 2.3 shows worst-case (surface normal always pointing in the flight direction) calculated erosion rates of Kapton as a function of altitude for very quiet and very active solar conditions.

Below 400–600-km altitude orbits, use of polyimide materials on exterior surfaces is not recommended, as a consequence of the high erosion rate. However, silicon dioxide (SiO_2), if deposited without surface defects or large internal stress, has an atomic oxygen erosion rate that is more than 3000 times lower than Kapton's.[47] For applications where surface charging may be a problem, somewhat thicker germanium or indium tin-oxide coatings can be applied. A 1-μm-thick layer of SiO_2 would last for at least 4 yr at an altitude of 400 km under active solar conditions. Coating thickness of at least 1 μm should be sufficiently durable for exterior and exposed

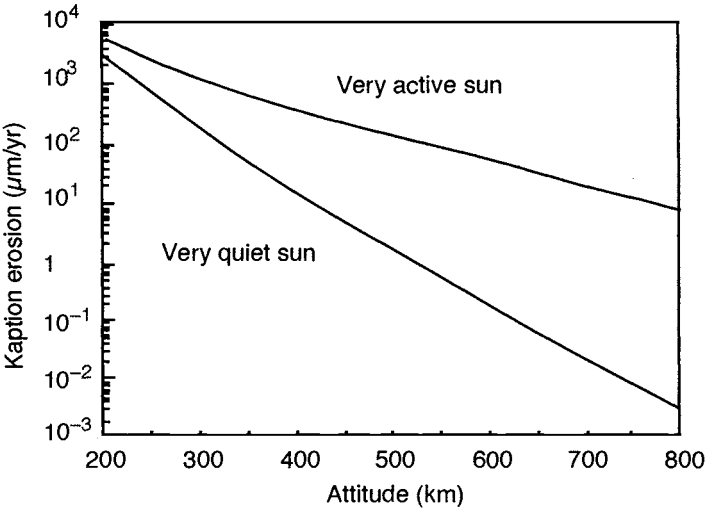


Fig. 2.3. Kapton erosion rate as a function of altitude under very quiet and very active solar conditions.

micromachined components (i.e., for microengineered active thermal surfaces, optical surfaces, and uncovered phased-array switches).

The "vacuum" of space has yet another important inhabitant: trapped ions and electrons that make up the Van Allen radiation belts. Damage in materials via defect formation can occur by the inelastic scattering of the high-energy particles. Defects in materials can form by atomic displacement, secondary particles, "showers" that create daughter products by fission, or by the ionizing tracks left in the wake of the particle pass-through. The latter type of damage can create abnormal charge concentrations. For MEMS, insulating surfaces can build up charge, which may upset electrostatic actuator and sensor operation. Similar effects on semiconductor circuits range from a temporary change in logic state, because of the sudden local appearance of charge, to permanent substrate atom and charge dislocations that produce altered current-voltage characteristics and possible device failure.[48] Single-event upsets (SEUs) are particle-induced "bit-flips"; while latch-ups are more serious high-current flow conditions generated by new low-resistance paths created by particle-induced ionization trails. Both SEUs and latch-ups can be controlled by appropriate choice of semiconductor technology, "watchdog" and error-correction circuits, and error-correction software. Continual accumulation of radiation damage, however, ultimately results in device failure.

Table 2.2, adapted from Griffin and French,[49] gives rough radiation hardness levels for different types of semiconductor devices. A rad is the amount of particle radiation that deposits 100 ergs of energy per gram of target material, and the radiation hardness level represents total dose required for device failure. In the spacecraft industry "total dose" is defined as the total dose absorbed and is therefore a material-dependent parameter. Typical low-power consumer electronic components, incorporating complementary metal oxide semiconductor (CMOS) technology, are designed to operate in our low-radiation biosphere (roughly 0.3 rad/yr) but can tolerate 1–10 kilorad integrated radiation doses. Unfortunately, the radiation tolerance varies widely from

Table 2.2. Radiation Hardness Levels for Semiconductor Devices

Technology	Total Dose in rads (silicon)
CMOS (soft)	$10^3 - 10^4$
CMOS (hardened)	$5 \times 10^4 - 10^6$
CMOS (silicon-on-sapphire: soft)	$10^3 - 10^4$
CMOS (silicon-on-sapphire: hardened)	$> 10^5$
ECL	10^7
I^2L	$10^5 - 4 \times 10^6$
Linear integrated circuits	$5 \times 10^3 - 10^7$
MNOS[a]	$10^3 - 10^5$
MNOS (hardened)	$5 \times 10^5 - 10^6$
NMOS	$7 \times 10^2 - 7 \times 10^3$
PMOS	$4 \times 10^3 - 10^5$
TTL/STTL	$> 10^6$

[a]Metal-nitride-oxide semiconductor.

design to design, so radiation testing should be performed on selected components. The particular semiconductor foundry process used also impacts radiation hardness; the 0.5-μm Hewlett-Packard CMOS process, as currently performed, can tolerate in excess of 100,000 rads.[50] Transistor-transistor logic (TTL) and emitter-coupled logic (ECL) circuits are inherently more radiation hard than CMOS, but they require more power. NMOS (n-type minority charge carrier MOS), PMOS (p-type minority charge carrier MOS), I^2L, and silicon-on-sapphire MOS circuits can be fabricated to be fully immune to latchup. CMOS circuitry fabricated onto silicon-on-sapphire substrates has traditionally provided radiation-tolerant electronics for space applications. The use of thin silicon over an insulator reduces the volume for charge collection along an ionizing particle track, thus reducing the amount of charge introduced into random gates. Thin-film silicon-on-insulator (TFSOI) technology is now being considered for commercial electronics because it can provide enhanced low-voltage operation, simplified circuit fabrication, and reduced circuit sizes relative to bulk silicon counterparts.[51] TFSOI would be particularly interesting for MEMS space applications because of its inherent radiation tolerance and its built-in etch stop for bulk silicon etching.

How much radiation shielding, that is, local packaging plus spacecraft structure, is required for a given mission? Dose rates for a silicon target are usually given as a function of grams/cm^2 or thickness of spherical aluminum shielding required for a given orbit and given solar conditions (i.e., for minimum or maximum solar activity). Figure 2.4 shows the yearly dose rate as a function of aluminum shielding thickness (full sphere shielding) for 700-km altitude orbits with orbit inclinations of 28.5 and 98.2 deg. CMOS circuits with an assumed total radiation dose tolerance of ~3000 rads will require at least 0.3 g/cm^2 aluminum (or 1.3 mm of silicon thickness) shielding for a 1-yr on-orbit lifetime in a 700-km, 28.5-deg inclination orbit. For the more interesting sun-synchronous (98.2-deg inclination) orbit, about 0.8 g/cm^2 (or 4-mm silicon thickness) is required for a 1-yr lifetime and about 3 g/cm^2 (1.3 cm silicon) for a 10-yr lifetime. At lower altitudes, significantly less shielding is required; while at higher altitudes, significantly more shielding may be required. Note that shielding effectiveness is not really linear with respect to thickness, especially

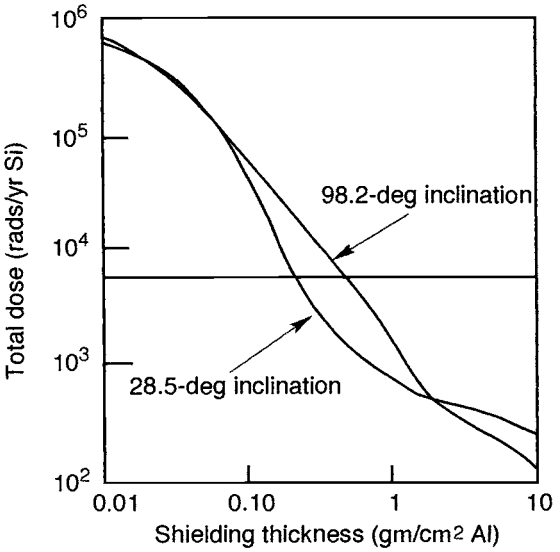


Fig. 2.4. Total yearly dose, under solar maximum conditions, in silicon as a function of aluminum shielding thickness for 700 km circular orbits.[52]

for low-inclination orbits. Use of more radiation-resistant technologies is the only solution for some orbits.

Figure 2.5, from Griffin and French,[49] shows the dose rate dependence as a function of circular equatorial orbit altitude inside spherical aluminum shields with densities of 0.5 g/cm^2 (0.18-cm-thick aluminum or 0.21-cm-thick silicon) and 3.0 g/cm^2 (1.1-cm-thick aluminum or 1.3-cm-thick silicon). Note the rapid rise in dose rate with altitude above about 2000 km and below 20,000 km; a hard-to-shield proton belt exists at ~4000 km and an easier-to-shield electron belt exists at ~20,000 km. At geostationary Earth orbit (GEO; 35,786-km altitude and 0-deg inclination) with a maximum dose of 3000 rads, 0.5 gm/cm^2 (0.22-cm silicon) and 3.0 gm/cm^2 (1.3-cm silicon) shielding give lifetimes of roughly 11 days and 3 yr, respectively. The real significance of Figs. 2.4 and 2.5 is that normal CMOS circuitry should be used only for low-altitude LEO missions; when designing smart MEMS and ASIMs for general space applications, hardened processes and designs must be used.

In addition to causing electronic upsets, on-orbit ions and electrons can also induce spacecraft charging of external surfaces. High-inclination orbits and high-altitude MEO and GEO orbits are particularly susceptible to this phenomenon. Without a slightly conducting path to spacecraft "ground," surface dielectric surfaces can charge up to kilovolt levels, resulting in a rapid local electrostatic discharge and potential device failure. Micron-scale MEMS structures probably will not tolerate this abuse; MEMS structures on exterior spacecraft surfaces should not be completely electrically isolated from their substrates. Resistive substrates and coatings should be used whenever possible. Additional information on launch system and space environment interactions with MEMS can be found in Muller *et al.*,[53] Barnes *et al.*,[54] and Stuckey.[55]

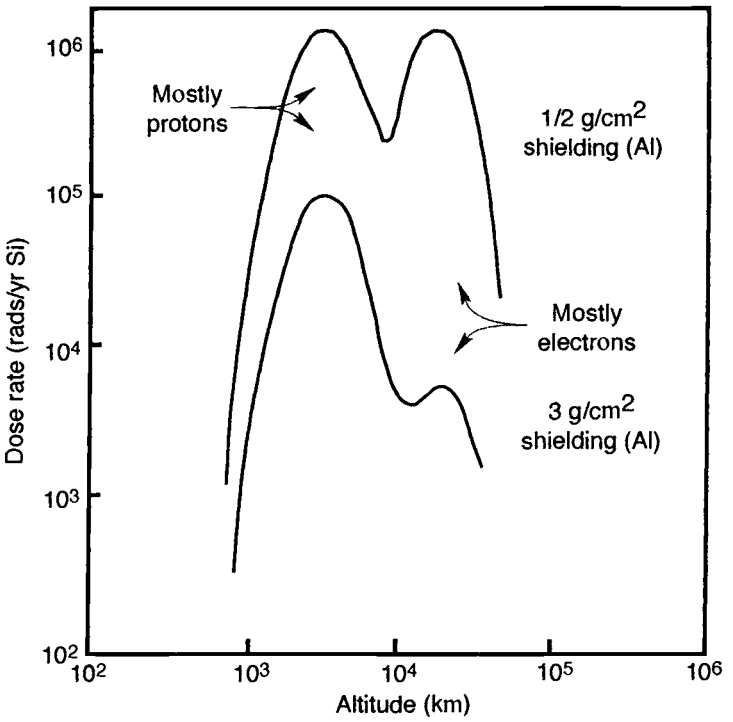


Fig. 2.5. Radiation environment for circular equatorial orbits.[49]

2.2 Spacecraft Applications

2.2.1 Electronic Systems

Electronics pervade almost all spacecraft systems. Individual electronic components can be classified as purely electronic or electromechanical. Purely electronic components (e.g., inductors, transistors, resistors) do not require any physical movement for proper operation; while electromechanical components (e.g., quartz crystal oscillators, relays, surface acoustic wave [SAW] filters, variable capacitors, and potentiometers) require translation, rotation, or vibration. Cofabrication of both purely electronic and electromechanical components on the same substrate is possible using a combination of MEMS and semiconductor fabrication techniques. This approach leads to a reduced parts count, volume, and the number of macroscopic electrical interconnects.

Resistors, capacitors, and inductors are typically considered passive components; whereas transistors and diodes are active components. MEMS changes the rules by allowing the fabrication of active capacitors, inductors, and resistors, and by offering micromachined switches and relays that could potentially compete with transistors in functionality for many applications. These "MEMtronic" devices are particularly interesting for space applications because they are inherently radiation-hard and could operate over a much wider temperature range (i.e., less than 50 K to more than 1500 K) than conventional circuitry.

Consider the active capacitor shown in Fig. 2.6.[56] This is a 190 × 190-μm-sq parallel plate capacitor with a nominal 1.5-μm variable air gap between the plates and a 300-fF capacitance. By applying a dc potential between the plates, the plates move closer together (normal to the page in Fig. 2.6), and a significant increase in capacitance results, that is, a 25% increase with a 4-VDC potential. This device can be used in an on-chip LC (inductor-capacitor) circuit to create variable frequency oscillators or filters. On-chip inductors will still need to be fabricated. Low-inductance spiral windings can be deposited on silicon substrates, but low-resistance silicon substrates produce capacitive loading. By fabricating the spiral inductor over an anisotropically etched pit, as shown in Fig. 2.7, much higher inductance and resonant frequencies are possible.[57] By exploiting MEMS cantilevered structures, variable-inductance coils are also possible.

Miniaturized communication systems may also use SAW devices for bandpass filters and time delay (or phase shift) generation. SAWs are usually discreet components composed of a piezoelectric substrate, quartz or lithium niobate, with patterned metallic electrodes acting as acoustic

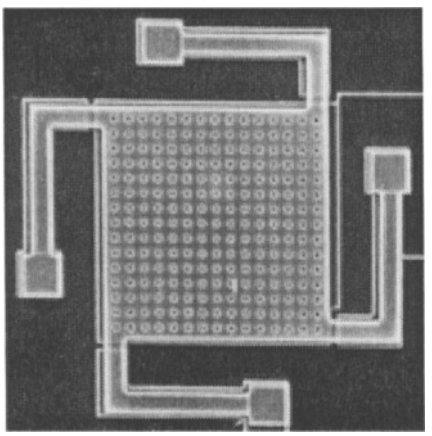


Fig. 2.6. Micromachined parallel plate capacitor with variable separation. (Courtesy B. Boser and D. Young.[56])

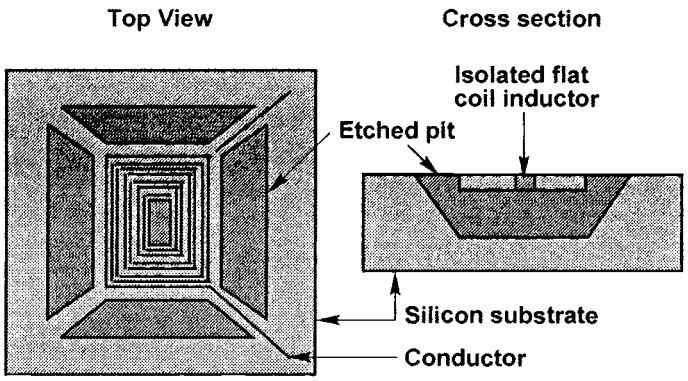


Fig. 2.7. Schematic design of low-capacitance coil for silicon substrates.

wave launchers, diffractors, and detectors.[58] The metallic patterns are created using photolithography, and a wide range of signal-processing functions can be accomplished by controlling electrode and surface geometry. A possibility for integrating SAW devices with silicon is to create an oxide layer on a silicon substrate, cover it with a piezoelectric layer such as zinc oxide, and top it with an appropriate metallization pattern.

Timing references are a key component of computer and communication systems. Quartz crystal oscillators consist of a specially cut crystal sandwiched between electrodes, which provides timing accuracy to better than 50 parts per million. An ultrastable quartz oscillator constructed by the Johns Hopkins Applied Physics Laboratory has a frequency stability of 7×10^{-14}, which is many orders-of-magnitude better than a digital watch oscillator. This device has a mass of 0.64 kg, a power requirement of 0.9 W, and a volume of 720 cm^3.[59] It may be possible to drastically reduce the volume, mass, and power requirements for precision oscillators using MEMS techniques to create single-crystal silicon resonators with mechanical isolation and micro heaters for precise temperature control. Cofabrication of the oscillator and electronics is also highly desirable to minimize the number of piece-parts and macroscopic interconnects. Designs for microheaters and hot plates compatible with CMOS processing can be found in Marshall *et al.*[60] Micromachined capacitively activated torsional resonators with Q factors greater than 500,000 in vacuum have also been demonstrated.[61] DARPA (Defense Advanced Research Projects Agency) currently sponsors a number of programs at the University of Michigan, Rockwell Science Center, and the California Institute of Technology[62] in the development of micromechanical filters and oscillators for communication systems operating at VHF, UHF, and S-band frequencies (100 MHz through 2500 MHz). Another option is to develop a miniaturized atomic clock that uses an atomic beam or static gas (for lower resolution units) in a cavity, an electromagnetic trap, and a scheme for excitation and sensing a resonance at a hyperfine splitting frequency (e.g., Mg^+ and Be^+ at 300Mhz, Hg^+, 40 GHz).[63]

Higher frequencies require smaller MEMS devices or nanoelectromechanical systems (NEMS). NEMS are MEMS with critical dimensions below 100 nm (0.1 μm). Current NEMS research uses specialized fabrication techniques to create submicron scale lengths in at least two dimensions. The semiconductor fabrication industry now fabricates devices with feature sizes down to 0.25 μm and is expected to break the 0.1 μm barrier in approximately the year 2007. Nanoelectronics will become commonplace, and mass-produced NEMS will become possible.

Microwave and millimeter-wave communication systems may use active antennas that integrate oscillators, amplifiers, or frequency conversion systems with microstrip antennas.[64-66]

The advantages of active antennas are reduced transmission line losses (which increase with frequency), and isolation of sensitive (low-noise preamplifier) or interference-generating (output amplifier) RF components from the digital electronics in the spacecraft. The individual radiators can be low-gain patch or micro stripline antennas, or they can use micromachined silicon horns or reflectors to boost gain.[67,68]

Phased-array systems take the active antenna concept one step further by using phase-controlled multiple transmit/receive antennas to produce and detect custom wave fronts.[69,70] This capability allows a fixed array of elements to simulate a single antenna of equivalent area with variable focusing characteristics; it allows electronic steering of a narrow beam, formation of multiple narrow beams, and controllable gain. Phased-array antennas add another degree of flexibility to communication systems by using fixed complex hardware under software control. Micromachined RF switches can be used to build true time delay lines and transmit/receive couplers for phased-array antennas.[71,72] In this application, micromachined switches can outperform transistor counterparts because of their inherent high bandwidth and low insertion loss.

MEMS switches can also replace transistors in digital circuits. The Air Force Institute of Technology (AFIT) has produced microrelays and microlatches for possible space applications.[73] Bimetallic (e.g., silicon and aluminum) thermal switches can be imbedded into electronic die to provide local overheating and latch-up protection. Figure 2.8 shows a schematic design of a simple MEMS switch designed at Northeastern University that is functionally equivalent to a field-effect transistor (FET) used in a digital mode; the gate potential determines if current can flow between the source and drain.[74] Figure 2.8 shows at the right a four-terminal microrelay version of this design, and Fig. 2.9 shows a scanning electron micrograph of the microswitch. Figure 2.10 shows a comparison between a FET-based dual-input NOT-AND (NAND) gate and a MEMtronic NAND gate based on the microswitch from Fig. 2.8. Note that they are almost identical. However, the same MEMtronic logic gate could operate deep within the Van Allen belts, on the surface of Venus, or within the icy fringes of our solar system. The disadvantages are the relatively slow speed (typically less than 100 MHz operation for MEMS, perhaps higher for NEMS), high operating voltages (tens to hundreds of volts using current technology), and increased surface area (the switches discussed in Zavracky, Majumder, and McGruer[74] are 30 × 65 μm in area). These disadvantages can be overcome by utilizing thinner structural layers and smaller device dimensions.

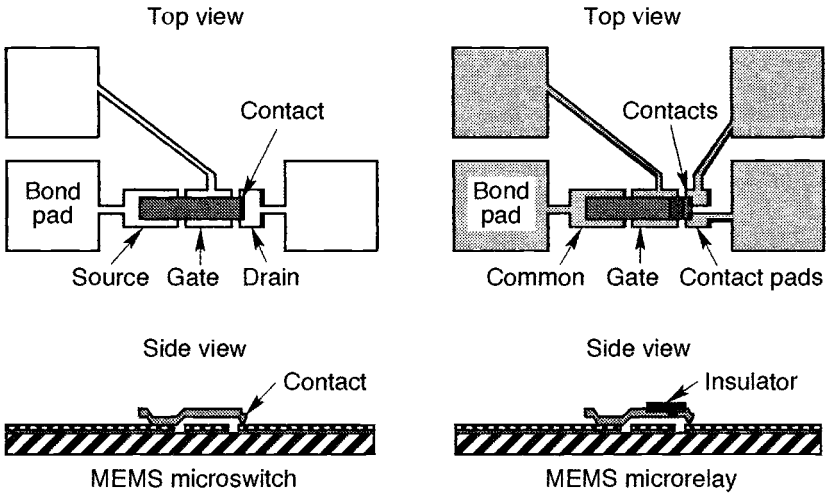


Fig. 2.8. Schematic of MEMtronic components.[74]

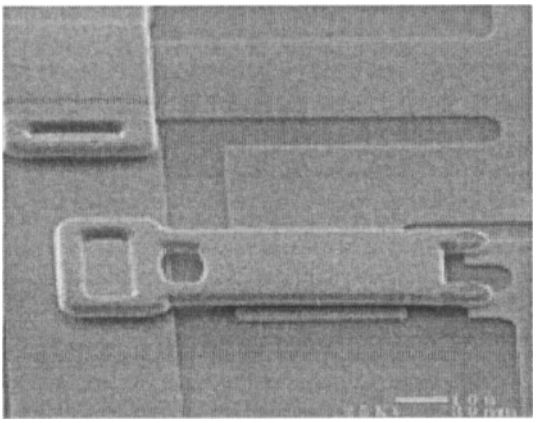


Fig. 2.9. Scanning electron micrograph of a MEMS microswitch. The "source" contact is on the left, the gate is in the middle, and the drain is under the two prongs. (Photo courtesy P. M. Zavracky.[75])

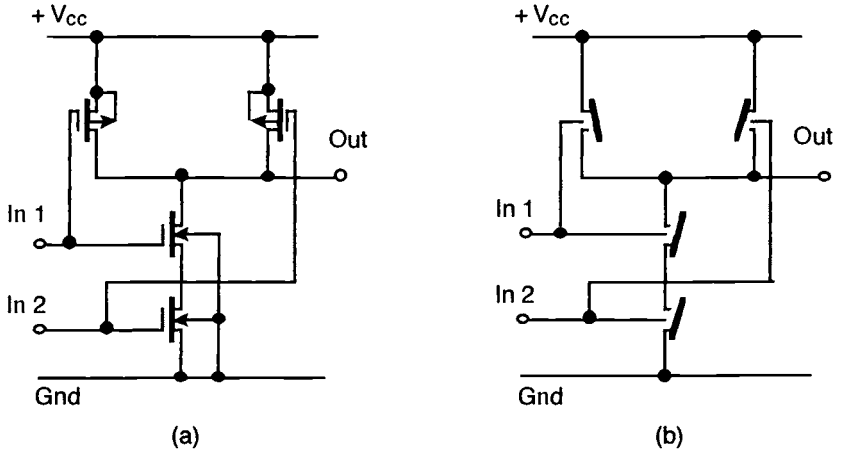


Fig. 2.10. Comparison of FET-based electronic and microswitch-based MEMtronic dual-input NAND gates.

2.2.2 Attitude Sensors

Spacecraft usually need to know their orientation in space to obtain maximum power from sunlight and to point high-gain communication antennas. Orientation can be determined by sighting against known references such as the sun, Earth, and stars; by measuring the local magnetic field vector; or by monitoring the phase shift in multiple antennas from different GPS satellite signals. Table 2.3 gives typical accuracy, mass, and power requirements for spacecraft attitude sensors. Optical sensors for locating the sun, Earth, and stars can have absolute accuracy much better than 0.1 deg and can operate from LEO to beyond GEO. Magnetic field sensors, on the other hand, work best in LEO and depend on a well-characterized magnetic field; above LEO they become more susceptible to transient magnetic events. GPS-based attitude determination is a promising technique that can provide absolute attitude and position determination. Once a "fix" has been established, on-board inertial navigation sensors can be used to estimate position and attitude at later times.

Table 2.3. Existing Attitude Sensors for Spacecraft[a]

Sensor	Accuracy (deg)	Mass (kg)	Power (W)
Sun sensors	0.005–3	0.05–2	0–3
Earth (horizon) sensors:			
Pulse generators	0.1–0.5	0.05–1	1
Passive scanners	0.5–3	1–10	0.5–14
Active scanners	0.05–0.25	3–8	7–11
Star Sensors	0.0003–0.1	1.5–10	1.5–20
Magnetic field sensors	0.5–5	0.6–2	0.5–2
GPS	0.1	2–10	15

[a]Data from Eterno *et al.*,[76] Pritchard and Sciulli,[77] and Johnson.[78]

Microoptoelectromechanical systems (MOEMS) can significantly decrease the mass, volume, and power requirements of optical navigation sensors, while MEMS could have a similar effect on inertial navigation sensors. A conceptual design for a single-chip, micromachined, single-axis sun sensor, designed by one of the authors, is given in Fig. 2.11.[30] The aperture is a slit 90 μm wide by 1.1 cm long, and the drive electronics are integrated with photodetectors. Photodetectors composed of n-doped regions in p-type silicon, or vice versa, are easily fabricated using

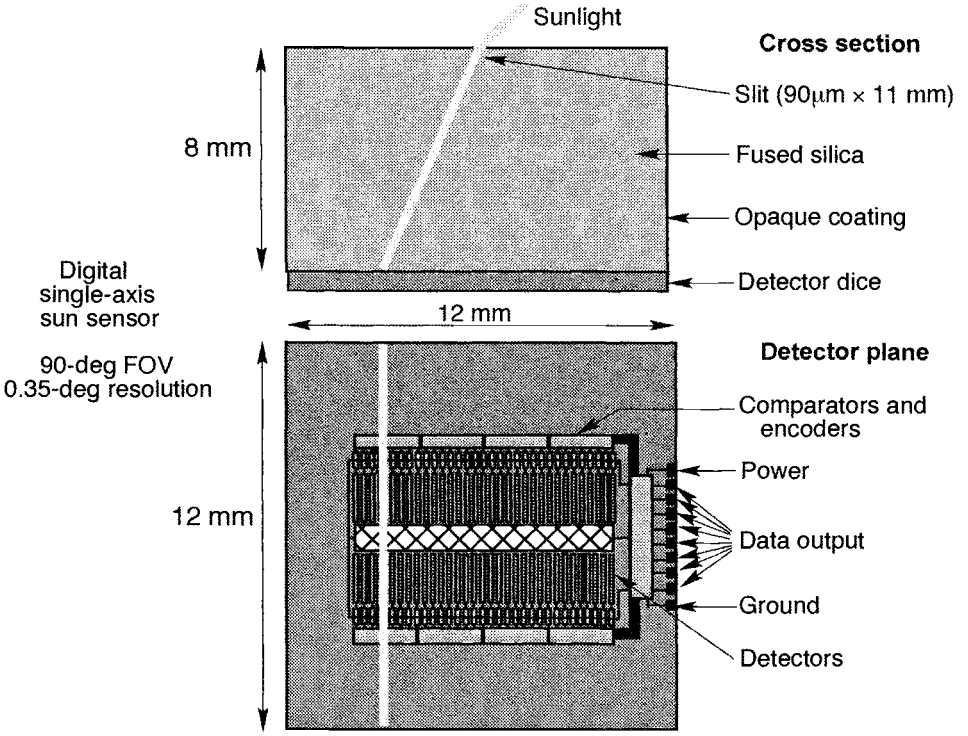


Fig. 2.11. Design of a micromachined sun sensor for a nonspinning satellite with a 90 deg field-of-view.

conventional CMOS processes. The large center detector provides a reference output against which the individual outputs of the 64 smaller interdigital detectors are compared. Coarse position (1.4-deg resolution) is determined in digital mode by locating which interdigital detector has the highest output; fine position (0.35-deg resolution) is determined in analog mode by ratioing the output powers from neighboring detectors. The fused silica provides radiation shielding for the detector electronics, and the opaque coating should be covered by a thin layer of aluminum, which oxidizes quickly and provides resistance to further atomic oxygen reactions. Estimated mass and power for a two-axis version are 5.5 g and 40 mW, respectively.

LEO spacecraft typically use flux-gate magnetometers to measure local magnetic field strength and direction. Flux-gate, magnetoresistive, and Hall-effect sensors are all suitable for developing microengineered magnetometers. The Honeywell HMC2003 is a three-axis magnetic sensor hybrid based on magnetoresistive transducers with a minimum detectable magnetic field of 100 μg and a range of ±2 g.[79] Nonvolatile Electronics, Inc., manufactures application-specific magnetic sensors based on the giant magnetoresistive ratio (GMR) effect, one of which has a ±10 g range.[80] Note that the Earth's magnetic field is less than 0.5 g in LEO. A novel magnetometer concept is being developed at Johns Hopkins University.[81] The operating principle of the magnetometer utilizes the Lorentz force to measure vector magnetic fields and is based on a classical resonating xylophone bar. The design is ideally suited for miniaturization, and the device has the potential for wide dynamic range and sensitivities down to applied fields of 1 nT. Figure 2.12 shows a scanning electron micrograph (SEM) of a polysilicon xylophone bar designed for capacitive pick-up. Although the device works, the high sheet resistance of the structural polysilicon layer limits the current-carrying capacity and sensitivity. An alternative material combination being considered is a metal/piezoelectric/metal (e.g., Pt/PZT/Pt) system.

Accelerometers and gyroscopes are key components of spacecraft inertial measurement units (IMUs). Spacecraft or launch vehicle accelerations can range from below 10^{-6} g to about 5000 g (the high levels are transient shocks), where g is the value of gravitational acceleration at the

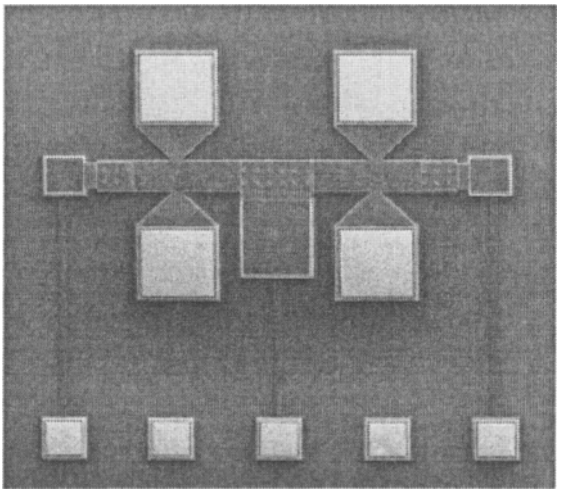


Fig. 2.12. Polysilicon xylophone magnetometer device fabricated by the MCNC MUMPS process. The poly0 layer is removed. The bar (poly1) dimensions are 1000 × 100 μm with the support legs 10 μm wide. The poly2 layer capacitive plates are placed at the ends and in the middle to enable differential capacitance measurements. (Photo courtesy D. K. Wickenden[81])

Earth's surface. Micromachined accelerometers monitor either the motion of a constrained proof mass or the force required to maintain an unconstrained proof mass at a fixed location within the instrument. The second approach usually provides higher bandwidth and accuracy. High sensitivity micromachined accelerometers such as the *Centre Suisse d'Electronique et de Microtechnique* (CSEM) ACSEM02-S and ACSEM02-T/6 force balancing sensors[82] or the silicon electron tunneling sensor built at NASA Jet Propulsion Laboratory (JPL) (sensitivity of 10^{-9} g/Hz$^{1/2}$)[83] offer micro-g and better sensitivity for on-orbit applications. This performance level requires temperature stability to within 1°C, which could be accomplished through integration of microheaters, silicon temperature sensors, and control electronics next to the sensing element. To survive launch loads and launch-related shock events, a safe "park" position may be required for the tunneling sensor. For launch vehicles and on-orbit propulsion monitoring, micromachined accelerometers in the range of 1 to 40 g can be used. A large number of such accelerometers are commercially available from a number of manufacturers, including Analog Devices, Kistler, Motorola, Silicon Designs, and EG&G IC Sensors.

Interestingly enough, within a three-axis-stabilized or rotating spacecraft, microaccelerometers with 10^{-7} g and better resolution can be used to determine spacecraft orientation by monitoring the gradient of the Earth's gravitational field. The radial component of the gravitational gradient, da/dr, is given by

$$\frac{da}{dr} = -2GMr^{-3} \tag{2.1}$$

where a is the local value of gravitational acceleration, r is the radial distance from the Earth's center, G is the gravitational constant, and M is the mass of the Earth ($GM = 3.98602 \times 10^{14}$ m^3/s^2). Values of $||da/dr||$ as a function of altitude above the Earth are given in Fig. 2.13. Note that the gravitational gradient is of the order of 10^{-6} m/s^2 per meter in LEO all the way down to the Earth's surface. Therefore, a 10^{-7} g resolution accelerometer could theoretically measure altitude to 1 m, if it was stationary with respect to the Earth's surface.

For the measuring of spacecraft orientation, consider an accelerometer mounted near the center-of-mass of a nonspinning spacecraft. The satellite is in free-fall, but the net local acceleration forces are zero, because of the balance between gravitational and orbit centrifugal forces. If the

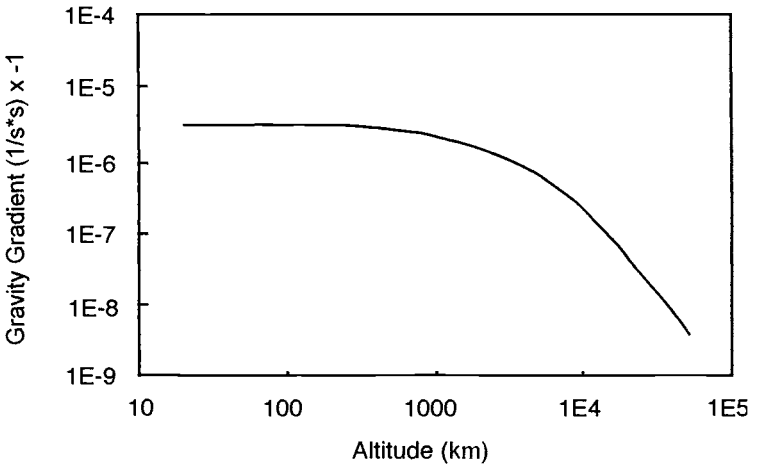


Fig. 2.13. The absolute value of the radial gravity gradient produced by the Earth as a function of altitude.

accelerometer is moved within the spacecraft radially outward from the Earth, the local gravitational acceleration is lower and the centrifugal acceleration is higher (larger orbit radius but higher velocity because the orbit period is the same), which results in a local tidal force directed away from the Earth. Similarly, a local tidal force is directed toward the Earth if the accelerometer is located closer to the Earth than the spacecraft center-of-mass. Figure 2.14 gives the radial tidal accelerations as a function of radial displacement from a spacecraft's center-of-mass for different orbit altitudes. If accelerometers could be produced with these sensitivities, determination of spacecraft orientation with respect to the Earth would be possible without using optical or RF (GPS) sensors. A gravity gradiometer with a sensitivity of 10^{-9} m/s^2/m was designed for the now-canceled ESA's ARISTOTELES mission.[84] This extraordinary sensitivity was to have been produced by 4 electrostatically controlled 320-g proof masses at the corners of a 1 m sq. If micromachined gravity-gradient sensors could approach this level of performance, batch-fabrication would allow proliferation of standardized attitude determination sensors across many LEO spacecraft series.

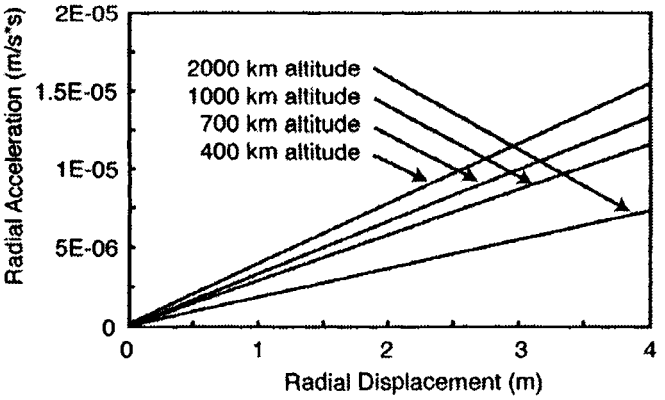


Fig. 2.14. Radial tidal accelerations as a function of radial separation from the center-of-mass of a spacecraft in circular orbit at different altitudes.

The established means of monitoring spacecraft attitude is to use rate gyros (gyroscopes) with optical sensors for absolute calibrations. Spacecraft gyroscopes are typically based on a rotating mass, a vibrating fork, or the continuous circulation of light around a closed loop (ring laser or fiber-optic gyros). Typical launch-vehicle or spacecraft-propulsion applications require drift rates of 0.1 deg/h or less, and typical spacecraft pointing requirements are at least an order of magnitude more demanding.[85] Micromachined gyros are based on "tuning forks" or vibrating structures that are excited in one plane and monitored for vibration at right angles to this plane. The Coriolis force, which is proportional to the angular rotation rate, generates these out-of-plane oscillations.

The Charles Stark Draper Laboratory has tested micromachined silicon gyroscopes with drift rates below 1 deg/h (at 0.1 Hz bandwidth).[86] Continuing research at Draper laboratories, JPL, and University of California, Berkeley, may drive drift rates down to 0.03 deg/h within a few years. If this performance cannot be obtained on a single gyro, perhaps applying signal-averaging techniques and a large number of gyros can reduce the drift rates. This approach is feasible if the drift is dominated by random factors. Drifts resulting from temperature changes are not random. Fiber-optic gyros (FOGs), which are replacing ring-laser gyros and spinning-mass gyros for many terrestrial applications, constitute the main competing technology for space applications. For example, Fibersense Technology offers a single-axis sensor with a 0.01 deg/h drift rate that consumes 5 W and weighs 10 oz.[87] The unit dimensions are 3.75 in. in diameter by 1.25 in. wide.

2.2.3 Propulsion

Propulsion is required for orbital maneuvering and can also be used for spacecraft attitude control. Once spacecraft attitude, position, and velocity are known, propulsion can be used for orbit raising, adjustment, and position maintenance. Currently, position and velocity are usually determined by ground station data and orbital mechanics. The range and the range-rate measurements are determined by radar or by relaying a known signal from a ground station, through the satellite's communications system, back to the ground station.

Propulsion requirements are expressed as a velocity increment (ΔV or delta-V) and the basic figures-of-merit for propulsion systems are thrust, minimum impulse bit, and specific impulse (I_{sp}), which is defined as the thrust divided by the mass-flow-rate of propellant through the thruster. If a time limit is imposed on a given mission, the minimum thrust can be determined from the ΔV, the mass of the spacecraft, and the thrusting time. Table 2.4 gives representative maneuvering missions, their associated ΔV, and the minimum thrust required in newtons per kilogram of spacecraft mass for two different mission times.

The mass of propellant to be expended is determined using the rocket equation:

$$\Delta V = g_o I_{sp} \ln\left(\frac{m_i}{m_f}\right) \tag{2.1}$$

where g_o is the gravitational acceleration at the Earth's surface (9.8 m/s^2), m_i is the initial spacecraft mass, and m_f is the final spacecraft mass ($m_i - m_f$ = propellant used). Figure 2.15 shows the propellant mass fraction (required propellant mass/initial spacecraft mass) as a function of specific impulse and ΔV. High specific impulse is desirable to minimize propellant mass or to maximize ΔV.

Table 2.4. Propulsion Requirements for Representative Missions

Mission	Time	ΔV (m/s)	Minimum Thrust (N/kg)
Increase altitude from 700 to 701 km	2 h	0.53	74.0
Increase altitude from 700 to 701 km	2 days	0.53	3.0
Move 10 km ahead at 700 km altitude	2 h	5.20	1100.0
Move 10 km ahead at 700 km altitude	2 days	0.04	8.3
Change inclination by 1 deg at 700 km altitude	2 h	131.00	40,000.0
Change inclination by 1 deg at 700 km altitude	2 days	131.00	2000.0
Change inclination by 1 deg at GEO	2 days	54.00	7500.0
Change inclination by 1 deg at GEO	2 days	54.00	750.0
Boost altitude by 100 km at GEO	1 day	3.65	42.0
Boost altitude by 100 km at GEO	1 week	3.65	6.0

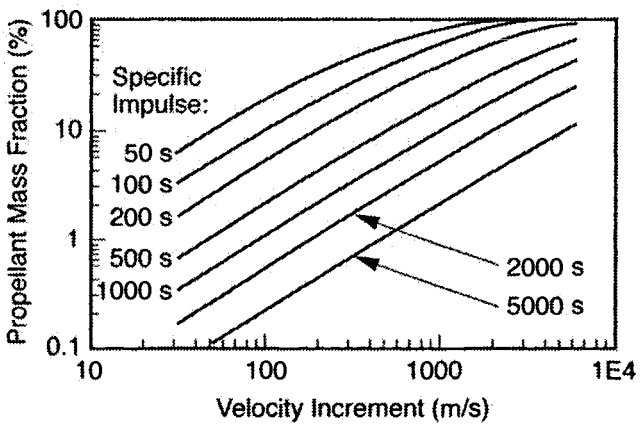


Fig. 2.15. Propellant mass fraction as a function of mission ΔV requirement for different I_{sp}.

How can micromachining techniques enable the building of propellant tanks, propellant lines, and valves? One solution is to bond several micromachined layers so that shallow surface cavities become tubes and deep cavities become propellant tanks. Figure 2.16 shows the basic concept in which three layers are bonded to form a propellant tank, associated plumbing, and two simple expansion nozzles. Multiple thrusters and propellant feed systems can be produced on the same substrate.

Micromachining offers new thruster design possibilities, which are presented in Chapter 17. As shown in Fig. 2.16, complete thruster systems need more than just thrusters; they also require propellant storage, propellant distribution, flow rate control, and health and status monitoring (temperature and pressure). MEMS nozzles and thrusters have already been demonstrated; yet to

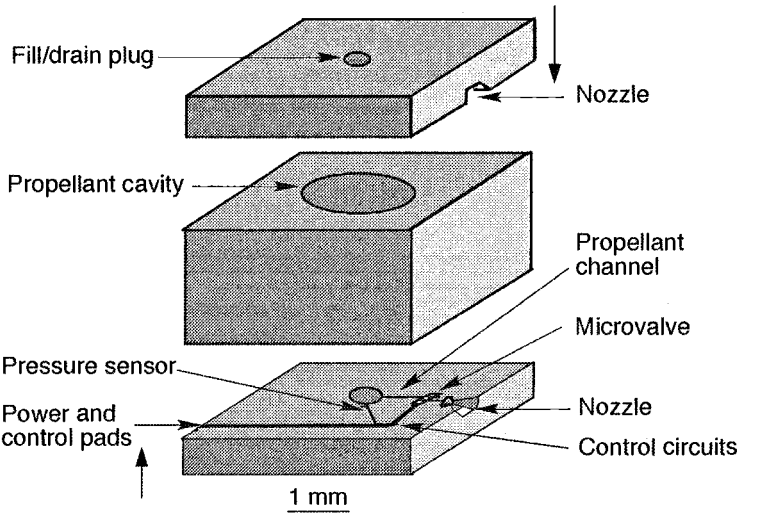


Fig. 2.16. Schematic assembly of a dual thruster micropropulsion system based on microfabrication techniques. The top and bottom wafers contain etched propellant channels with 100 to 1000 μm widths, multiple microfabricated valves, sensors and control electronics, and thrusters. The center wafer contains the propellant cavity (1mm to 1 cm diam) and may support additional microfabricated components.

be addressed are the leak rates of MEMS valves, the relatively slow response time of thermally actuated valves (typically 0.1–1 s), the operating pressure ranges of currently available valves (up to 100 psig), and the need for filtration of micron-scale particles within a propellant feed system.

MEMS valves typically use silicon-silicon or silicon-glass valve seats that do not have adequate seals for space applications; the leak rates can be 0.02 sccm or larger. At this rate, about 40 mg of propellant will be lost per day through each valve, or about 0.5 g per day through a 12-valve attitude control system. While this loss rate may be tolerable for spacecraft with mass greater than 50 kg, it would be intolerable for a 1 to 10-kg-class spacecraft that had to function for 5 years or longer. Elastomeric or "soft goods" seals, which are standard in macroscopic spacecraft valves, have just recently appeared in a MEMS valve produced by Redwood Microsystems.[88] There are other approaches to circumvent the traditional MEMS "leaky valve" problem. The JPL approach is to use resistive heaters to sublimate an otherwise low-vapor-pressure solid or liquid on demand.[89] Another approach, funded by DARPA and executed by TRW, Inc., Aerospace, and the California Institute of Technology, is to construct an array of single-shot microthrusters.[90,91] In its simplest form, this "digital" propulsion concept uses individually addressable sealed microcavities containing propellant, an internal heating resistor, and a micromachined silicon or silicon nitride burst disk as shown in Fig. 2.17. Each microcavity provides an impulse when the contained propellant is ignited and the gases exhausted. The diaphragm is designed to burst at a preset pressure, and for additional thrust the exhaust gases are made to flow through a converging/diverging nozzle. Preliminary "burst tests" have shown that a 0.5-μm-thick, roughly 500-μm-sq silicon-nitride diaphragm can be made to burst cleanly without clogging the flow channel. Polysilicon resistors can be placed directly on a thin oxide layer without regard to thermal loss because the firing time is so fast, on the order of 25 μs, that heat penetration into the oxide layer and substrate is minor. Micromachining enables the fabrication of thousands of similar microthrusters so that hundreds of complex propulsion maneuvers can be accomplished. Chapter 17 gives additional details of the digital propulsion system.

Micromachined pressure sensors can be integrated into conventional or micromachined propulsion systems once the materials compatibility issues have been addressed. Hydrazine (N_2H_4) is widely used as a space-storable propellant, but becomes an anisotropic etchant for silicon if water is present. The basic process involves formation of hydrated silica, which gets dissolved in the hydrazine/water mixture.[92,93] Water acts as a catalyst to generate OH^- ions, which can oxidize silicon. Monopropellant grade (MIL-P-2653C Amendment 2) hydrazine can contain up to 1% (by

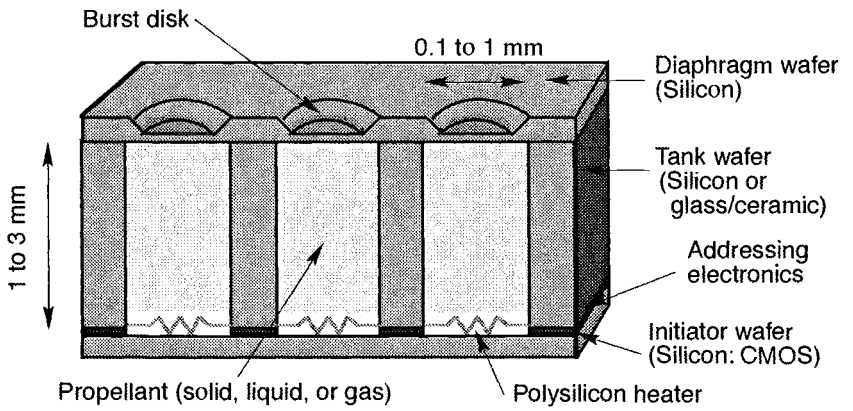


Fig. 2.17. Schematic cut-away view of a "digital" propulsion thrust system.

weight) water, which can be reduced to ~0.07% by passing it through an activated alumina column.[94] Our experience is that dry hydrazine will not etch bulk silicon with a native oxide layer. Materials compatibility testing of propellant-grade hydrazine with doped silicon, undoped silicon, and polysilicon at spacecraft temperatures is still needed.

2.2.4 Optical Systems

In current spacecraft, optical components are primarily used for imaging systems. These systems include Earth-imaging sensors and optical attitude determination sensors. MEMS and MOEMS will not replace macroscopic lenses and mirrors, but they could be used in controlling the image focal plane and to direct light beams for inter/intra satellite optical communications.

Near-term applications should include fiber-optic data buses, FOGs, and laser communication systems. MEMS and MOEMS can be used in all these applications. For example, light output from diode lasers and VCSELs (vertical cavity surface emitting laser) could be used more efficiently with "on-chip" optics for focusing (i.e., Fresnel lens) into a fiber or with micromachined scanning mirrors for beam steering (e.g., laser to fiber coupling module [95]). MEMS technology has successfully fabricated such components.[96] The devices are initially fabricated planar to the surface, but can be rotated out of the surface plane under microactuator control and locked into position.[97] Figure 2.18 shows an example of a micromachined beam steering system produced by the University of California, Berkeley.[98] The polysilicon reflector or mirror is near the bottom right of the photo and has been popped out of the plane of the silicon substrate. MEMS vibromotors, visible as flat structures with comblike features, control mirror orientation about a single axis (in plane of substrate) and mirror translation along a perpendicular axis (also in plane of substrate). The roughly 200-μm-sq mirror has an angular travel range of 90 deg, a translation range of 60 μm, and a maximum angular scan rate of 10.2 radians/s.

Variable gratings such as the vertical-motion "Grating Light Valve" phase grating by Silicon Light Machines[99] (vertical-motion phase grating) and the horizontal-motion grating device designed at AFIT, presented in Chapter 12, can be used to construct miniature programmable spectrometers for visible and infrared radiation. Simple versions could be used in Earth horizon sensors; while more complex imaging versions could be used for Earth observation, that is, cloud

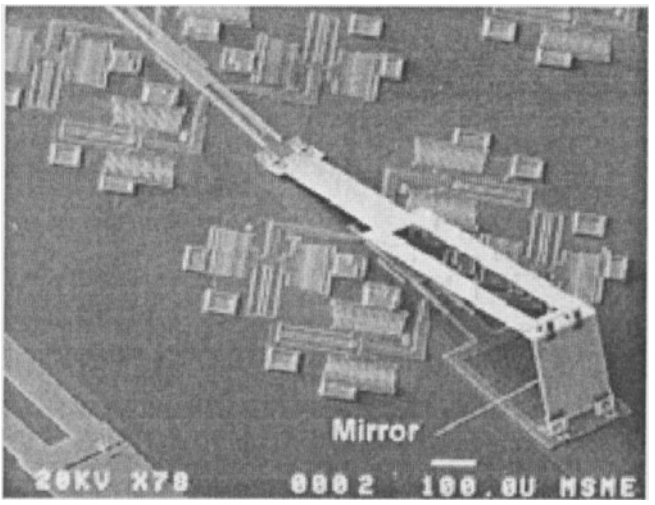


Fig. 2.18. Scanning electron micrograph of a microreflector with two degrees of freedom. (Photo courtesy R. S. Muller.[98])

cover, vegetation, and surface-temperature monitoring. Remote sensing from space-based platforms is becoming commercially viable.[100]

The growing demand for high-speed and high-density communication networks, both within and between spacecraft, and the potential impact that photonics technology may have in realizing that demand, will force both MEMS and MOEMS technology to be utilized on orbit. Examples of potentially useful devices include corner-cube microreflectors,[101] tunable optical filters,[102] and deformable mirrors for aberration control.[103] Figure 2.19 shows a hexagonal array of electrostatic-driven micromirrors, designed at AFIT, which is intended for aberration control in optical systems. The nonsegmented continuous-membrane array design of Boston University is also of great interest.[104]

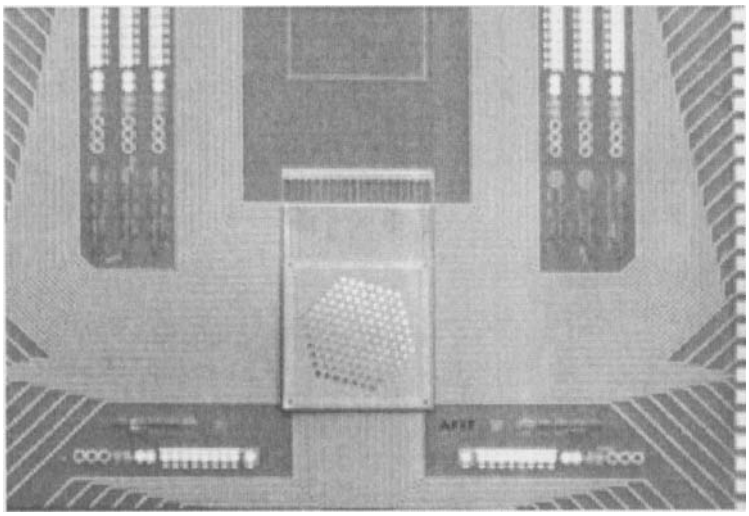


Fig. 2.19. Array of segmented micromirror piston-actuators designed at AFIT and fabricated at the North Carolina's MCNC MUMPs (Multi-User MEMS Processes). (Photo courtesy V. M. Bright.[103])

2.2.5 Thermal Control

The effective absorptivity/emissivity ratio, which determines the temperature of exposed spacecraft surfaces, can be modified using thermal louvers.[105] Many current spacecraft control heat rejection within a ~6:1 range by employing rectangular blade (venetian blind) and pinwheel designs driven by bimetallic springs. Figure 2.20 shows a conceptual micromachined implementation of the thermal louver concept based on the Texas Instruments Digital Micromirror Device.[106] The vanes and exposed silicon surfaces are coated with vapor-deposited aluminum to give a solar absorptivity of ~0.1 and an emissivity of ~0.05. When a vane is rotated out of the surface plane, a high-emissivity surface of either high or low absorptivity is exposed to the outside environment. Since silicon is transparent to infrared radiation between ~1.2 and 6.5 μm, elimination of the high emissivity coating would allow a warm object located below the silicon substrate to radiate to space while the vane was open. The hinge line is offset from the center-of-surface area to allow wide opening angles without requiring a large-gap height separation. The advantages of micromachined louvers are rapid response time and the ability to tailor the emitted thermal spectrum; the cavity can act as a high-pass (in frequency) filter if suitably designed.

Another approach to spacecraft thermal control is to use mechanical "thermal switches" that open or close a thermal conduction path between a heat source and a heat sink. An international

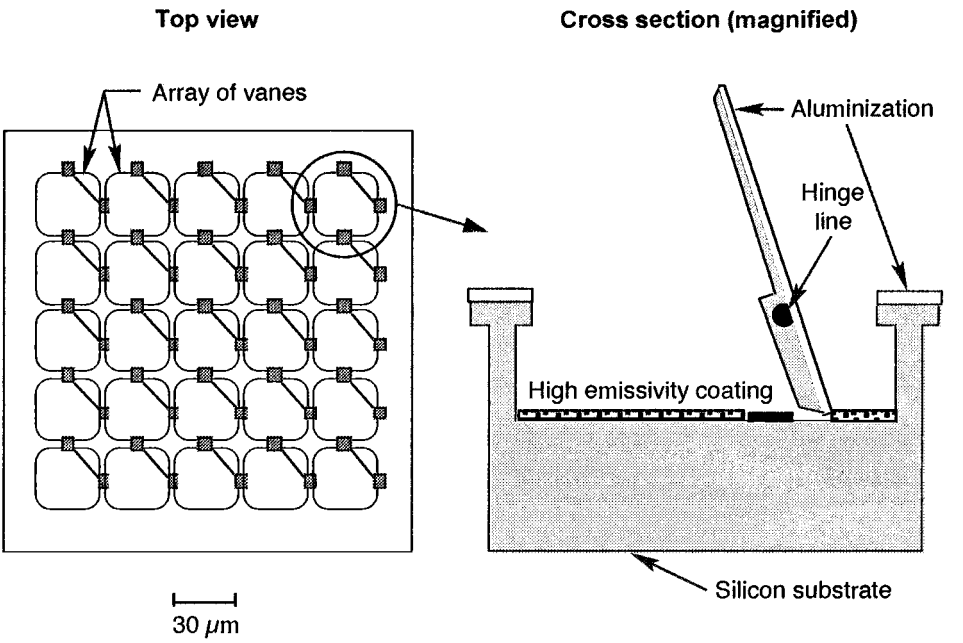


Fig. 2.20. Conceptual design of a micromachined thermal louver array for control of surface heat rejection capability.

team has demonstrated a micromachined active radiator tile (ART).[107] These 1-in.-sq (2.5 cm × 2.5 cm) prototypes are composed of two bulk-etched silicon wafers and use electrostatic attraction to pull a flexible upper diaphragm into thermal contact with the base wafer. The thermal gap between the diaphragm and base plate is 10–20 μm thick, and about 40 V is required to pull the diaphragm across that gap. A new design is under development that should withstand launch vibrations and accelerations.

Within the spacecraft, heat pipes are normally used to provide high thermal conductivity paths. Heat pipes are sealed tubes that transfer heat from one location to another, using vaporization of a working liquid at the "hot" end followed by convective transport of the vapor and condensation at the "cold" end. The condensed liquid returns to the hot end via a wicking or surface tension process. Miniature heat pipes have hydraulic diameters on the order of 1 mm; while micro heat pipes have diameters on the order of 10 μm. Additional information on miniature and micro heat pipes can be found in Cao *et al.*[108] and Khrustalev and Faghri.[109] The miniaturization of heat pipe technology using MEMS fabrication techniques allows heat dissipation to be enhanced over small distances for individual integrated circuits, detectors, or actuators. Micromachined heat pipes have been investigated by a number of researchers with some promising results.[110,111] Fabrication is relatively straightforward using a (100) silicon wafer. A long, thin exposed region of silicon can be anisotropically etched to produce a "V" groove, which becomes a sealed tube when bonded against a flat surface. Methanol has been used as the working fluid. Figure 2.21 gives the dimensions and geometry used.[110,111] The results show an increase in effective thermal conductivity of up to 81%, compared with a standard silicon wafer, and a significantly improved transient thermal response. Micromachined heat pumps may provide an effective way of removing heat from integrated circuits without using metallic heat radiator elements.

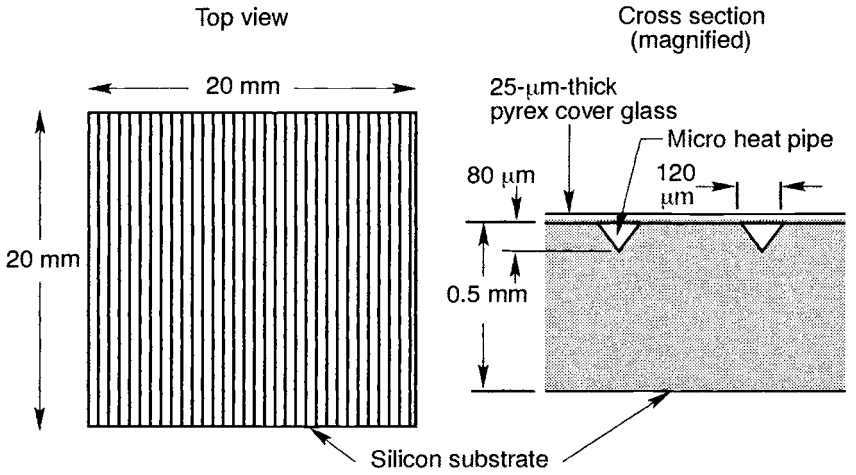


Fig. 2.21. Micro heat pipe construction.

2.3 Silicon Satellites

2.3.1 Basic Concept

Figure 2.22 shows a rendering of an Earth-observation, silicon satellite (also known as the nanosatellite) to be used in LEO. Introduced in Janson, Helvajian, and Robinson,[112] this concept presents a new paradigm for space system design, construction, testing, architecture, and deployment. Integrated spacecraft that are capable of attitude and orbit control for complex space missions can be designed for mass-production using adaptations of semiconductor batch-fabrication techniques. Integrated circuits for command and data handling (C&DH), communications, power conversion and control, on-board sensors, attitude sensors, and attitude control devices can be manufactured on 1 to 4-mm-thick silicon substrates that simultaneously provide structure, radiation shielding, and thermal control. Silicon compares favorably with aluminum in terms of thermal conductivity, radiation-shielding ability, and mass density, yet it is stronger than steel (~7 GPa maximum stress vs ~1 GPa for steel) and transparent to IR radiation between 1.2- and 6.5-μm and also between 25- and 100-μm wavelengths. Diamond is better on almost all counts, but silicon is readily available and easily processed. Silicon's main weakness is its brittleness; impact and shock loading must be controlled during fabrication, assembly, testing, and launch. Batteries and solar cells for the nanosatellite will still need to be fabricated using conventional materials. The spacecraft shown in Fig. 2.22 is essentially a stacked multiwafer package. A multichip module approach combined with partial wafer-scale integration would be used to fabricate the wafers. Useful silicon satellites will have dimensions of 10 to 30 cm; while more complex configurations using additional nonsilicon mechanical structure (i.e., truss beams, honeycomb panels, and inflatable structures) will be much larger. The benefits of batch-fabricated silicon satellites are:

- Radically increased functionality per unit mass
- Ability to produce 10,000 or more units for "throw-away" and dispersed satellite missions
- Decreased material variability and increased reliability because of rigid process control
- Rapid prototype production capability using electronic circuit, sensor, and MEMS design libraries with existing (and future) computer-assisted design (CAD)/CAM tools and semiconductor foundries
- Reduced number of piece-parts
- Ability to tailor designs in CAD/CAM to fabricate mission-specific units

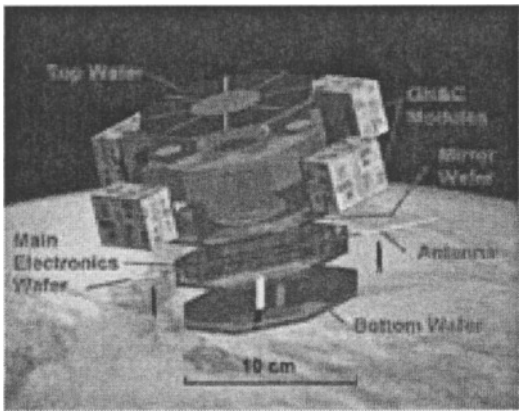


Fig. 2.22. Rendering of a hypothetical Earth observation silicon satellite.

Initial nanosatellite designs use two types of processed wafers: wafers that incorporate a sparse number of electronic devices (i.e., low interconnect density, as opposed to memory or microprocessor fabrication wafers), numerous micro channels, plus MEMS and MOEMS; and wafers that are essentially multichip modules (MCMs) that contain most of the centralized signal processing, command, and control electronics and the RF communications. These MCM wafers will need a high density of interconnects and the capability for mixed-signal processing. In the near-term nanosatellite designs, communication between wafers will be via metal and polysilicon lines routed to the wafer edges with a perimeter connection system that constitutes the satellite bus. In long-term nanosatellite designs, communication between wafers is more likely to be via local RF or free-space, optoelectronic switching technology, for example, use of heterojunction phototransistors (HPTs) integrated with vertical cavity surface-emitting lasers (VCSELs).[113]

2.3.2 Size Impacts: Feasibility of a Pico-Femto Satellite

Silicon satellites can be classified as microsatellites (1–100 kg mass), nanosatellites (1 g–1 kg mass), picosatellites (1 mg–1 g mass), or femtosatellites (1 μg–1 mg mass). While picosatellites and femtosatellites would have seemed absurd 10 years ago, they are now conceptually feasible because of continuing decreases in electronic gate size and the emergence of MEMS and MOEMS. These technologies permit the integration of the C&DH and communication systems, low-resolution attitude sensors, inertial navigation sensors, and a propulsion system into a 1-cm-cube or smaller size satellite. On the other hand, by removing propulsion, for example, picosatellites and femtosatellites would be ideal as simple space environment sensors. Using only solar radiation and depending on the overall configuration, picosatellites through microsatellites can produce power levels in the 1–100 W range. On the other hand, femtosatellites can only generate microwatts to milliwatts. This directly affects how much power is available for power-hungry communication and data-processing systems. Thermal control is also an issue for these lilliputian satellites. Simple lumped-parameter models of silicon satellite temperature swings between fully lit and Earth-eclipsed conditions have shown that passive thermal control is possible for nearly spherical nanosatellites and microsatellites.[114] When dimensions drop below 2 cm, the temperature extremes exceed typical electronics and battery limits. Femtosatellites, with their extremely low mass, can reach the equilibrium sunlight (or eclipse) temperature within minutes. As a consequence, picosatellites and femtosatellites will require some form of thermal control.

Small size also affects radiation shielding ability and orbit lifetimes. For constant altitude and spacecraft density, the ratio of air drag to spacecraft mass is inversely proportional to scale length.

As spacecraft shrink in size, the deceleration due to the air-drag becomes stronger, resulting in more rapid orbital decay. At altitudes below 500 km where radiation shielding (0.38 mm maximum length for a 1-mg mass cubic femtosatellite) may be adequate for radiation-tolerant electronics (about 10^4 rads total dose with error detection and correction), the orbit lifetime is only a few days. At higher altitudes, rapidly increasing radiation levels limit the lifetime to a few days unless special radiation-hard (e.g., silicon-on-sapphire) electronics are used. Femtosatellites should be nearly spherical in shape to minimize air drag and maximize radiation shielding. Maximum power generation levels will therefore be in the submilliwatt range. Femtosatellites are an extremely difficult challenge because of their low thermal mass and wild temperature swings as they enter and exit Earth's shadow.

Picosatellites are the smallest useful satellites, but active thermal control will be required. A thermally passive picosatellite will have temperature swings of 90 K between sunlight and eclipse in low Earth orbit. Cubic picosatellites made of silicon can have as much as 0.18 cm radiation shielding and orbit lifetimes of several years at 700-km altitude under solar-maximum conditions. Nearly spherical satellites are needed again to provide radiation shielding, and if low-inclination orbits are used (below 700-km altitude), use of radiation-soft CMOS electronics may even be feasible. Available power will be in the tens of milliwatts range. Picosatellites may be good for disposable or short-duration (i.e., 1-week) missions (i.e., as space probes).

2.3.3 Missions

Silicon nanosatellites and microsatellites with 10 cm and larger dimensions, micromachined attitude sensors, and micropropulsion for attitude and orbit control could perform useful missions with on-orbit lifetimes of 1 to 5 years. Possible mission applications are communication relay, cloud cover monitoring, geolocation, and space environment monitoring. Mission applications can be grouped into three broad categories:

- Disposable missions that use silicon satellites for a short period of time followed by deorbit
- Global coverage missions that use hundreds of silicon satellites in LEO to provide continuous Earth coverage for communications or Earth observation
- Local cluster missions that utilize hundreds of silicon satellites in a sparse array configuration to provide a large effective aperture

An example of a "disposable" mission is the untethered flying observer (UFO) that was analyzed during the workshop portion of the First International Conference on Integrated Micro/Nanotechnology for Space Applications.[115] A UFO, shown in Fig. 2.23, could be deployed on command and flown about the host vehicle to provide a visual assessment of the larger spacecraft health and physical attributes, for example, after an operational anomaly is detected. The UFO would be mounted on the surface of a LEO spacecraft in a "cocoon" and would lie dormant until activation. The workshop effort produced a conceptual design with a mass less than 1 kg, a maximum power level of 1.6 W, and an operational lifetime of 48 h. A lithium primary battery supplies power, a 2000 × 2000 pixel CCD imager provides images and high-resolution attitude information, and ammonia cold gas microthrusters provide maneuvering and attitude control. Image and telemetry data would be transmitted at S-band using omnidirectional antennas (the stubs protruding from the UFO in Fig. 2.23) to Space Ground Link Subsystem (SGLS) stations in the Air Force Satellite Control Network (AFSCN) using 0.5 W of RF power. Following the mission, the UFO does not return to the mother ship but is deorbited.

Silicon satellites can also be dispersed as local clusters. One approach, analyzed by researchers at MIT, is to use random clusters in which individual nanosatellites move with respect to each other. The Aerospace approach is to utilize orbital mechanics to create configurations that

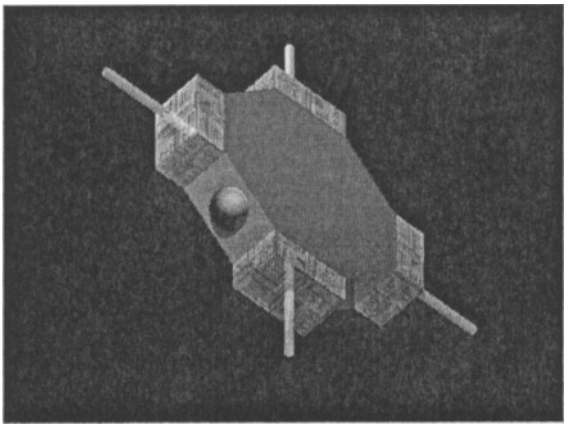


Fig. 2.23. Artist's concept of an untethered flying observer. The main body is 10 cm in diameter.

maintain a fixed geometry without requiring continuous thrusting.[116] A circular ring of physically unconnected satellites will maintain its geometric configuration, to first order, if the ring diameter is orders of magnitude smaller than the orbit radius, the ring itself is in a circular orbit, and the surface-normal vector of the ring points 30 deg away from the nadir (toward the Earth's center) in the orbit-normal direction. The major orbit perturbations to these clusters result from so-called "J2" effects (the Earth's mass distribution cannot be adequately represented by a point mass because the Earth is slightly flattened because of rotation and "J2" represents the gravitational perturbation that results from this flattening), which decrease rapidly with altitude. Once established, this ring of satellites will rotate once per orbit period as seen in the reference frame of the orbiting cluster; for example, the ring rotates about its center while the whole cluster rotates about the Earth. Since the ring rotation rate is independent of radius, multiple concentric rings of different diameters will rotate together, thus producing a rotating disk of fixed geometry. These local clusters, composed of hundreds to thousands of individual silicon spacecraft, can operate in a concerted fashion as a single large phased array at radio frequencies. Each cluster would operate as a local area network with short-range optical or RF communication links between the nanosatellites and a central mother ship.

Silicon satellites have not yet been built, but they offer radically new ways to perform space missions. MEMS and MOEMS advanced microelectronic processing and packaging make them possible. More development effort is required for miniaturizing space systems, in particular micromachined gyros, micropropulsion, and micromachined laser communications. MEMS sensors for on-board health and status monitoring are also needed, fulfilling similar tasks as that required in larger satellites.

2.4 Manufacturing Future Space Systems

2.4.1 Manufacturing Challenges and Limits of MEMS/MOEMS Technology Insertion

MEMS and MOEMS technology will inevitably be used in future space systems. Investigations are already under way to reduce spacecraft size and weight and to modularize the subsystems to enable new technology insertion in future block changes of existing satellite programs. Studies that look further into the future than the next block change already know that space is a strategically lucrative platform for conducting business–both civilian and military. Pragmatically, some of these missions can only be accomplished by orbiting a large constellation of satellites. The

nanosatellite concept is one solution to meet this challenge. The infrastructure necessary to assemble a mass producible nanosatellite does not yet exist within the space community; however, elements of this required infrastructure do exist in the commercial world, for example, in the manufacturing of laptop computers, personal information systems, cellular phones, and hand-held video cameras. Regardless of how the necessary manufacturing infrastructure is mobilized, future satellite systems will be designed to process more data on board, to operate more autonomously, and to be manufactured by automated assembly-line processes as opposed to the current piecemeal building approaches used. In addition, to reduce the cost of building satellites, statistical quality-control methods must be implemented as a requirement for achieving overall high-quality systems.[117] These criteria alone provide an avenue for technologies like MEMS and MOEMS to be inserted into space systems, either as monitoring instruments (e.g., satellite manufacturing process line, onboard satellite health and welfare systems management) or to provide new and enhanced capabilities. The extent to which MEMS and MOEMS can be inserted into future satellite designs will depend on how rapidly microengineering prototyping centers can be established and how rapidly microdevices can be fabricated on materials not within the conventional microelectronics industry repertoire. The latter requirement arises because besides semiconductors, materials such as ceramics, glasses, diamond, polymers, and composite materials are typically used in space systems. The fabrication of microdevices and complete ASIMs on these materials is crucial to satellite design approaches for a fully integrated system. The alternative is to implement a macro-scale package for each individual micro device, which negates the desire to reduce excessive packaging. In reality, if new satellite design paradigms are implemented, the most likely path space system engineers will follow is to design for full integration but incorporate nonintegrated components as add-on systems and only if there are compelling benefits to satellite operations.

2.4.2 Need for Rapid Prototyping Centers

The success of micro/nanotechnology to revolutionize our world will, in general, depend on the development of effective rapid prototyping centers and the networking of these prototyping centers to enable users to draft process sequences that can be cycled through physically separated sites. For example in the United States, the Multi-User MEMS Processing Service (MUMPS), the Metal Oxide Semiconductor Implementation Service (MOSIS), other "virtual" foundries, and most university and industry research centers offer an excellent path to accelerated component prototyping, and the recent DARPA-initiated MEMS-Exchange program[118] could establish the environment for distributed MEMS fabrication and manufacturing. In most fabrication centers the tools and fabrication processes are geared for semiconductor materials processing. This fact will certainly influence the design of many terrestrial and space instruments such that wherever feasible, components, devices, and complete subsystems will be designed to leverage the use of existing microelectronics technologies.

For space applications, however, the use of materials other than semiconductors can be advantageous. Combustion chambers must withstand high temperatures and possible chemical attack. Silicon may work for hydrazine monopropellant microthrusters, but bipropellant thrusters have combustion temperatures far in excess of silicon's melting temperature. High thermal conductivity and electrically insulating materials (e.g., diamond) should be used around high-power circuits while polymers or other ductile materials are preferred in valve seats to limit leakage. As a result, processing tools and techniques that can efficiently micromachine/process nonsemiconductor materials may become necessary. The laser is one example of such a processing tool.[119] Laser material-processing technology has experienced a robust growth in the past decade. This is mostly because the reliability of laser systems has increased, higher repetition rate lasers are now

commercially available (e.g., kilohertz to megahertz), and a variety of wavelength (e.g., vacuum UV—far IR) and pulse-width (e.g., femtoseconds to continuous wave) choices are on the market. As a material-processing tool, the laser is a nonintrusive, in-situ material-processing tool, which in principle can remove material, deposit material, and anneal the surface. It can serve as a diagnostic of the surface quality, morphology, surface adsorbates, and gas phase reactant; and it can "micromachine" structures on the surface or imbedded in the bulk. Lasers can process not only silicon but also other semiconductors, ceramics, metals, glasses and composite materials. However, unlike semiconductor processing in which the processing tools are automated and the process recipes refined over the past three decades, laser-based tools are just becoming commercially available with comparable automation capability and process control.[120]

2.4.3 Need for Mixed Technology Integration and CAD

Rapid-prototyping centers alone will not advance MEMS technology. To adequately capitalize on investments in the manufacture, design, and test of semiconductor integrated circuits toward future terrestrial and aerospace applications, computer-assisted manufacturing programs for mixed technology systems must be developed. Mixed-technology systems are defined to include mixed-energy domains (electronic, kinematic, optical, fluidic, and electromagnetic domains, etc.) and mixed-signal integrated microsystems. Mixed-technology systems may incorporate hundreds to thousands of integrated microdevices that create new system capabilities unachievable through more traditional hybrid integration. Their design represents unique challenges and opportunities. The tight integration of microdevices in mixed-technology systems, however, requires more than just an electronic domain analysis to understand and optimize the functionality of the design. Issues such as coupled energy domain simulation, three-dimensional shape analysis before and possibly after integration, and mixed-technology-interconnect design and analysis need to be addressed as part of the design process. Key to enabling mixed-technology systems is the development of a design environment that supports both design and manufacture based upon many available mixed-technology and electronic building blocks. In addition, design trade-offs, optimizations, and synthesis need to be explored from an overall systems perspective in a mixed-domain design and layout environment. The DARPA-funded Composite CAD[121] program is an attempt to create a design environment that encompasses these challenges. Figure 2.24 shows the paradigm shift in CAD, which is enabled by Composite CAD. In effect, the approach to a system design changes from the current bottom-up process to a top-down process. In the top-down process the overall system requirements are first defined and then reduced to the component level specifications. This is visualized in the spiral model shown in Fig. 2.25. Starting from the center

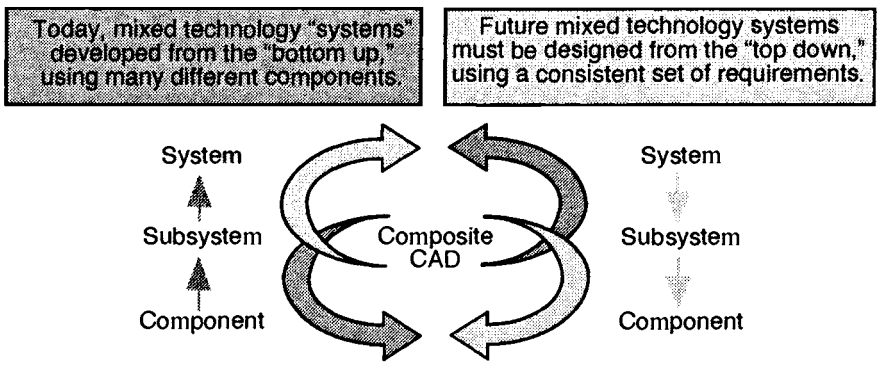


Fig. 2.24. Comparison of today's and future mixed technology design. (Drawing courtesy H. Dussault.[121])

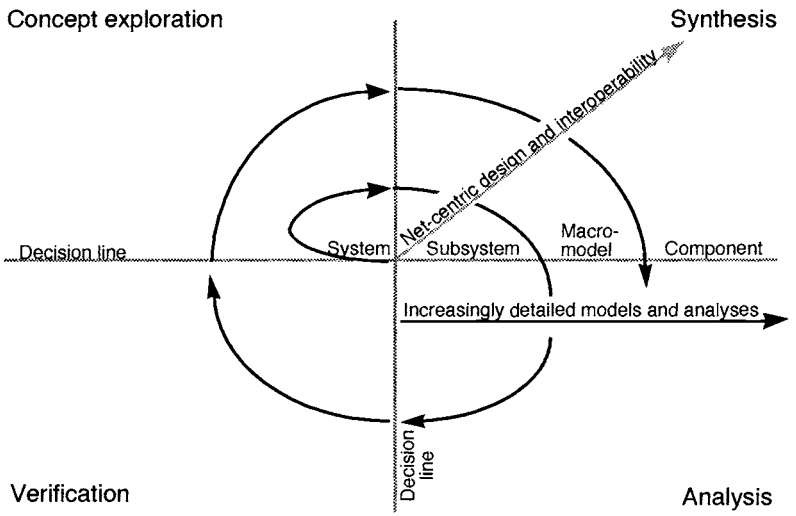


Fig. 2.25. Adapted from the spiral development model for software development and applied to CAD. (First proposed by B. Boehm in 1988 and the Software Productivity Consortium's Evolutionary Spiral Process [ESP] Model in 1991. Drawing courtesy H. Dussault.[121])

and a high-level definition of the required system, there is a sequential process of concept exploration, synthesis, analysis, and verification with decision points at every boundary. With every full cycle the models and analysis become progressively more detailed. Much as CAD has helped to foster new generations of highly complex digital VLSI systems, the Composite CAD program will enable designers to create complex, highly integrated, mixed technology "systems on a chip" by rapidly exploring multiple design alternatives. Efforts similar to the DARPA program are also being explored in Europe and are presented in Chapter 7.

2.4.4 Need for Flight Demonstrations

New technology is fundamentally risky. It must be tested and verified under relevant conditions before being accepted by the aerospace community. "I don't want to be the first to fly that device, but I'll be the second," is commonly heard by technologists trying to get new systems and subsystems used on-orbit. Experimental test flights are required, and Aerospace is trying to shorten the laboratory-to-operational-use time lag for MEMS by inserting emerging devices onto space platforms. The NASA Johnson Space Center, in collaboration with Aerospace, has developed a MEMS testbed that can be flown on the U.S. Space Shuttle.[123] The testbed, shown in Fig. 2.26, is due for flight in 1999 (STS-93) inside a middeck locker. It incorporates multiple MEMS accelerometers, several rate gyros, chemical sensors, nanoelectronics, and a variable surface emissivity device into an industrial PC card frame. Rate gyros and accelerometers are mounted on the rear wall of the middeck locker (top left in Fig. 2.26) to measure Shuttle angular accelerations, vibrations, and linear accelerations. Additional accelerometers and rate gyros are mounted on ISA bus cards within the PC card cage (middle and lower right in the figure) to characterize the experiment environment. The card cage is wrapped in foam and inserted into the middeck locker to provide acoustic and vibration damping. Data are obtained and logged during launch, on-orbit operations, reentry, and landing. The intent is to provide a standard and easy-to-use experiment infrastructure for MEMS researchers; integration of devices into the testbed and integration of the testbed onto the Shuttle are performed by Aerospace, NASA, and Air Force personnel. The middeck

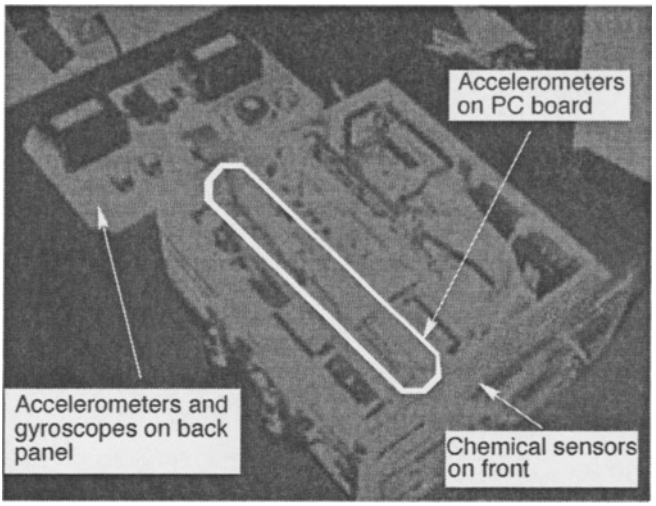


Fig. 2.26. Photograph of the NASA-Aerospace MEMS testbed. The card cage measures 50 × 50 × 25 cm. (Photo courtesy of The Aerospace Corporation.)

implementation provides exposure to launch and reentry loads, microgravity conditions on orbit, and on-board atmosphere (composition and pressure).

2.5 Conclusions

In the United States, both NASA and the DOD have recognized the potential of using microengineered systems in space applications. A similar conclusion has been reached by technology pundits at the ESA[122] and by independent space-system contractors. Major programmatic funding for space applications, outside of that from NASA's New Millennium Program (NMP), still remains in the realm of systems analysis and reliability studies. For example, the U.S. Air Force Research Laboratory in Albuquerque, New Mexico, is performing radiation effects testing of MEMS, and technology demonstration experiments are under way at NASA JPL and Aerospace. Many universities, including those not traditionally involved in MEMS research, are entering the "MEMS for space" arena. Many spacecraft designs will continue to shrink in mass and size, given the resounding success of the Mars Pathfinder mission. Experimental spacecraft currently on the docket will increase the confidence of the space community that small can be good. Additional confidence in "small and capable" spacecraft will be attained as the NMP spacecraft complete their missions to Mars, Pluto, and the asteroids. Finally, there have been numerous workshops hosted by JPL, Round Table discussions hosted by ESA, and a focused conference sponsored by NASA and Aerospace at the Johnson Space Center, Houston, Texas (1995).[122] In April 1999 The Aerospace Corporation, DARPA, JPL, and the Air Force Research Laboratory Vehicle Systems Directorate will host the Second International Conference on Integrated Micro/Nanotechnology for Space Applications, in Pasadena, California. Interest and momentum are increasing steadily.

This chapter has focused on technologies, which if applied to space systems, can result in revolutionary changes in current and future space systems. The specific technologies presented are primarily in the microengineering realm and show clear evidence for worldwide terrestrial use. The underlying assumptions are three: the best means for attracting the space community attention to these new technologies is to identify examples that present distinct advantages when incorporated into space systems; the identified technologies can be incorporated in both a revolutionary and evolutionary manner; and there exists a significant terrestrial application base from

which to draw upon. Other technology areas currently less mature in development will also have revolutionary impact on space systems. Two deserve brief mention: nanotechnology and micro-robotics.

Nanotechnology deals with the development of processes whereby a strong level of atomic or molecular control is exercised in the device fabrication. Micro-robotics is an interdisciplinary technology area whose objective is to assemble a class of limited-"intelligence," autonomous robots of sizes ranging from millimeters to centimeters. Both technologies have identified terrestrial applications. For nanotechnology the industrial drivers are pharmaceuticals, bioengineering, advanced lithography, nanoelectronics (e.g., resonant tunneling devices), and functionalized surfaces. For micro-robotics, the industrial applications appear to be in toys, micro inspection systems (e.g., pipelines), microsurgery, and miniature information devices. Both areas have experienced rapid growth in interest, and both have benefited from MEMS/microsystems technology. Nanotechnology benefits because microsystems used in large arrays permit the nanofabrication/processing over practical areas; Microrobotics, because of the implementation of micro-actuation and the resulting capability to interact with the physical world. For space applications, manned and unmanned, both technologies can potentially revolutionize the deployment, assembly, and governance of space systems. In the near term, nanotechnology will be useful in space as nanoelectronics (multivalued logic circuits); basic components (e.g., resonant tunneling diodes, transistors) and some circuits have already been fabricated. Multilevel logic nanoelectronic circuits can provide the same function as binary circuits but with reduced component count, and they offer distinct advantages in computation-intensive tasks (e.g., image processing). The applications of micro-robotics to space systems will strongly depend on the level of capability endowed. Based on the developments for terrestrial applications (i.e., providing local diagnostics in pipelines), similar applications could be used in the ISS and other satellites, for example, in monitoring the integrity of the ISS hull, fuel tanks, and other critical surfaces that could develop stress, fracture, or sustain a micrometeorite impact. Longer term applications of nanotechnology and micro-robotics lead to speculative answers and deserve the benefit of observation for a few more years.

Microengineered devices will inevitably reduce the size of spacecraft or increase its functionality manifold. The path of size reduction will in turn address the use of a smaller launch vehicle, and this combination will undoubtedly reduce the total cost of launching to orbit. Another expected outcome is that the incorporation of these devices will also increase the autonomy in operations and increase availability through the use of condition-based maintenance protocols. Perhaps the most profound result from this revolution will be that satellites will become truly mass-producible commodities much like dynamic RAM chips are today.

2.6 Acknowledgments

The authors gratefully acknowledge the MEMS and MOEMS communities for providing information, The Aerospace Corporation Corporate Research Initiative Program for supporting much of our research efforts, and The Aerospace Institute for making this publication possible. We also gratefully acknowledge the DARPA MEMS program for supporting the digital microthruster effort at TRW, Aerospace, and California Institute of Technology.

2.7 References

1. H. Helvajian, ed., *Microengineering Technology for Space Systems*, Monograph 97-02 (The Aerospace Press. El Segundo, CA, 1997), p. 1. First published as The Aerospace Corp. Report no. ATR-95(8168)-2 (1995).

2. J. P. Aguttes, J. Sombrin, and E. Conde, "Charting the Course for Radar Sail," *Aerospace America*, **35** (9), 30–33 (September 1997).
3. W. C. Tang, "Micromechanical Devices at JPL for Space Exploration," *1998 IEEE Aerospace Conference* (Snowmass, CO, March 1998).
4. D. Collins, C. Kukkonen, and S. Venneri, "Miniature, Low-Cost, Highly Autonomous Spacecraft—A Focus for the New Millenium," IAF paper 95-U.2.06, *46th International Astronautical Congress* (Oslo, Norway, October 1995).
5. H. W. Price, K.B. Clark, C. N. Guiar, J. M. Ludwinski, and D. E. Smyth, "X2000 Flight Missions Utilizing Common Modular Components," *1998 IEEE Aerospace Conference* (Snowmass, CO, March 1998).
6. *Proceedings 2nd Round Table on Micro/Nanotechnologies for Space* (ESTEC, The Netherlands, 15–17 October 1997). ESA WPP-132; See also A. Martinez de Aragón, "Future Applications of Micro/Nanotechnologies in Space Systems," *ESA Bull.*, No. 85, 65 (February 1996).
7. Brite/Euram II Project, "MAGNIFIT," Reference no. BRE20536.
8. Brite/Euram III Project, "Thick Film Ferroelectric Actuators for New Design Industrial Applications," Reference no. BRPR960318.
9. Brite/Euram III Project, "Basic Research on the Use of Magnetic Fluids in Microsystems," Reference no. BRPR970598.
10. Brite/Euram III Project, "A Novel Method for the Synthesis of Microsize Permanent Magnets," Reference no. BRPR970488.
11. Brite/Euram III Project, "Thick Oxide Films for Passive And Active Optical Components," Reference no. BRPR970434.
12. Esprit-4 Project, "Field Programmable System on Chip," Reference no. 21625.
13. T. Diehl, W. Ehrfeld, M. Lacher, and T. Zetterer, "Electrostatically Operated Micromirrors for a Hadamard Transform Spectrometer," *1998 IEEE/LEOS Summer Topical Meetings on Optical MEMS* MOEMS '98 (Monterey, CA); See also R. Riesenberg, T. Seifert, and J. Schöneich, "Slit Array Made by Microsystem Technology for Performance Improvement of CCD-Detector Arrays and Mini-Spectrometer-Instruments," *Proceedings of 2nd Round Table on Micro/Nano Technologies for Space*, ESA Report WPP-132, p. 145.
14. Fine Sun Sensor Literature, Daimler-Benz Aerospace: Jena-Optronik GmbH, Prüssingstraße 41, D-07745 Jena, Germany.
15. J. P. Krebs, O. Brunel, C. Guerin, D. Guillon, and J. M. Niot, "A New IR Static Earth Sensor for Micro and Nanosatellites," *Proceedings of 2nd Round Table on Micro/Nano Technologies for Space*, p. 187.
16. C. Cavadore, P. Magna, Y. Degerli, A. Gautrand, F. Lavernhe, J. Farre, F. Solhusvik, and R. Davencens, "Active Pixel Image Sensors for Space Applications," *Proceedings of 2nd Round Table on Micro/Nano Technologies for Space*, p. 115; See also W. Ogiers, B. Dierickx, D. Scheffer, D. Uwaerts, G. Meynants, and C. Truzzi, "Compact CMOS Vision Systems for Space Use," *Ibid.*, p. 123; and J. Josset, P. Plancke, G. Boucharlat, and C. Val, "Digital 1k x 1k Microimager for Planetary Surface Exploration," *Ibid.*, p. 143.
17. J. P. Pekola, M. M. Leivo, M. J. Peltomäki, and A. J. Manninen, "NIS Microrefrigerators and Microcalorimeters on Silicon and Silicon Nitride Membranes," *Proceedings of 2nd Round Table on Micro/Nano Technologies for Space*, p. 61.
18. L. Stenmark, J. Köhler, M. Lang, and U. Simu, "Micro Machined Propulsion Components," *Proceedings of 2nd Round Table on Micro/Nano Technologies for Space*, p. 69.
19. R. Linnemann, M. Richter, and P. Woias, *Proceedings of 2nd Round Table on Micro/Nano Technologies for Space*, p. 83.
20. F. Eckhard, B. H. va der Schoot, K. Fluri, Ch. A. Paulus, R. H. Huijser, P. Reijneker, D. Va den Assem, A. J. Kramer, H. Leeuwis, and A. Prak, "CAELIS, Capillary Electrophoresis in Space," *Proceedings of 2nd Round Table on Micro/Nano Technologies for Space*, p. 37.

21. R. Born, Ph. Arquint, B. va der Schoot, F. van Steenkiste, V. Spiering, J. Cefai, and K. Schumann, "Sensor and Integration Technologies for a Multisensor Array," *Proceedings of 2nd Round Table on Micro/Nano Technologies for Space*, p. 45.
22. P. Woias, M. Richter, K. Hauser, E. Yacoub-George, H. Wolf, T. Abel and S. Koch, "Silicon Microreactors for Space-Bound Chemical Microanalysis," *Proceedings of 2nd Round Table on Micro/Nano Technologies for Space*, p. 49.
23. A. Prak, M. Richter, J. Naundorf, M.Eberl, H. Leeuwis, P. Woias, and A. Stechenborn, *Proceedings of 2nd Round Table on Micro/Nano Technologies for Space*, p. 55.
24. "Sensors 2000! Sensor Technology for the Next Millenium," NASA Program at Ames Research Center, http://s2k.arc.nasa.gov/welcome.html (August 1997).
25. S. Santoli, "Nanobiotechnology: An Emerging Source of Innovation for Competitive Space Strategies," *ESA-ESTEC International Workshop on Innovations for Competitiveness* (Noordwjik, The Netherlands, 19–21 March 1997).
26. A. Hansson, "From Microsystems to Nanosystems," Special issue: "From Microsystems to Nanotechnology for Space Missions," *J. of the British Interplanetary Soc.* **51** (4), 123 (1998).
27. A. Teshigahara, M. Watanabe, N. Kawahara, Y. Ohtsuka, and T. Hattori, "Performance of a 7 mm Microfabricated Car," *J. MEMS* **4**, 76 (1995).
28. Asian Technology Information Program, http://www.atip.org/MEMS.
29. S. Konishi and H. Fujita, "A Conveyance System Using Air Flow Based on the Concept of Distributed Micro Motion Systems" *J. MEMS* **3**, 54 (1994).
30. S. W. Janson, "Spacecraft as an Assembly of ASIMS," in *Microengineering Technology for Space Systems*, edited by H. Helvajian, Monograph 97-02 (The Aerospace Press, El Segundo, CA, 1997), pp. 143–201.
31. E. Y. Robinson, "ASIM Applications in Current and Future Space Systems," in *Microengineering Technology for Space Systems*, edited by H. Helvajian, Monograph 97-02 (The Aerospace Press, El Segundo, CA, 1997), p. 21.
32. "Smart Structures and Materials: Implications for Military Aircraft of New Generation," Advisory Group for Aerospace Research & Development (AGARD), AGARD-LS-205 (NATO/AGARD) 1996.
33. W. R. Davis Jr., B. B. Kosicki, D. M. Boroson, and D. F. Kostishack "Micro Air Vehicles for Optical Surveillance," *Lincoln Laboratory J.* **9** (2), 197 (1996).
34. D. A. Lorenzini and C. Tubis "Vehicle Tracking System Using Nanotechnology Satellites and Tags," *Proceedings of the International Conference on Integrated Micro/Nanotechnology for Space Applications* (Houston, TX, 30 October–3 November 1995).
35. P. W. Siglin and G. F. Senn, "The Courier Satellite," in *Communication Satellites. Proceedings of a Symposium Held in London*, edited by L. J. Carter, 1962.
36. R. Steele, C. McCormick, K. Brandt, W. Fornwalt, and R. J. Bonometti, "Utilization of the Multiple Access Communications Satellite (MACSAT) in Support of Tactical Communications," *Proceedings 5th Annual AIAA/USU Conference on Small Satellites* (Utah State University, Logan, Utah, 26–29 August 1991).
37. N. P. Bean, "A Modular Small Satellite Bus for Low Earth Orbit Missions," *Proceedings 2nd Annual AIAA/USU Conference on Small Satellites* (Logan, Utah, September 1988).
38. J.A. King, R. McGwier, H. Price, and J. White, "The In-orbit Performance of Four Microsat Spacecraft," *Proceedings 4th Annual AIAA/USU Conference on Small Satellites*, Vol. 1 (Logan, Utah, 27–30 August 1990).
39. S. Handler, "Birth of a Satellite," *Satellite Times* **1** (4), 10–14 (March/April 1995).
40. P. R. K. Chetty, *Satellite Technology and its Applications* (TAB Professional and Reference Books, Blue Ridge Summit, PA, 1991), pp. 457–468.
41. J. Wallach, "Automatic Pictures via NOAA 14," *Satellite Times*, **1** (4), 66–68 (March/April 1995).
42. R. Tessier, "The Meteosat Programme," *ESA Bull.*, No. **58**, 44–57 (May 1989).
43. S. Bruzzi and M. Wooding, "ERS-1: A Contribution to Global Environmental Monitoring in the 1990s," *ESA Bull.*, No. 62, 11–21 (May 1990).

44. E.W. Mowle, "Landsat-D and D'—The Operational Phase of Land Remote Sensing from Space," in "Monitoring Earth's Ocean, Land, and Atmosphere from Space—Sensors, Systems, and Applications Progress," edited by Abraham Schnapf, *Astronautics and Aeronautics* **97**, 349–370 (1985).
45. M. Courtois and G. Weill, "The SPOT Satellite System," in "Monitoring Earth's Ocean, Land, and Atmosphere from Space—Sensors, Systems, and Applications Progress," edited by Abraham Schnapf, *Astronautics and Aeronautics* **97**, 493–523 (1985).
46. S. J. Isakowitz, *International Reference Guide to Space Launch Systems*, 2nd ed. (American Institute of Aeronautics and Astronautics, Washington, D.C., 1991).
47. B. A. Banks and C. LaMoreaux, "Performance and Properties of Atomic Oxygen Protective Coatings for Polymeric Materials," *24th International SAMPE Technical Conference*, T165-T173 (Toronto, Canada, 20–22 October 1992).
48. R. D. Rasmussen, "Spacecraft Electronics Design for Radiation Tolerance," *Proceedings of the IEEE*, Vol. 76 (November 1988), pp. 1527–1537.
49. M. D. Griffin and J. R. French, *Space Vehicle Design* (American Institute of Aeronautics and Astronautics, Washington, D. C., 1991), p. 70.
50. J. V. Osborn, D. C. Mayer, R. L. Lacoe, S. C. Moss, S. D. Lalumondier, and G. Yabiku, "Total Ionizing Dose and Single Event Latchup Characteristics of Three Commercial CMOS Processes," *Sixth NASA VLSI Design Symposium* (5 March 1997).
51. M. Alles and S. Wilson, "Thin Film Silicon on Insulator: An Enabling Technology," *Semiconductor International*, 67–74 (April 1997).
52. M. M. Gates, M. J. Lewis, and W. Atwell, "Rapid Method of Calculating the Orbital Radiation Environment," *J. Spacecraft and Rockets* **29**(5), 646–652 (September–October 1992).
53. L. Muller, M. H. Hecht, L. M. Mitler, H. K. Rockstad, and J. C. Lyke, "Packaging and Qualification of MEMS-Based Space Systems," *Proceedings IEEE 9th Annual Workshop on Micro Electro Mechanical Systems* (San Diego, CA, February 1996), pp. 503–508.
54. C. Barnes, A. Johnston, and G. Swift, "The Impact of Space Radiation Requirements and Effects on ASIMS," *Proceedings of the International Conference on Integrated Micro/Nanotechnology for Space Applications* (Houston, TX, 30 October–3 November 1995).
55. W. Stuckey, "Environmental Effects Guidelines for Materials Selection," in *Microengineering Technology for Space Systems*, edited by H. Helvajian, Monograph 97-02 (The Aerospace Press, El Segundo, CA, 1997), pp. 51–57.
56. B. Boser and D. Young, "Monolithic Tunable RF Oscillators based on Micromachined Variable Capacitors," http://kowloon.EECS.Berkeley.edu/~boser/vco.html (1997). D. J. Young and B. E. Boser, "A Micromachined Variable Capacitor for Monolithic Low Noise VCOs," *Proceedings of the 7th IEEE Solid State Sensor and Actuator Workshop* (Hilton Head Island, SC, 3–6 June 1996), pp. 86–89.
57. J. Y. Chang, A. A. Abidi and M. Gaitan, "Large Suspended Inductors on Silicon and Their Use in a 2-m CMOS RF Amplifier," *IEEE Electron Devices Letters* **14**, 246–248 (May 1993).
58. D. P. Morgan, *Surface-Wave Devices for Signal Processing* (Elsevier Publishing, NY, 1985), pp. 1–14.
59. J. R. Norton and J.M. Clocren, "Precision Quartz Oscillators and Their Use in Small Satellites," *Proceedings of the 6th Annual AIAA/UTAH State University Conference on Small Satellites* (Logan, Utah, 21–24 September 1992).
60. J. Marshall, M. Gaitan, M. Zaghloul, D. Novotny, V. Tyree, J.-I. Pi, C. Piñà, and W. Hansford, "Realizing Suspended Structures on Chips Fabricated by CMOS Foundry Processes Through the MOSIS Service," Report no. NISTR 5402 (National Institute of Standards and Technology, Gaithersburg, MD, June 1994).
61. R. A. Buser and N.F. deRooij, "Capacitively Activated Torsional High-Q Resonator," *Proceedings of the IEEE Micro Electro Mechanical Systems Conference* (Napa Valley, CA, 11–14 February 1990), pp. 132–135.

62. "Microelectromechanical Systems," R & D Programs, Electronics Technology Office, DARPA, http://web-ext2.darpa.mil/ETO/MEMS/.
63. S. J. Wineland, "Trapped Ions, Laser Cooling and Better Clocks," *Science* **226** (4673), 395 (1984).
64. H. An, B. K. J. C. Nauwelaers, G. A. E. Vandenbosch, and A. R. Van De Capelle, "Active Antenna Uses Semi-Balanced Amplifier Structure," *Microwaves and RF* **33** (13), 153–156 (December 1994).
65. T. Razban, M. Nannini and A. Papiernik, "Integration of Oscillators With Patch Antennas," *Microwave J.* **1993**, 104–11 (January 1993).
66. J. Lin, "Active Integrated Antennas," *IEEE Transactions on Microwave Theory and Techniques* **42** (12), 2186–2194 (December 1994).
67. A. Yarbrough, "Applying Micro-/Nanotechnology to Satellite Communications Systems," in *Micro- and Nanotechnology for Space Systems: An Initial Evaluation*, edited by H. Helvajian and E. Y. Robinson, Monograph 97-01 (The Aerospace Press, El Segundo, CA, 1997), pp. 17–30. First published as The Aerospace Corp. Report no. ATR-93(8349)-1 (1993).
68. C. C. Ling, "A 94 GHz Planar Monopulse Tracking Receiver," *IEEE Transactions on Microwave Theory and Techniques* **42** (10), 1863–1871 (October 1994).
69. G. L. Lan *et al.*, "Millimeter-Wave Pseudomorphic HEMT MMIC Phased Array Components for Space Communications," *Proceedings of the SPIE: Monolithic Microwave Integrated Circuits for Sensors, Radar, and Communications Systems* (Orlando, FL, 2–4 April 1991), pp. 184–192.
70. K. A. Shalkhauser and C. A. Raquet, "System-Level Integrated Circuit (SLIC) Development for Phased Array Antenna Applications," *Proceedings of the SPIE: Monolithic Microwave Integrated Circuits for Sensors, Radar, and Communications Systems*, pp. 204–209.
71. H. J. De Los Santos, R. A. Brunner, J. F. Lam, L. H. Hackett, R. F. Lohr Jr., L. E. Larson, R. Y. Loo, M. Matloubian, and G. L. Tangonan, "MEMS-Based Communications Systems for Space-Based Applications," *Proceedings of the International Conference on Integrated Micro/Nanotechnology for Space Applications* (Houston, TX, 30 October–3 November 1995).
72. T-H. Lin, P. Congdon, G. Magel, L. Pang, C. Goldsmith, J. Randall, and N. Ho, "Silicon Micromachining in RF and Photonic Applications," *Proceedings of the International Conference on Integrated Micro/Nanotechnology for Space Applications* (Houston, TX, 30 October–3 November 1995).
73. M. W. Phipps, "Design and Development of Microswitches for Microelectromechanical Relay Matrices," Master's thesis, Air Force Institute of Technology (AFIT/GE/ENG/95J-02), 1995.
74. P. M. Zavracky, S. Majumder, and N.E. McGruer, "Micromechanical Switches Fabricated Using Nickel Surface Micromachining," *J. Microelectromechanical Systems* **6** (1), 3–9 (March 1997).
75. P. M. Zavracky, "Electrostatically Actuated Micromechanical Switches Using Surface Micromachining," http://www.ece.neu.edu/edsnu/zavracky/mfl/programs/relay/relay.html, 1997.
76. J. S. Eterno, R. O. Zermuehlen and H. F. Zimbelman, "Attitude Determination and Control," in *Space Mission Analysis and Design*, 2nd ed., edited by J. R. Wertz and W. J. Larson (Kluwer Academic Publishers, Boston, 1992), pp. 616–629.
77. W. L. Pritchard and J. A. Sciulli, *Satellite Communication Systems Engineering* (Prentice-Hall, Englewood Cliffs, NJ, 1986), p. 114.
78. W. Johnson, "Attitude Adjustment; GPS Innovation Keeps Satellites Oriented," *Satellite Commun.* **1995**, 19–21 (June 1995).
79. Honeywell HMC2003 Three-Axis Magnetic Sensor Hybrid data sheet, Solid State Electronic Center, Honeywell Inc., Plymouth, MN, 1995.
80. "Rapid Prototype Integrated GMR Magnetic Sensors," "GMR Magnetic Bridge Sensor—NVSB," and "Integrated GMR Magnetic Sensors—NVSI," Nonvolatile Electronics, Inc., Eden Praire, MN, 1995.
81. D. K. Wickenden, R. B. Givens, R. Osiander, J. L. Champion, D. A. Oursler, and T. J. Kistenmacher, "MEMS-based Resonating Xylophone Bar Magnetometer," *Proceedings SPIE Micromachining* (1998).
82. ASEM02-S and ASEM02-T/6 data sheets, *Centre Suisse d'Electronique et de Microtechnique SA*, P.O. Box 41, CH-2007, Neuchatel, Switzerland, 1994.

83. C. A. Kukkonen, presentation at the *Micro/Nano-Technologies for Space Workshop* (European Space Agency ESTEC, Nordwijk, The Netherlands, 27–28 March 1995).
84. M. Schuyer, P. Silvestrin, and M. Aguirre, "Probing the Earth from Space—The Aristoteles Mission," *ESA Bull.*, No. 72, 67–75 (1991).
85. G. N. Smit, "Performance Thresholds for Application of MEMS Inertial Sensors in Space," *Proceedings of the International Conference on Integrated Micro/Nanotechnology for Space Applications*, (Houston, TX, 30 October–3 November 1995).
86. J. H. Connelly, J. P. Galmore, and M. S. Weinberg, "Micro-Electromechanical Instrument and Systems Development at the Charles Stark Draper Laboratory," *Proceedings of the International Conference on Integrated Micro/Nanotechnology for Space Applications* (Houston, TX, 30 October–3 November 1995).
87. IK/80 High Accuracy FOG data sheet, Fibersense Technology Corporation, Canton, MA, 1996.
88. A. K. Henning, "Microfluidic MEMS for Semiconductor Processing," in *Proceedings, Second Annual International Conference on Innovative Systems in Silicon* (IEEE Press, Piscataway, NJ, 1997), pp. 340–349.
89. J. Mueller, *Proceedings of the JPL Micropropulsion Workshop* (Pasadena, CA, 7–9 April 1997).
90. S. W. Janson, R. B. Cohen, H. Helvajian, W. W. Hansen, E. J. Beiting III, B. B. Brady, P. Fuqua, R. Robertson, and M. Abraham, "Digital Micropropulsion Program Status, January 1998," The Aerospace Corp. Report no. ATR-99(7527)-1 (November 1998).
91. S. W. Janson, R. B. Cohen, H. Helvajian, W. W. Hansen, E. J. Beiting III, B. B. Brady, P. Fuqua, R. Robertson, and M. Abraham, "Digital Micropropulsion Program Status, July 1998," The Aerospace Corp. Report no. ATR-99(7527)-2 (November 1998).
92. M. J. Declercq, L. Gerzberg, and J. D. Meindl, "Optimization of the Hydrazine-Water Solution for Anisotropic Etching of Silicon in Integrated Circuit Technology," *J. Electrochem Soc.* **122** (4), 545–552 (April 1975).
93. D. B. Lee, "Anisotropic Etching of Silicon," *J. of Appl. Phys.* **40** (11), 4569–4574 (October 1969).
94. E. W. Schmidt, *Hydrazine and Its Derivatives* (John Wiley & Sons, NY, 1984), pp. 84, 452.
95. M. J. Daneman, O. Solgaard, N. C. Tien, K. Y. Lau, and R. S. Muller, "Laser-to-Fiber Coupling Module Using a Micromachined Alignment Mirror," *IEEE Photonics Technol. Lett.* **8** (3), 396 (1996).
96. M. C. Wu, L. Y. Lin, and S. S. Lee, "Micromachined Free-Space Integrated Optics," *SPIE* **2291**, 40 (1994).
97. V. M. Bright, J. H. Comtois, J. R. Reid, and D. E. Sene, "Surface Micromachined Micro-Opto-Electro-Mechanical Systems," *IEICE Trans. on Electron. Micromachine Technol.* **E80** 206–213 (February 1997).
98. M. J. Daneman, N. C. Tien, O. Solgaard, K. Y. Lau, and R. S. Muller, "Linear Vibromotor-actuated Micromachined Microreflector for Integrated Optical Systems," *Proceedings of the 1996 Workshop on Solid State Sensors and Actuators* (Hilton Head, SC, 1996).
99. D. M. Bloom, "The Grating Light Valve: Revolutionizing Display Technology," *Proceedings of the SPIE—The International Society For Optical Engineering*, Vol. 3013 (San Jose, CA, 10–12 February 1997), pp. 165–171.
100. D. L. Glackin, "Future Space Systems," Invited paper, *2nd International Symposium on The Expansion of the Remote Sensing Market* (Paris, March 1997).
101. L.Y. Lin, S.S. Lee, K. S. J. Pister, and M.C. Wu, "Three-Dimensional Micro-Fresnel Optical Elements Fabricated by Micromachining Technique," *Electron. Lett.* **30** (5), 448–449 (March 1994).
102. J. Klemic, J. M. Sirota, and M. Mehregany, "Fabrication Issues in Micromachined Tunable Optical Filters," *Proceedings SPIE Micromachining and Microfabrication 1995*, p. 89.
103. M. C. Roggemann, V. M. Bright, B. M. Welsh, S. R. Hick, P. C. Roberts, W. D. Cowan, and J. H. Comtois, "User of Micro-Electro Mechanical Deformable Mirrors to Control Aberrations in Optical Systems: Theoretical and Experimental Results," *Opt. Engin.* (May 1997).

104. T. G. Bifano, R. Krishnamoorthy Mali, J. K. Dorton, J. A. Perriault, N. Vandelli, M. N. Horenstein, and D. A. Castanon, "Continuous Membrane, Surface Micromachined, Silicon Deformable Mirror," *Opt. Engin.* (May 1997).
105. B. E. Hardt, R.D. Karam, and R. J. Eby, "Louvers," in *Satellite Thermal Control Handbook*, edited by D. G. Gilmore (The Aerospace Press, El Segundo, CA, 1994), pp. 4-97–4-121.
106. M. A. Mignardi, "Digital Micromirror Array for Projection TV," *Solid State Technol.*, July 1994, 63–68.
107. H. Zaid, P. Van Gerwen, K. Baert, T. Slater, E. Masure, and F. Preud'homme, "Thermal Switch for Satellite Temperature Control," *Proceedings of the International Conference on Integrated Micro/Nanotechnology for Space Applications* (Houston, TX, 30 October–3 November 1995).
108. Y. Cao, A. Faghri, and E. T. Mahefkey, "Micro/Miniature Heat Pipes and Operating Limitations," *HTD 29th National Heat Transfer Conference, Heat Pipes and Capillary Pumped Loops American* (Society of Mechanical Engineers, Atlanta, GA, 1993), pp. 236, 328–335.
109. D. K. Khrustalev and A. Faghri, "Thermal Analysis of a Micro Heat Pipe," *J. of Heat Transfer* **116** (1), 189–198 (1994).
110. G. P Peterson, A. B. Duncan, and M. H. Weichold, "Experimental Investigation of Micro Heat Pipes Fabricated in Silicon Wafers," *J. of Heat Transfer* **115**, 751–756 (August 1993).
111. A. B. Duncan and G. P. Peterson, "Charge Optimization for a Triangular-Shaped Etched Micro Heat Pipe," *J. of Thermophys. and Heat Transfer* **9** (2), 365–368 (April–June 1995).
112. S. W. Janson, H. Helvajian, and E. Y. Robinson, "The Concept of Nanosatellite for Revolutionary Low-Cost Space Systems," Paper IAF-93-U.5.573 presented at *44th Congress of the International Astronautics Federation* (Graz, Austria, October 1993).
113. J. Cheng, P. Zhou, S. Z. Sun, S. Hersee, D. R. Myers, J. Zolper, and G. A. Vawter, "Surface Emitting Laser-Based Smart Pixels for Two-Dimensional Optical Logic and Reconfigurable Optical Interconnections" *IEEE J. Quantum Electron.* **29**, 741 (1993).
114. S. W. Janson, "Silicon Satellites: Picosats, Nanosats, and Microsats," *Proceedings of International Conference on Integrated Micro/Nanotechnology for Space Applications* (Houston, TX, 30 October–3 November 1995).
115. *Proceedings of the International Conference on Integrated Micro/Nanotechnology for Space Applications* (Houston, TX, 30 October–3 November 1995).
116. J. G. Walker, "The Geometry of Satellite Clusters," *J. British Interplanetary Soc.* ***35***, 345–354 (1982).
117. Presidential Blue Ribbon Commission on Acquisition (known as the Packard Commission), 1986.
118. MEMS-Exchange http://www.mems-exchange.org.
119. M. J. Kelley, H. F, Dylla, G. R. Neil, L. J. Brillson, D. P. Henkel, and H. Helvajian, "UV Free Electron Laser Light Source for Industrial Processing" *SPIE* **2703**, 15 (1996); See also H. Helvajian, "Laser Material Processing: A Multifunctional in Situ Processing Tool for Microinstrument Development," *Microengineering Technology for Space Systems*, edited by H. Helvajian, Monograph 97-02 (The Aerospace Press, El Segundo, CA, 1997), p. 67.
120. C. P. Christensen, "Laser Processing Works on a Micro Scale," Industrial Laser Review, Application Report, June 1994; See also R. R. Kunz, M. W. Horn, T. M. Bloomstein, and D. J. Ehrlich, "Application of Lasers in Microelectronics and Micromechanics," *Appl. Surf. Sci.* **79/80**, 12–24 (1994); R. V. Belfatto, F. Sansone, and E. Young, "Laser Based Machines for Hybrid and Semiconductor Manufacturers," *Proceedings SouthCon*, Vol. 90 (1990), p. 90; Revise Corporation literature, Revise Inc., 79 Second Ave., Northwest Park, Burlington, MA, 01803; Potomac Corporation literature, Potomac Photonics Inc., 4445 Nicole Dr., Lanham, MD, 20706.
121. *Proceedings, Design for Mixed Technology Integration (Composite CAD) Program, Semi-Annual Principal Investigator's Meeting* (1–3 December 1997). See also H. Dussault, Program Manager; DARPA Composite CAD home page, http://web-ext2.darpa.mil/eto/CompCad/; R. Goering, "Funding for Advanced Tools to Push Envelope in Sensing, Measurement, Optical Devices—DARPA program Seeds Mixed-Technology CAD," *EE Times*, No. 937, 4 (20 January 1997).

122. A. Martínez de Aragón, "Future Applications of Micro/Nanotechnologies in Space Systems," *ESA Bull.*, No. 85, 65 (February 1996).
123. D. G. Sutton and R. S. Smith, "MEMS Spaceflight Testbed," The Aerospace Corp. Report no. ATR-99(2101)-1 (to be published).

3

Mechanical Analysis and Properties of MEMS Materials

D. J. Chang* and W. N. Sharpe, Jr.†

3.1 Introduction

Microelectromechanical systems (MEMS) is a revolutionary technology involving micro-optics, micromechanics, and microelectronics. The MEMS concept is not about one single application or device, nor is it defined by a single fabrication process or limited to a few materials. Rather, it is a "smart" microinstrument incorporating multiple technologies. MEMS can greatly benefit both the launch and operation of space systems, reducing mass, power consumption, volume, cost of hardware manufacture, and cost of testing. The use of this technology is expected to significantly advance the state of the art in ultra-large-scale integrated systems.

Many types of MEMS devices are needed in both launch-vehicle and space-satellite systems. These include pressure transducers, accelerometers, actuators, and gas-detection sensors, to name a few. These different devices can be combined and assembled in a compact, self-contained manner to replace the current heavy, less flexible, and costly systems. For example, the Liftoff Instrumentation System (LOIS) and the Wideband Instrumentation System (WIS) are now employed to monitor motor takeoff pressure, vehicle acceleration, and pyrotechnic information. LOIS provides data for the first 1.5 s of flight for the Titan launch vehicle.[1] This system uses umbilical cords to record the data. The WIS in-flight system monitors most of the ascent flight and provides a limited number of wideband channels for dynamic environments, such as acoustics and vibration, for which a radio-frequency signal transmission technique is used. Corresponding weights for LOIS and WIS are shown in Table 3.1.

Requirements for sensors for Titan IV LOIS and WIS are shown in Table 3.2. Both systems are limited by the number of telemetry channels available and the high cost of moving sensors from one location to another. However, the same types of information can be collected through the installation of MEMS systems that are assembled in wrist-watch-sized packages. These systems can be mounted next to or on critical locations with virtually no impact on the environment and can make measurements using the vehicle's power and telemetry systems. Since the data are recorded in local memories, they can either be sent back in real time or at a later time.

Table 3.1. Corresponding Weights for LOIS and WIS

System	Weight (lb)
LOIS	234–290
WIS A vehicle	217–399
WIS B vehicle	391–587

*Mechanics and Materials Technology Center, The Aerospace Corporation, El Segundo, California.
†Department of Mechanical Engineering, The Johns Hopkins University, Baltimore, Maryland.

Table 3.2. Sensor Requirements for Titan IV LOIS and WIS

Measurement Parameters	
Three-axis vibration	10-2000 Hz, ± 300 g max. amplitude
Acoustics	10-4000 Hz, ± 185 dB max. range
Acceleration	0-50 Hz, ± 10 g max. amplitude
Pressure	0-50 Hz, 0–16 psia range
Strain	0-50 Hz, 900 μin./in. or 2000 psi stress
Survivability and Environmental Parameters	
Vibration, shock, EMI, radiation, temperature, atmospheric to vacuum pressure, contamination	

Another application of MEMS in space systems is its use in Global Positioning System (GPS) satellites. For example, the so-called integrated GPS/IMU includes an inertial measurement unit (IMU), which is regarded as an important unit that provides acceleration and angular-rate information through accelerometers and gyroscopes. The measured acceleration and angular-rate information can be integrated to obtain both velocity and altitude. MEMS technology could certainly help reduce the size, weight, and required power consumption of both the IMU and receivers.

Mission reliability for space systems is another area where MEMS can play a major role. Currently, Air Force spacecraft programs use thin films on optical, microelectronics, and structural systems. Reliable long-duration performance of these components is critical to the mission success of all spacecraft systems. However, the ability to assess the thin-film factors, such as fatigue life, thermo-optical performance, and to develop countermeasures for withstanding radiation threats, has been less than desirable. Likewise in microelectronics, voids and hillocks appear in aluminum interconnects; cracks are observed in silicon dioxide passivation layers; and high tensile and compressive stresses occur in the transistor gate structures. The magnitudes of these stresses can be as high as 500 MPa (70 ksi). These are mostly caused by stress migration resulting from a combination of thermal-expansion mismatch of various materials and thermal cycling. There are also stress failures induced by electron movement resulting from applied current in the line, so-called "electromigration." Consequently, an accurate assessment of the reliability at both component and system levels has yet to be achieved.

It is apparent that MEMS devices will be used in the near future as a diagnostic method to record needed data and to detect anomalies. Their long-term reliability in meeting this objective is vitally important. Currently, however, no work addresses this issue, and therefore, the reliability issues associated with MEMS components such as structural margin and fatigue life need to be addressed first.

As MEMS technology evolves, the mechanics of fluids and solids, as well as materials properties, become more important than ever. The designer needs to achieve a component design with the objectives of lighter weight and reliable service life while still meeting the excellent performance requirement. The light weight requirement results in thinner or shorter dimensions. When the dimensions reduce the size of 0.1 to 2 μm, which has occurred in many current devices, the geometrical dimension becomes comparable to that of the grain size. The implication of this

phenomenon is that the material may not be assumed homogeneous and isotropic. The stress uniformity assumption that has been accepted for large structures, that is, where the geometrical dimension-to-material grain size is very large, may be violated. As will be discussed in Sec. 3.9, the size of the polysilicon grain is between 0.5 and 1 μm. Because the material is nonhomogeneous and anisotropic, the stress can no longer be expected to be uniform.

Similarly, when the grain-size-to-geometrical-dimension ratio is on the order of 1, the principles of classical fracture mechanics may not be directly applicable for the same reason.

Many current MEMS materials such as thin films are fabricated using different techniques, for example, sputtering and chemical vapor deposition (CVD), to name a couple. The grain size depends strongly on the techniques used as well as on the depositing temperature or annealing process. The different film deposition processes will result in a different material with different fracture properties for the film and different residual stresses between the film and the substrate.

For all types of microsensors and devices, mechanical stresses are involved; for example, bending stresses in accelerometers and biaxial membrane stresses in pressure transducers. The magnitudes of these stresses depend upon the materials and the environment. A MEMS designer must understand the basic theories of continuum mechanics, fracture mechanics, and fatigue. The MEMS designer should also follow the development and the advancement of mechanics on the MEMS scale. Only with such an understanding can one accurately predict the stress fields and range of application of the devices, and thereby obtain higher operational reliability. It is therefore the intent of the authors to introduce the basic principles of continuum mechanics, fracture mechanics, fatigue, failure theories, and other related mechanics topics to MEMS designers and users so that the quality of MEMS devices can be improved.

In the sections to follow, various topics pertinent to solid mechanics are discussed:

- **Section 3.2**: Stress and strain, illustrating the various types for different types of MEMS devices
- **Section 3.3**: Classical constitutive relations between stress and strain components based on energy consideration
- **Section 3.4**: Linear piezoelectricity
- **Section 3.5–3.6**: Failure theories and the concept of fracture mechanics for solids
- **Section 3.7**: More detailed material elastic properties, specifically, coefficients of thermal expansion and failure properties in MEMS applications
- **Section 3.8**: Dynamic induced-fatigue behavior of materials
- **Section 3.9**: Microstructure formation of some MEMS materials
- **Section 3.10**: Other MEMS-related subjects
- **Section 3.11**: Sample applications pertinent to different devices
- **Section 3.12**: Current research in mechanics relating to MEMS applications

3.2 Stress and Strain in MEMS

In the discussion of mechanics in solids, some basic terms first need to be defined, the most important of which are "stress" and "strain." Next, the linear stress and strain relationship characterized by Hooke's law will be described. The linear part of the stress-strain curve should be used in the MEMS design for long service life. When the stress-strain gets into the nonlinear region, inelastic strain would most likely exist, and the required service life would not be assured.

It should be mentioned that for certain devices such as a membrane pressure transducer, even though the displacement may get large, the strain is small. As long as the strain is small, the linear stress-strain relation still works for the design.

We assume that the material body in our discussion is a continuum medium. Under the application of forces, the body will deform from its original or "unstressed" shape. A body is called elastic if it possesses the property of recovering its original shape when the applied forces causing deformation are removed. Further, the elastic body is linear when the deformation is proportional to the applied forces. Our discussion will be limited to linear elastic behavior. Readers are encouraged to read Refs. 2, 3, and 4 for more details.

3.2.1 State of Stress

Let us define a rectangular Cartesian coordinate system for a continuum medium. Define x_1, x_2, and x_3 as the three mutually perpendicular right-hand coordinate axes. Assume that V represents the volume occupied by the medium and ΔV is an element of V. There are two types of forces acting on the volume element ΔV:

- Body forces: forces proportional to the mass contained in ΔV, designated as F with components F_1, F_2, and F_3,
- Surface forces: forces acting on the surface ΔS of ΔV

Consider a force ΔT acting on the surface element ΔS. The stress is defined as the limiting value of $\Delta T/\Delta S$. Since T is in general a vector, the stress values also have directional preference. Figure 3.1 shows that there are nine stress components associated with the three surface force vectors T_1, T_2, and T_3. These components are mathematically referred to as the elements of a second-order stress tensor. These stress tensor elements are σ_{11}, σ_{12}, σ_{13}, σ_{21}, σ_{22}, σ_{23}, σ_{31}, σ_{32}, and σ_{33}, respectively. The first index i in a stress component σ_{ij} refers to the direction of the coordinate axis normal to the element surface on which T acts; while the second index j indicates the direction of the stress component. For example, in σ_{23}, the subscript 2 indicates that the normal-to-the-element surface on which T_2 acts is x_2; while subscript 3 indicates that the direction of this stress component is parallel to x_3.

The components σ_{11}, σ_{22}, σ_{33} are called the normal components of stresses; the others are called the tangential, or shearing, components. Normal stresses act normal to a surface. Hydrostatic pressure is an example. Shear stresses act parallel to the surface, such as those generated by friction between two surfaces. The nine stress components can be expressed in matrix form, Eq. (3.1).

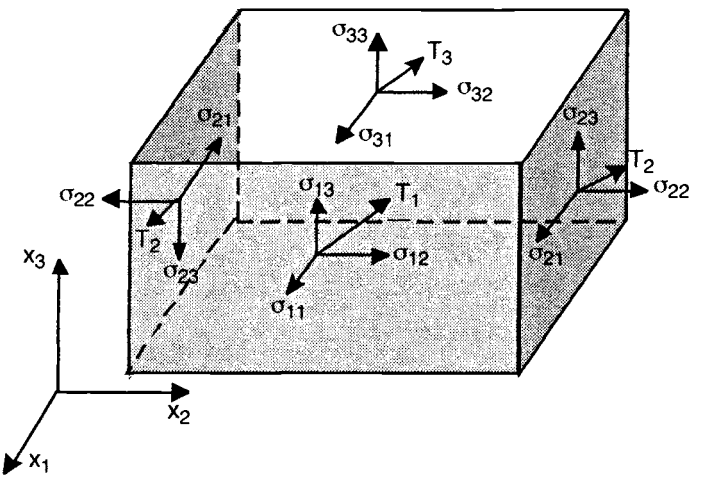


Fig. 3.1. Force vectors and stress tensors in a Cartesian coordinate.

$$\begin{bmatrix} \sigma_{11} & \sigma_{12} & \sigma_{13} \\ \sigma_{21} & \sigma_{22} & \sigma_{23} \\ \sigma_{31} & \sigma_{32} & \sigma_{33} \end{bmatrix} \tag{3.1}$$

The following examples are different stress types.

- Pinned truss element: uniform normal tension or compression
- Prismatic cantilever bar under bending: normal tension and compression varied linearly across a lateral dimension such as thickness
- Circular shaft under torsion: shear stresses with linearly varying magnitude with respect to the center of the shaft
- Circular membrane under normal pressure: equal biaxial normal stresses
- Cube under hydrostatic pressure: triaxial normal stresses with equal magnitude. No shear stresses present.

It can be proved that the stress components must satisfy the equilibrium equations, Eq. (3.2).

$$\begin{aligned} \frac{\partial \sigma_{11}}{\partial x_1} + \frac{\partial \sigma_{21}}{\partial x_2} + \frac{\partial \sigma_{31}}{\partial x_3} + F_1 &= 0 \\ \frac{\partial \sigma_{12}}{\partial x_1} + \frac{\partial \sigma_{22}}{\partial x_2} + \frac{\partial \sigma_{32}}{\partial x_3} + F_2 &= 0 \\ \frac{\partial \sigma_{13}}{\partial x_1} + \frac{\partial \sigma_{23}}{\partial x_2} + \frac{\partial \sigma_{33}}{\partial x_3} + F_3 &= 0 \end{aligned} \tag{3.2}$$

3.2.2 State of Strain

When the positions of material points in a continuous body are changed due to an external applied force, we say the body has been displaced. If the displacement has not produced any relative position change between any pair of material points, then it is referred to as rigid, or nondeformable, displacement. If the relative position between any pair of points in the body is altered, the body is then deformed.

When the displacement vector ΔL lies in the same direction of the unstrained material vector L, the material is then said to be extensionally strained. The extensional strain is defined as the limiting value of $\Delta L/L$. For example, given a constant cross-section rod that is 2 cm long, if the total elongation of the rod is 0.1 cm, then the extensional strain of the rod is $0.1/2 = 0.05$, or 5%. If the displacement vector is perpendicular to the unstrained material vector L, the strain $\Delta L/L$ is of shear type. In other words, the shear strain is defined as an angular change. For example, given a 2- × 1-cm rectangle, if the two shorter sides slide 0.1 cm in a parallel manner, the rectangle becomes a parallelogram. But there is an angle change of $0.1/2 = 0.05$ radian, which is the shear strain. It is required that the displacements, and thus strains, be small in treating linear elastic problems. This means that $\Delta L/L << 1$. This requirement also suggests that the principle of superposition holds. That is, $L(p1 + p2) = L(p1) + L(p2)$, where L is a linear operator and $p1$ and $p2$ are either two applied-load or displacement fields. The equation indicates that the sequence of various load (displacement) applications does not affect the answer.

Analogous to stress, there are nine strain components, ε_{11}, ε_{12}, ε_{13}, ε_{21}, ε_{22}, ε_{23}, ε_{31},ε_{32}, ε_{33}, which are the elements of the second order of strain tensor. Also, similar to the stress components ε_{11}, ε_{22}, ε_{33} are normal strains, and all other components are shear components of the strain tensor. These nine strain components can also be expressed in matrix form, Eq. (3.3).

$$\begin{bmatrix} \varepsilon_{11} & \varepsilon_{12} & \varepsilon_{13} \\ \varepsilon_{21} & \varepsilon_{22} & \varepsilon_{23} \\ \varepsilon_{31} & \varepsilon_{32} & \varepsilon_{33} \end{bmatrix} \tag{3.3}$$

Examples of different strain types include

1. Pinned truss element: uniform normal extensional or contraction strain
2. Prismatic cantilever bar under bending: normal extensional and contraction strains varying linearly across the lateral dimension, such as thickness
3. Circular shaft under torsion: shear strains with linearly varying magnitude with respect to the center of the shaft
4. Circular membrane under normal pressure: equal biaxial extensional strains
5. Cube under hydrostatic pressure: triaxial normal strains with equal magnitude. No shear strains present.

It should be noted that both stresses and strains are real. Based on the strain energy consideration, both stress and strain fields are symmetrical. This means that

$$\sigma_{12} = \sigma_{21} \qquad \varepsilon_{12} = \varepsilon_{21}$$
$$\sigma_{23} = \sigma_{32} \qquad \varepsilon_{23} = \varepsilon_{32}$$
$$\sigma_{31} = \sigma_{13} \qquad \varepsilon_{31} = \varepsilon_{13}$$

The stress symmetry relation may not be valid for the case when there are body moments distributed through the material medium. For example, when grains become dumbbell shaped under the influence of externally applied electric or electromagnetic fields, the body moments may be induced. Because continuum mechanics was developed without consideration for the length dimension of average grain, predictions based on continuum mechanics start to deviate at the micro level. The derivation of "couple stress" reduces the errors associated with this discrepancy. More detailed information can be found in the literature.[5,6]

It will become apparent that it is more convenient to express the symmetrical stress $\{\sigma\}$ and strain $\{\varepsilon\}$ tensors using the following compressed notations.

$$\sigma_1 = \sigma_{11} \qquad \varepsilon_1 = \varepsilon_{11}$$
$$\sigma_2 = \sigma_{22} \qquad \varepsilon_2 = \varepsilon_{22}$$
$$\sigma_3 = \sigma_{33} \qquad \varepsilon_3 = \varepsilon_{33}$$
$$\sigma_4 = \sigma_{23} \qquad \varepsilon_4 = \varepsilon_{23}$$
$$\sigma_5 = \sigma_{31} \qquad \varepsilon_5 = \varepsilon_{31}$$
$$\sigma_6 = \sigma_{12} \qquad \varepsilon_6 = \varepsilon_{12}$$

3.3 Constitutive Relations

3.3.1 Stiffness Matrix

We will use the principal material axes as the coordinate system. Based on Hooke's law, the stress-strain relations for a linear elastic material can be expressed as

$$\{\sigma\} = [C]\{\varepsilon\} \tag{3.4}$$

where

$$\{\sigma\} = \begin{bmatrix} \sigma_1 \\ \sigma_2 \\ \sigma_3 \\ \sigma_4 \\ \sigma_5 \\ \sigma_6 \end{bmatrix} \quad \text{and} \quad \{\varepsilon\} = \begin{bmatrix} \varepsilon_1 \\ \varepsilon_2 \\ \varepsilon_3 \\ \varepsilon_4 \\ \varepsilon_5 \\ \varepsilon_6 \end{bmatrix} \quad \text{and} \quad [C] = \begin{bmatrix} C_{11} & C_{12} & C_{13} & C_{14} & C_{15} & C_{16} \\ C_{21} & C_{22} & C_{23} & C_{24} & C_{25} & C_{26} \\ C_{31} & C_{32} & C_{33} & C_{34} & C_{35} & C_{36} \\ C_{41} & C_{42} & C_{43} & C_{44} & C_{45} & C_{46} \\ C_{51} & C_{52} & C_{53} & C_{54} & C_{55} & C_{56} \\ C_{61} & C_{62} & C_{63} & C_{64} & C_{65} & C_{66} \end{bmatrix}$$

The C_{ij} are components of the elastic moduli matrix [C] representing the material properties of the continuous body and are referred to as the material stiffness matrix. They are also components of a fourth-order tensor. All the components are real since they represent actual material properties. The subscripts follow the short notations described earlier for stress and strain. For example, C_{11} is formally C_{1111}, C_{66} is formally C_{1212}, and C_{34} is formally C_{3323}. In general, the values of [C] are a function of the material point. However, the C_{ij} become invariant for a fixed coordinate system when the continuous body is homogeneous.

There are 36 independent elastic constants in the matrix [C]. It has been proved again from the strain energy consideration that the elastic moduli tensor is symmetrical. This leaves only 21 independent constants for a fully anisotropic elastic material.

Consider a body elastically symmetric with respect to the x_1x_2 plane. Symmetry reduces the number of constants to 13. For this example, the C_{ij} that are identically zero are C_{14}, C_{15}, C_{24}, C_{25}, C_{34}, C_{35}, C_{46}, and C_{56}. When there is additional symmetry about the x_2x_3 plane, the number of elastic constants is further reduced to nine. This means that C_{16}, C_{26}, C_{36}, C_{45} become zero. When a material has two planes of symmetry, it also has symmetry about three orthotropic planes and is called an orthotropic material. Many semiconductor materials such as silicon and germanium are orthotropic, and nine independent elastic constants are needed to describe the material property.

An orthotropic material is called transversely isotropic on the x_1x_2 plane when the material properties are independent of the orientation on the plane. In this case, the number of material constants is further reduced to five. Hence, relationships between C_{ij} are

$$C_{11} = C_{22}$$
$$C_{44} = C_{55}$$
$$C_{13} = C_{23}$$
$$C_{66} = (C_{11} - C_{12})/2$$

The number of elastic constants can again be reduced if they are independent of orientation. This is so-called isotropic material. Only two constants are required to describe the material property. If we choose C_{11} and C_{12} as the independent constants, the other constants can be expressed as follows:

$$C_{22} = C_{33} = C_{11}$$
$$C_{13} = C_{23} = C_{12}$$
$$C_{44} = C_{55} = C_{66} = (C_{11} - C_{12})/2$$

When the material is either orthotropic, transversally isotropic, or isotropic, there is no coupling between the normal stress (strain) and the shear strain (stress). The manipulation of the stress-strain can be done using the 3×3 subset of [C] matrices involving normal stresses (strains) only. (See Eq. [3.4].)

3.3.2 Compliance Matrix

The strain-stress relation can be obtained by inverting $[C]$ in Eq. (3.4) and is expressed in Eq. (3.5).

$$\{\varepsilon\} = [C]^{-1}\{\sigma\} = [S]\{\sigma\} \tag{3.5}$$

where

$$[S] = \begin{bmatrix} S_{11} & S_{12} & S_{13} & S_{14} & S_{15} & S_{16} \\ S_{21} & S_{22} & S_{23} & S_{24} & S_{25} & S_{26} \\ S_{31} & S_{32} & S_{33} & S_{34} & S_{35} & S_{36} \\ S_{41} & S_{42} & S_{43} & S_{44} & S_{45} & S_{46} \\ S_{51} & S_{52} & S_{53} & S_{54} & S_{55} & S_{56} \\ S_{61} & S_{62} & S_{63} & S_{64} & S_{65} & S_{66} \end{bmatrix} \tag{3.6}$$

and $[S]$ is called the compliance matrix. Its elements are components of a fourth-order tensor, and the product of $[C][S]$ is equal to $[I]$, which is an identity matrix.

When the material is an orthotropic material, $[S]$ has the following form:

$$[S] = \begin{bmatrix} S_{11} & S_{12} & S_{13} & 0 & 0 & 0 \\ S_{21} & S_{22} & S_{23} & 0 & 0 & 0 \\ S_{31} & S_{32} & S_{33} & 0 & 0 & 0 \\ 0 & 0 & 0 & S_{44} & 0 & 0 \\ 0 & 0 & 0 & 0 & S_{55} & 0 \\ 0 & 0 & 0 & 0 & 0 & S_{66} \end{bmatrix}$$

where

$$\begin{bmatrix} S_{11} & S_{12} & S_{13} \\ S_{21} & S_{22} & S_{23} \\ S_{31} & S_{32} & S_{33} \end{bmatrix} = \begin{bmatrix} C_{11} & C_{12} & C_{13} \\ C_{21} & C_{22} & C_{23} \\ C_{31} & C_{32} & C_{33} \end{bmatrix}^{-1} \quad \text{and} \quad \begin{aligned} S_{44} &= 1/C_{44} \\ S_{55} &= 1/C_{55} \\ S_{66} &= 1/C_{66} \end{aligned}$$

3.3.3 Relations Between [C], [S], and Engineering Material Properties Terms

Uniaxial tests are the most common tests for generating material properties. Introduced here are the commonly used terminologies associated with uniaxial testing. During a uniaxial test, the specimen is usually in a cylindrical-shaped dog-bone configuration for thick material or a flat dog-bone specimen if the body layer is thin. The American Society for Testing and Materials (ASTM) has issued specifications for the methods of preparation, testing, data acquisition, and determination of the required engineering values for various types of materials and environments. For example, ASTM specification E8 provides the required testing apparatus, specimen configuration, and test procedure for tensile testing of metallic materials. A typical stress-strain relationship is depicted in Fig. 3.2. The small strain portion of the curve is usually linear between stress and strain. The slope E of the linear portion of the curve is called "modulus of elasticity" or "Young's modulus." The maximum stress in which strain remains directly proportional to stress

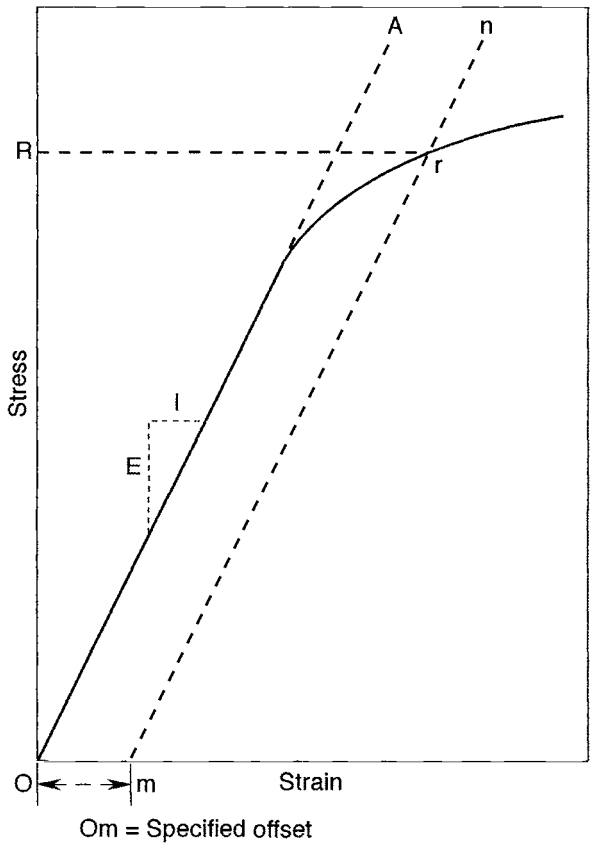


Fig. 3.2. Stress-strain diagram for determination of yield strength by the offset method.

is called "proportional limit." Yield strength F_{ty} of a metallic material is defined as the point r on a stress-strain curve, when a parallel line mn drawn from this point r intersects the horizontal (strain) axis at a strain of 0.002 (0.2%) om, offset. Ultimate strength F_{tu} is the stress level on the stress-strain curve at which the specimen is no longer capable of resisting any load. The yield strength at the micro level is defined in the same fashion with the offset strain equal to 10^{-6} m/m.

Tensile testing is generally conducted using a specimen that has a uniform cross section in the midsection of a known gauge length segment. Under the tensile load, elongation ε is defined as the increase in gauge length, measured after fracture of the tensile specimen occurred within the gauge length segment, expressed as a percentage of the original gauge length. Since the gauge length is rather arbitrary, this number does not represent the actual strain-to-failure at the vicinity of the failure. Rather, it is an average of the inelastic strain (nonrecoverable strain) at failure within the material length in which the strain is measured (gauge length). In other words, a gauge length of 1 cm will correspond to an average strain-to-failure that is different from the average strain-to-failure for a gauge length of 2 cm.

Reduction of area (RA) is the difference between the original cross-sectional area of the tensile test specimen and the minimal cross-sectional area measured after fracture of the specimen, expressed as the percentage of the original cross-sectional area. Both elongation and reduction of area are quantities used to measure the ductility of the material.

Similarly, modulus of rigidity, or shear modulus G, is defined as one-half of the ratio of shear stress to shear strain in the linear portion of the shear test. The shear modulus is used to calculate shear stress or shear strain. The ASTM specifications for shear testing are ASTM specification D1002 for metal-to-metal tension shear, D3518 for inplane shear of reinforced plastics, and D3528 for double lap shear adhesive joints for tension shear.

Last, we need to define "Poisson's ratio." As expressed in Eq. (3.5), the strain in direction x_2 will be affected by the stresses exerted in all three directions. For example, in Eq. (3.5) the off-diagonal terms S_{12}, S_{13}, S_{23} are the coupling coefficients relating the stress to the strain. This is the strain-stress coupling effect, also called Poisson's effect. Equation (3.5) can also be rewritten as Eq. (3.7):

$$S = \begin{bmatrix} \frac{1}{E_{11}} & -\frac{\nu_{21}}{E_{22}} & -\frac{\nu_{31}}{E_{33}} \\ -\frac{\nu_{12}}{E_{11}} & \frac{1}{E_{22}} & -\frac{\nu_{32}}{E_{33}} \\ -\frac{\nu_{13}}{E_{11}} & -\frac{\nu_{23}}{E_{22}} & \frac{1}{E_{33}} \end{bmatrix} \tag{3.7}$$

where E_{ii} is the modulus in i direction and ν_{ij} is defined as the negative strain in the j direction when the strain in the i direction is 1 m/m and the applied load is in the i direction. In an orthotropic material, there are six Poisson's ratios, but only three are independent. The other three are determined by the relations:

$$\begin{aligned} \frac{\nu_{12}}{E_{11}} &= \frac{\nu_{21}}{E_{22}} \\ \frac{\nu_{13}}{E_{11}} &= \frac{\nu_{31}}{E_{33}} \\ \frac{\nu_{23}}{E_{22}} &= \frac{\nu_{32}}{E_{33}} \end{aligned} \tag{3.8}$$

For isotropic materials, Eq. (3.8) is reduced to ν/E.

3.3.4 Transformation of Stress and Strain Tensors

As was discussed earlier, all components of stress, strain, stiffness, and compliance matrices are elements of tensors and can be transformed from one coordinate system to another. Without discussing the details of the law of tensor transformation, the relations between transformed and original components of these four tensors follow.

Let the original Cartesian coordinate system be x_1, x_2, and x_3; while the new coordinate system is y_1, y_2, and y_3. The relations between the two systems are:

$$\begin{Bmatrix} y_1 \\ y_2 \\ y_3 \end{Bmatrix} = \begin{bmatrix} l_1 & m_1 & n_1 \\ l_2 & m_2 & n_2 \\ l_3 & m_3 & n_3 \end{bmatrix} \begin{Bmatrix} x_1 \\ x_2 \\ x_3 \end{Bmatrix} \tag{3.9}$$

where the square matrix represents the directional cosines between the two systems. The same stress field originally expressed in the x_1, x_2, x_3 system can now be expressed in the y_1, y_2, y_3 coordinate system by:

$$\begin{bmatrix} \sigma_{11} & \sigma_{12} & \sigma_{13} \\ \sigma_{21} & \sigma_{22} & \sigma_{23} \\ \sigma_{31} & \sigma_{32} & \sigma_{33} \end{bmatrix} = \begin{bmatrix} l_1 & m_1 & n_1 \\ l_2 & m_2 & n_2 \\ l_3 & m_3 & n_3 \end{bmatrix} \begin{bmatrix} \sigma_{11} & \sigma_{12} & \sigma_{13} \\ \sigma_{21} & \sigma_{22} & \sigma_{23} \\ \sigma_{31} & \sigma_{32} & \sigma_{33} \end{bmatrix} \begin{bmatrix} l_1 & l_2 & l_3 \\ m_1 & m_2 & m_3 \\ n_1 & n_2 & n_3 \end{bmatrix} \tag{3.10}$$

Equation (3.10) also applies to the strain field transformation simply by replacing the stress matrix [σ] with the strain matrix [ε].

By proper choice one can obtain a y_1, y_2, y_3 system such that the shear stress components at that material point disappear. Similarly, a y_1', y_2', y_3' system can be found such that the shear strain components disappear. If the material is at most orthotropic (with nine or less elastic constants), the y_1, y_2, y_3 and y_1', y_2', y_3' systems coincide and the three axis directions are called principal directions and the corresponding stress and strain fields are called principal stresses and principal strains. Determination of the principal directions can be found in the literature.[3]

3.3.5 Thermal Stress

Structures such as microelectronics components usually are subject to electrical power. The power-generated heat produces either uniform temperature or thermal gradients throughout the structure. Thermal stress will be induced by either (1) uniform temperature when the structure has multiple materials stacked monolithically in layers with different coefficients of thermal expansion (CTE), defined as the amount of elongation per unit length of material per unit temperature increase, or (2) a structure with nonuniform thermal gradient. In the first case, the material with lowest CTE is in tension; while the material of highest CTE is in compression. In the second case, if the CTE is positive, the cooler side is in tension and the warmer side in compression.

In the formulation of a thermal stress problem, the mechanical strain tensor $\{\varepsilon_{ij}\}$ in Eqs. (3.3), (3.4), and (3.5) needs to be replaced by $\{\varepsilon_{ij} - \delta_{ij}\alpha\Delta T\}$, where δ_{ij} is the Kronecker delta taking the value of 1 when $i = j$ and 0 when $i \neq j$, α is the material CTE, and ΔT is the temperature rise above the reference temperature. Some simple thermal stress examples are presented here. The first example is a statically determinate thin strip with length L, and a rectangular cross section of width w and thickness $2h$. Assume that the temperature is uniform in the length and width directions and varies only in the thickness direction with a profile $T(y)$, $(-h < y < h)$. The only nonvanishing stress is the longitudinal direction normal stress and is expressed as:[4]

$$\sigma_x = -\alpha E T + \frac{\alpha E}{2h}\int_{-h}^{h} T dy + \frac{3\alpha E y}{2h^3}\int_{-h}^{h} T y dy \tag{3.11}$$

As a second example, let us assume a statically bimetallic strip with E_1, α_1, and thickness h/2 for material 1 and E_2, α_2 and thickness $h/2$ for material 2. Given a temperature rise ΔT throughout the entire strip, the stress field across the thickness is:[7]

$$\sigma_X = \frac{E_1 E_2(\alpha_2 - \alpha_1)\Delta T}{2(E_1 + E_2)}\left[-1 + 6\left(\frac{y}{h}\right)\right] \quad \text{for strip 1} \quad 0<y<h/2$$

$$\sigma_X = \frac{E_1 E_2(\alpha_2 - \alpha_1)\Delta T}{2(E_1 + E_2)}\left[1 + 6\left(\frac{y}{h}\right)\right] \quad \text{for strip 2} \quad -h/2<y<0 \tag{3.12}$$

3.3.5.1 Simple Geometries

Microdevices, or parts of them, often have shapes that are very simple and therefore amenable to easy calculation of the stresses and deflections. Finite element analysis is not needed, and a "back-of-the-envelope" computation can be made as part of a brainstorming session or initial design. This approach is taught at the undergraduate level in a course entitled "Mechanics of Deformable Solids," or known by another title, "Strength of Materials." Numerous textbooks are available.[8,9]

The basic assumption about the geometry of the component is that it is long and thin with a uniform cross section, meaning the length is on the order of 10 times the largest cross-sectional dimension and the cross-sectional dimensions are roughly the same. A component 100-μm long having a cross section 5 × 2 μm would be appropriate, but one 20-μm long with the same cross section would not. Neither would a cross section 20-μm wide by 2-μm thick be suitable. Handbooks giving stresses and deflections for geometries of various simple shapes with different loadings are available, such as the classic book by Roark.[10] We present here only the very simplest cases of axial loading, bending, and torsion for illustration. Geometry and loading are shown in Fig. 3.3. Assume that the left end of the member is fixed; that is, this would be a cantilever beam.

Consider the response of the long, thin rod to an axial load P_1, a transverse load P_2, and a torque T. The following stress cases are discussed.

3.3.5.2 Axial Loading

If only axial load P_1 is present, the axial stress σ everywhere in the rod is

$$\sigma = \frac{P_1}{A},$$

where A is the cross-sectional area (bh, in this case).

The deflection of the end is

$$\delta_x = \frac{P_1 L}{AE},$$

where E is the Young's modulus of the material.

3.3.5.3 Bending

If the rod is subjected to a transverse load P_2, the stress will be nonuniform both along its length and across its cross section. One treats this as a two-dimensional problem, with the stress varying in the x and y directions, but not in the z direction. The stress at any point is given by

$$\sigma_x(x, y) = \frac{M(x)y}{I},$$

where I is the moment of inertia of the cross section about its centroidal axis parallel to the z direction. In this case,

$$I = \frac{1}{12}bh^3$$

The moment here is $P_2 (L - x)$, generating the maximum moment at the left, or fixed, end of the beam. The maximum stress occurs at the top of the beam (tension) or the bottom (compression) where $y = h/2$ or $-h/2$.

Often only maximum stress is of interest; the formula then is simply

$$\sigma_{max} = \frac{M_{max}c}{I}$$

where c is the largest dimension in the y direction.

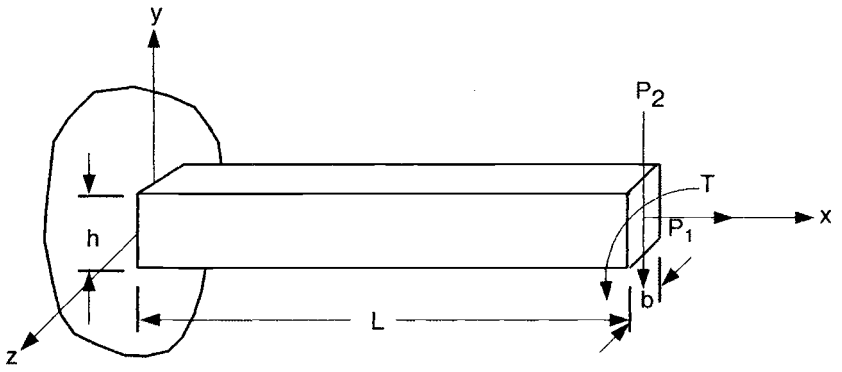


Fig. 3.3. Geometry and loading of a simple member.

The calculation of the maximum bending stress can become complicated when the cross-sectional dimensions as well as the moment vary with x.

The deflection of the end of the rod is for this case,

$$\delta_y = \frac{P_2 L^3}{3EI}$$

3.3.5.4 Torsion

The situation when the rod is subjected to torsion is similar to that of bending. Assuming that the rod is circular in cross section instead of rectangular, the stress is

$$\tau(x, r) = \frac{T(x)r}{J}$$

where τ is now the shear stress, r is the radial distance from the center of the cross section, and J is the polar moment of inertia of the cross section about its centroidal axis, x in this case. If R is the outer radius of the rod, then J is

$$J = \pi \frac{R^4}{2}$$

The maximum stress occurs on the outer surface and is

$$\tau_{\max} = \frac{TR}{J}$$

The angular deflection at the end of the rod, θ_x, is

$$\theta_x = \frac{TL}{JG}$$

where G is the shear modulus of the material.

A few general comments can be made based on these simple examples. Stresses are dependent upon only the geometry and loading, not the material properties, because a cantilever is a statically determinate structure. Stresses are linearly related to the applied loading and distance from the neutral axis, but inversely proportional to the moment of inertia, which has units of cross-sectional dimensions to the fourth power. Deflection does depend on the stiffness of the material through either Young's modulus E or the shear modulus G.

3.4 Piezoelectricity

Piezoelectricity is the capability of certain crystalline materials to change their dimensions (strained) when subjected to an electric field or, conversely, to produce electrical signals when mechanically deformed. Piezoelectric materials find wide use in microsystems.

Depending upon their degree of symmetry, crystals are commonly classified into seven systems: triclinic, monoclinic, orthorhombic, tetragonal, hexagonal, trigonal, and isometric. They are in turn divided into 32 point classes according to their symmetry with respect to a point. Twenty of the 32 classes can be piezoelectric. The linear piezoelectric constitutive equations are[11]

$$\sigma_{ij} = C_{ijkl}\varepsilon_{kl} - e_{kih}E_k$$

$$D_i = e_{ikl}\varepsilon_{kl} + \lambda_{ik}E_k \tag{3.13}$$

where e_{kij}, elements of a third-order tensor, represent the piezoelectric coefficients; E_k, a vector field, is the electric field in voltage; λ_{ik}, elements of a second-order tensor, are the dielectric coefficients; and D_i is the electric displacement. The term e_{kij} can be expressed in a matrix form using a compressed format (Sec. 3.2.2) as follows:

$$e = \begin{bmatrix} e_{1x} & e_{1y} & e_{1z} \\ e_{2x} & e_{2y} & e_{2z} \\ e_{3x} & e_{3y} & e_{3z} \\ e_{4x} & e_{4y} & e_{4z} \\ e_{5x} & e_{5y} & e_{5z} \\ e_{6x} & e_{6y} & e_{6z} \end{bmatrix} \tag{3.14}$$

Symmetry once more reduces the number of independent constants required from a possible 27. For example, quartz has two nonzero independent piezoelectric constants $e_{6y} = e_{2x} = -e_{1x}$ and $e_5 = -e_{4x}$. Cubic crystals such as gallium arsenide and germanium have only one nonzero constant $e_{4x} = e_{5y} = e_{6z}$. An isotropic material, or a cubic crystal with a center of symmetry such as silicon, is not piezoelectric.

There are several types of piezoelectric materials, such as the ceramic type, the natural crystalline type, and the polymer type. The ceramic type, such as lead zirconate titanate (PZT), is polycrystalline in nature. Initially, the ceramics do not have piezoelectric properties. The piezoelectricity is induced by a polarizing treatment (poling) that aligns the polar axes of individual crystallites.

Kocharyan *et al.*[12] investigated a number of polar and nonpolar polymers subjected to a high-poling voltage and a high-frequency field. They found that the higher the polarity of the unit cell of the polymer, the higher the "induced" piezoelectric effect. Based on this finding, many polymer films with piezoelectricity have been synthesized using polyvinylidene fluoride (PVDF). The PVDF films offer some advantages over ceramic films. The advantages include mechanical flexibility, low mechanical and acoustic impedance, higher resistance to moisture and contaminants, and easy lamination for producing bimorph and multimorph elements.

When a voltage of proper polarity is applied to a sheet of piezoelectric film, the film becomes thinner and elongates, as shown in Fig. 3.4(a). Laminated piezoelectric films can be fabricated. A bimetallic strip of two layers of film with opposite polarity will generate bending motion when a voltage is applied across the thickness, as depicted in Fig. 3.4(b).

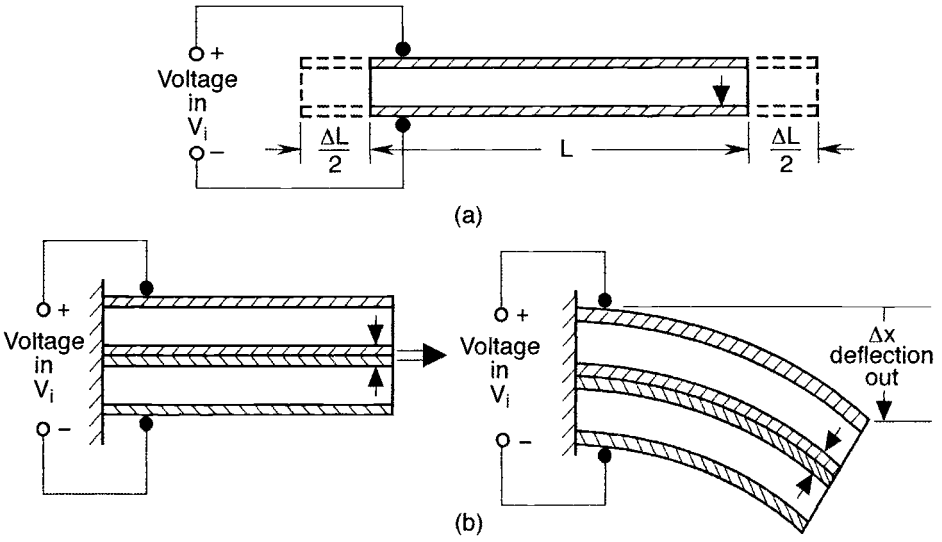


Fig. 3.4. Bimorph cantilever voltage in (V_i) results in deflection out (Δx).

Piezoelectric materials generally exhibit varying degrees of nonlinearity. If Eq. (3.14) is used to determine the stress and displacement fields associated with the applied electric field, then the material constants are a function of applied fields as well as the temperature of the material.

Mukherjee *et al.*[13] have investigated the nonlinear behavior of PZT material as a function of the applied electric field and temperature. They found that there is hystereses between the applied voltage and the induced stress, indicating the piezoelectric coefficients are nonlinear.

The temperature effects of PZT material were investigated by Sherrit *et al.*,[14] who found that the slope in a voltage-versus-time plot increases with temperature. For more detailed information, the readers are referred to the references.

3.5 Failure Theories

3.5.1 General Failure Theories

Failure modes and failure strengths are important parts of structural mechanics. Failure modes identify the types of failure of structural components or elements; while failure strengths represent the ability to resist externally applied loads, including those induced by temperature. Depending on the types of materials, many different failure criteria have been proposed. A few frequently used theories are briefly described below.

3.5.1.1 Maximum Stress Theory

This theory states that a material will fail when the maximum tensile stress or maximum shear stress in a structure reaches a critical value $F_{tu}(F_{ts})$. In this criterion, there is no coupling from the stresses in other directions. It is most applicable to brittle materials. The equations defining this theory are

Tension failure	Shear failure
$\sigma_1 = F_{tu}$, or	$\sigma_1 - \sigma_2 = \pm 2F_{ts}$, or
$\sigma_2 = F_{tu}$, or	$\sigma_2 - \sigma_3 = \pm 2F_{ts}$, or
$\sigma_3 = F_{tu}$	$\sigma_3 - \sigma_1 = \pm 2F_{ts}$.

3.5.1.2 Maximum Strain Theory

This theory assumes that the material failure is controlled by the maximum tensile (shear) strain values $\varepsilon_f(\varepsilon_s)$. Strain criteria are more applicable to materials with large ductility. Their stress-strain curves have a bilinear shape, that is, the curve is composed of two line segments with different slopes, with a relative flat stress-versus-strain slope beyond the yield strength. Their equations are:

Tension failure	Shear failure
$\varepsilon_1 = \varepsilon_f$, or	$\varepsilon_1 - \varepsilon_2 = \pm 2\varepsilon_s$, or
$\varepsilon_2 = \varepsilon_f$, or	$\varepsilon_2 - \varepsilon_3 = \pm 2\varepsilon_s$, or
$\varepsilon_3 = \varepsilon_f$.	$\varepsilon_3 - \varepsilon_1 = \pm 2\varepsilon_s$.

3.5.1.3 Effective Stress Theory (von Mises)

This theory states that a structure will fail when the effective stress,[15] σ_f, as defined below, reaches a critical value, say F_{tu}.

$$\sigma_f = \left[\frac{(\sigma_1-\sigma_2)^2 + (\sigma_2-\sigma_3)^2 + (\sigma_3-\sigma_1)^2 + 3(\sigma_{12})^2 + 3(\sigma_{23})^2 + 3(\sigma_{31})^2}{2}\right]^{\frac{1}{2}} = F_{tu}$$

It is noted that this criterion considers the coupling effect between normal stresses and shear stresses. However, it assumes that a hydrostatic stress field does not contribute to the failure. In fact, this theory works for metallic materials where yielding and failures are caused by the slip in the grain. The slip in the grain is caused by the application of the shear stresses. For a hydrostatically stressed medium, there is no shear stress, which, therefore, will not contribute to the failure.

3.5.2 Statistical Failure Theories

For brittle materials, the failure strength is affected by the presence of defects such as internal porosity and surface cracks. As a result, the measured failure strength values vary within a range depending upon the preparation of the specimens. The failure strength can, therefore, be treated statistically. It is a very useful tool in estimating high confidence/reliability strength values with a reasonable amount of experimental data.

The most popular statistical failure theory is Weibull's two-parameter and three-parameter formulas.[16] They are expressed as:

$$P_s = \exp\left(-K\left(\frac{\sigma}{\sigma_o}\right)^m\right) \text{ (two parameters)}$$

$$P_s = \exp\left(-K\left(\frac{\sigma-\sigma_u}{\sigma_o}\right)^m\right) \text{ (three parameters)} \tag{3.15}$$

where P_s is the probability of survival; σ_u is the threshold stress below which the material does not fail, σ_o is a normalizing stress parameter, and m is the defect-diversity exponent. The constant K can be regarded as a specimen geometry parameter ratio, such as surface area, specimen length, or specimen volume.

Batdorf[17] has generalized Weibull's three-parameter representation of experimental data and made more accurate fitting possible by the use of a Taylor's expansion. He also expanded the representation to consider multiple-direction stress fields. Batdorf and Chang[18] proved that biaxial statistics can be obtained from either a volume-distributed crack theory or a surface-distributed crack theory.

There is no definite preference regarding the choice of failure theories. In general, maximum stress theory is more suitable to more brittle types of materials under mechanical loads; whereas, the maximum strain theory is more suitable when the driving forces come from prescribed strain such as temperature-induced dimensional changes or CTE mismatch. The von Mises criteria are more applicable to conditions where inelastic stress (strain) becomes significant. Statistical theories are generally believed to be applicable to all classes of materials. Their weakness is, however, the requirement of a large data base before a confident criterion can be established for a given material. The appropriate model should be based on the understanding of the material behavior and the driving environments (forces, pressure, temperature, electrical field, etc.).

3.6 Fracture Mechanics

Materials in general have defects that cause reduction in their apparent strength. Many premature failures of bridges, railroads, and pressure mains occurred during the period of the 19th and early 20th century when the effects of defects on material strength were not realized. The development of fracture mechanics did not start until the late 1950s, even though the concept was introduced by Griffith[19,20] in the early 1900s. A brief description of the fracture mechanics theory for brittle materials will be presented here, and some discussion regarding the role of fracture mechanics in the development of MEMS systems will be provided.

3.6.1 Stress at a Crack Tip

We will begin by examining the case of a two-dimensional thin plate of an isotropic material with an elliptical hole. The hole has a major axis diameter of $2a$ and a minor axis diameter of $2b$. The plate is subjected to a uniform uniaxial tension, σ_0, normal to the direction of the crack direction, as depicted in Fig. 3.5. The maximum stress occurs at the end of the major axis in the y direction and is expressed as:[4]

$$\sigma_y = K\sigma_0 = \sigma_0\left(1 + 2\sqrt{\frac{a}{\rho}}\right) \qquad \text{or} \qquad \sigma_y = \sigma_0\left(1 + \frac{2a}{b}\right) \quad \rho = \frac{b^2}{a}, \tag{3.16}$$

where K is the stress concentration factor and ρ is the radius of curvature at the end of the major axis with a value of b^2/a.

When the b/a ratio of the ellipse reduces to zero, the elliptical hole becomes a flat "mathematical crack," and the maximum stress σ_y is unbounded. The stress distribution near the tip of a mathematical crack has been worked out as in Eq. (3.17).[21]

$$\sigma_x = \sigma_o\sqrt{\frac{a}{2r}}\cos\frac{\theta}{2}\left[1 - \sin\frac{\theta}{2}\sin\frac{3\theta}{2}\right] + \text{2nd term} + \text{3rd term} + \ldots$$

$$\sigma_y = \sigma_o\sqrt{\frac{a}{2r}}\cos\frac{\theta}{2}\left[1 + \sin\frac{\theta}{2}\sin\frac{3\theta}{2}\right] + \text{2nd term} + \text{3rd term} + \ldots$$

$$\sigma_{xy} = \sigma_o\sqrt{\frac{a}{2r}}\sin\frac{\theta}{2}\cos\frac{\theta}{2}\cos\frac{3\theta}{2} + \text{2nd term} + \text{3rd term} + \ldots, \tag{3.17}$$

where r is the distance from the crack tip and θ is the angle between the r vector and the major axis. For small r all terms are finite or bounded, but the first term in each equation tends to infinity. Therefore, the stress field at the tip of a crack is of the form $f_{ij}(\theta)\sigma\sqrt{a/(2r)}$, as defined in Eq. (3.17), and we say that the stress field has a singularity of the order of $r^{1/2}$.

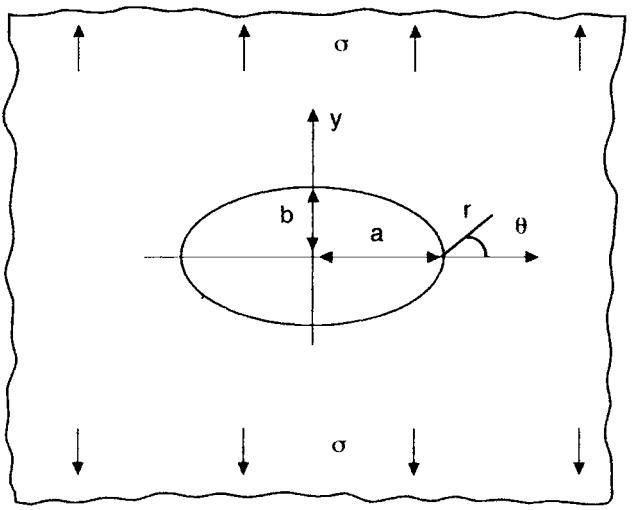


Fig. 3.5. A two-dimensional plate with an elliptical hole (major axis $2a$, minor axis $2b$), subject to a far-field uniaxial tensile stress σ perpendicular to the major axis.

3.6.2 Plane Strain Fracture Toughness

By examining Eq. (3.17), it is seen that the stress field can be expressed as

$$\sigma_{ij} = \frac{K_I}{\sqrt{2\pi r}} f_{ij}(\theta) , \tag{3.18}$$

where K_I is equal to $\sigma_o(\sqrt{\pi a})$ and is called the "stress intensity factor."

The stress intensity factor K_I captures the effects of both magnitude of stress σ_o and the size of the crack $2a$. In other words, K_I can be regarded as a measurement of the magnitude of the stress for given values of r and θ at the vicinity of the crack tip. When a flat panel with a known crack length ($2a$) fails at a stress level σ_f that is below the ultimate tensile strength F_{tu}, then the calculated K_I becomes the "critical stress intensity factor" or "fracture toughness" and is given by:

$$K_{cr} = K_I \text{ (at failure)} = \sigma_f \sqrt{\pi a} \tag{3.19}$$

The mechanics community has recognized that fracture toughness for any alloy is one of the inherent material properties such as Young's modulus and Poisson's ratio. The critical stress σ_{cr} that should cause incipient failure when applied to a flat panel with a crack of length $2a$ can be determined from:

$$\sigma_{cr} = \frac{K_{cr}}{\sqrt{\pi a}} \tag{3.20}$$

Or alternatively, the critical crack size a_{cr} that a flat panel, which is under a uniaxial stress field σ, may have to induce incipient failure is given by:

$$a_{cr} = \frac{1}{\pi}\left(\frac{K_{cr}}{\sigma}\right)^2 \tag{3.21}$$

Therefore, the experimentally determined critical stress intensity factor from one panel can be used to determine the critical stress value or crack size for a different panel. In fact, the critical stress value versus critical crack size for a given critical intensity factor can be visualized from a typical plot, as shown in Fig. 3.6

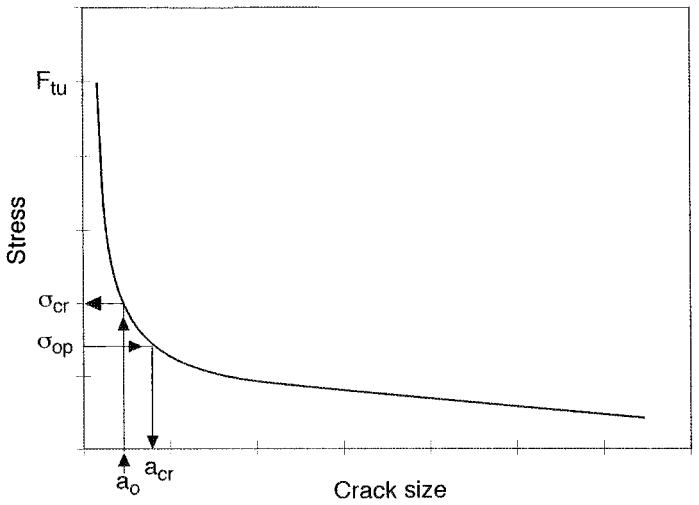


Fig. 3.6. Critical stress and critical crack size, where σ_{op} is the operational stress.

Different structures with different crack geometries and orientations under different loadings will have different expressions for stress intensity factors. However, all the critical stress intensity factors (K_I at failure) associated with different structures, crack geometries, and loadings will be equal to the fracture toughness, which can be determined experimentally using simple specimen geometries and stress fields. Some formulas for stress intensity factor expression are listed in Table 3.3. For complicated structures and loadings, the finite element method, such as that in the ABAQUS computer program, can be performed to determine the stress intensity values.

There are three different modes, as illustrated in Fig. 3.7. Normal stresses give rise to "opening mode," or mode I loading. In-plane stress results in mode II, or "sliding mode." The "tearing mode," or mode III, is caused by out-of-plane shear. Mode I is technically the most important, since most of the stress fields in applications cause mode-I failures.

The critical-stress intensity factor is found to be high for thin specimens. The value decreases as the specimen thickness increases and approaches an asymptotic value. This asymptotic critical stress intensity factor is referred to as "plane-strain fracture toughness, K_{Ic}." The specimen needs to meet a minimum thickness to qualify for a plane-strain requirement. ASTM specification E399

Table 3.3. Some Formulas for Stress Intensity Factors

Type of crack	Applied stress	Mode	Stress Intensity factor
Central, of length 2a in infinite plate	$\sigma_{yy} = \sigma$	I, opening	$K_I = \sigma\sqrt{\pi a}$
	$\sigma_{xy} = \tau$	II, sliding	$K_{II} = \tau\sqrt{\pi a}$
	$\sigma_{xz} = q$	III, tearing	$K_{III} = q\sqrt{\pi a}$
Central, of length $2a$ in plate of width W	$\sigma_{yy} = \sigma$	I, opening	$K_I = \sigma[W\tan(\pi a/W)]^{1/2}$
Central, penny-shaped, of radius a in infinite body	$\sigma_{zz} = \sigma$	opening (radially symmetric)	$K_I = (2/\pi)\sigma\sqrt{\pi a}$
Edge, of length a in semi-infinite plate	$\sigma_{yy} = \sigma$	I, opening	$K_I = 1.12\,\sigma\sqrt{\pi a}$,
	$\sigma_{xz} = q$	III, tearing	$K_{III} = q\sqrt{\pi a}$

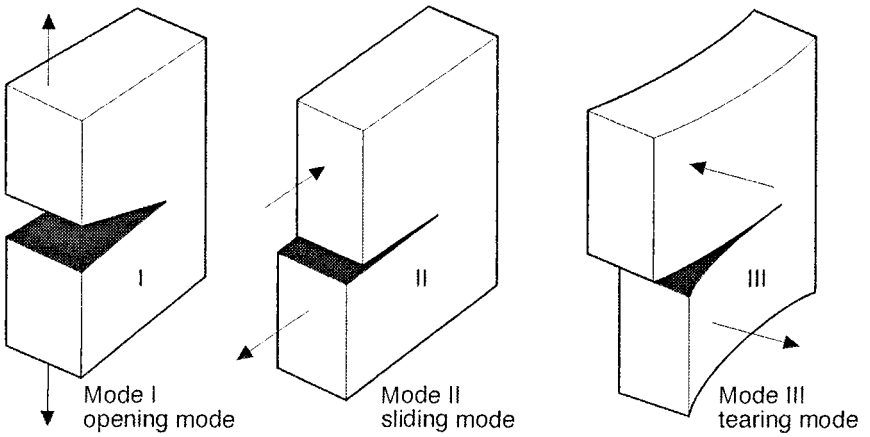


Fig. 3.7. Fracture modes.

describes the specimen configuration and procedure for plane-strain fracture toughness testing and specifies the minimum thickness B as shown in Eq. (3.22).

$$B = 2.5\left(\frac{K_{Ic}}{F_{ty}}\right)^2 \tag{3.22}$$

where F_{ty} is the material yield strength defined in Sec. 3.3.3. Let us take 2219-T87 aluminum as an example. It has a yield strength of 400 MPa and plane-strain fracture toughness of 40.7 $(\text{MPa})\text{m}^{1/2}$. Hence the required minimum specimen thickness for a valid plane-strain fracture toughness test will be 25.8 mm.

When the material thickness of a structure is less than required by Eq. (3.22), the critical stress intensity factor is not the K_{Ic} value. For this case, K_{max} would be more suitable in describing the critical stress intensity factor.

3.6.3 Strain Energy Release Rate

Griffith stated in his criterion for fracturing a body containing a crack that the rate of potential energy loss with respect to the crack length is equal to the surface tension at the plane of the crack extension, shown in Eq. (3.23).

$$\frac{\partial U}{\partial a} = 2\gamma \tag{3.23}$$

where U is the potential energy and γ is the surface tension along the crack extension surface.

The term $\partial U/\partial a$ is defined as the strain energy release rate and is given by the scalar symbol G. It was determined by Griffith that for an infinite plane under a uniform stress field, as discussed previously, G can be expressed as in Eq. (3.24).

$$\begin{aligned} G &= \frac{\sigma^2 \pi a}{E} \text{ (plane stress)} \\ &= \frac{\sigma^2 \pi a(1-\nu^2)}{E} \text{ (plane strain)} \end{aligned} \tag{3.24}$$

At incipient fracture, G attains its critical value G_{Ic}. For a linear elastic plane-strain condition, the relation of G_{Ic} and K_{Ic} is expressed in Eq. (3.25).

$$G_{Ic} = \frac{K_{Ic}^2 (1 - \nu^2)}{E} \tag{3.25}$$

3.6.4 Crack Growth Under Cyclic Loading

It has been known that a subcritical crack, i.e., a crack smaller than the critical size, will grow in size when a structure is under repeated cyclic loading. It was postulated by various investigators that the crack growth per cycle (or crack growth rate, da/dN, where N is number of cycles) is a power law function of the stress intensity range ΔK defined as $(K_{max} - K_{min})$. The K_{max} and K_{min} correspond to the stress intensities at maximum and minimum stresses, thus leading to a simple functional relationship of the form, referred as Paris law:

$$\frac{da}{dN} = C(\Delta K)^n \tag{3.26}$$

where C and n are constants associated with individual materials and are determined from experiments. By a simple integration, the cyclic life N_f for a structure with an initial crack size a_o can be obtained. Therefore, an infinite plate with an initial crack $2a_o$ under uniaxial stress field σ would have a life N_f,

$$N_f = \frac{2}{(2n+1)C(\Delta\sigma)\pi^{n/2}}\left[a_{cr}^{n/2+1} - a_o^{n/2+1}\right] \tag{3.27}$$

where a_{cr} is the critical crack size and $\Delta\sigma$ is the stress range or $\sigma_{max} - \sigma_{min}$.

There are many other crack growth models proposed (see Broek[21]). Table 3.4 illustrates some typical values of K_{Ic}, C, and n for Eq. (3.26). As determined from Eq. (3.27), the dimension of C is in (meters/cycle) / $[(MPa)m^{1/2}]^n$.

Table 3.4. Fracture Toughness Values and Constants for Paris Crack Growth Equation

Material	K_{Ic}MPa-$m^{1/2}$	C	n
2219-T87 aluminum	40.7	0.67×10^{-11}	4
Titanium 6Al-4V Eli	105	0.27×10^{-10}	3.7
Si	0.94	NA	NA
Ge	0.60	NA	NA
GaAs	0.44	NA	NA

3.7 Mechanical Properties of MEMS Structures

Examples of materials now used in MEMS are thin-film polysilicon manufactured by chemical vapor deposition (surface micromachining), single-crystal silicon fabricated into shape by bulk micromachinings, and electrodeposited metals such as nickel produced by the LIGA method in which thick molds are prepared using x-ray exposure. The thin films are on the order of 1 to 10 μm thick; the other two materials can be as thick as 1 mm.

Mechanical properties of materials refer to responses to stress or deformation and are therefore different from physical properties such as density and chemical properties such as corrosion resistance. Like physical and chemical properties, mechanical properties such as modulus should

be independent of the specimen size and shape in order to be true "material" properties. This independence is easily accomplished for measurement of modulus properties where the tension/compression test is the standard, but is not achieved for fatigue and fracture behavior where the specimen's size and shape influence the results.

3.7.1 Elastic Properties

Young's modulus E and Poisson's ratio ν were defined in Sec. 3.3 as the slope of the linear part of a uniaxial stress-strain curve and the negative ratio of transverse to axial strain. In this section we review measurement techniques and present some representative results. More detail is given in Sharpe, Jr., *et al.*[22] and Sharpe, Jr., Yuan, and Edwards.[23]

Young's modulus and Poisson's ratio are important to designers because they determine the amount of deflection from applied forces as long as the material is homogeneous and isotropic. Each can be measured by a so-called inverse approach in which the deflection of a structure is measured and compared with the predictions of an analytical solution or a finite element analysis. The elastic property (in most cases Young's modulus) is then extracted via this comparison. This is an entirely valid approach when the boundary conditions on the structure are met.

The concept of a static beam test to measure E for polysilicon is quite straightforward; one prepares a cantilever beam by surface micromachining and measures the load and deflection at its free end.[24–26] Simple formulas from elementary mechanics of deformable solids enable one to easily compute the modulus. A resonant beam test consists of a beam fabricated so that it is attached to some sort of excitation structure, usually a capacitive comb actuator.[27–30] It is easy to excite the structure over a range of frequencies and also easy to detect the first resonance and extract the modulus from an analytical or numerical solution. Another method that has been developed and refined is the "bulge test" of a thin membrane, in which the specimen material is deflected by pressure on one side and the deflection at the center is measured.[31–33]

Most mechanical properties are obtained from a tensile test. The standards set by the ASTM specify that the specimen be subjected to a uniaxial and uniform stress field and that the strain be measured directly on the specimen with a suitable extensometer.[34] Koskinen[35] tested long thin filaments of polysilicon using crosshead motion as a measure of strain. Read[36] introduced the concept of releasing tensile specimens by etching away the substrate. That approach has been extended by Sharpe *et al.*[22] to polysilicon specimens on which gold lines are deposited to enable the direct measurement of strain.

The formulas for determining the modulus E are:

Cantilever Static Beam	Resonant Frequency ω of a Uniform Weightless Cantilever Beam with Mass M at Tip	Membrane	Tensile Test
$\frac{4PL^3}{\delta bh^3}$	$\frac{4ML^3\omega^2}{bh^3}$	$\frac{p(1-\nu)a^4}{\delta^3 hc(\nu)}$	$\frac{P}{bh\varepsilon}$

where h, b, and L are the thickness, width, and length of the specimen; P and p are the applied force and pressure; M is the effective mass; a is the dimension of a square membrane; and δ and ε are the measured deflection and strain. The function of Poisson's ratio, $c(\nu)$, in the membrane formula depends upon the geometry, and ν must be assumed. The beam and the membrane formulas all involve lengths raised to exponents. This means that these dimensions must be measured with a precision commensurate with the desired results. That may be quite difficult for the small sizes involved in MEMS specimens. The simplicity of the tensile test formula originates in the uniaxiality of the stress and strain states.

Poisson's ratio is difficult to measure by its very nature; it is the ratio of two strains that are small. It can be extracted from tests of two different shapes of membranes,[37] but there are as yet no published results for polysilicon from that approach. Sharpe *et al.*[38] made the first measurements of Poisson's ratio for polysilicon by recording the transverse as well as the axial strains in accordance with the ASTM standard.[39] A typical result is shown in Fig. 3.8, which shows polysilicon to be both linear and brittle.

A more detailed discussion of the variation of mechanical properties of polysilicon appears in Sharpe,[40] so only summary information is given here. Koskinen's measurements of E for three sets of 15 polysilicon fibers each give values of 176, 164, and 164 GPa, with a standard deviation of 25 GPa. Sharpe *et al.*[38] tested 48 specimens from five different production runs of MCNC (Microelectronics Center of North Carolina) MUMPs (multi-user MEMS processes) and measured $E = 169 \pm 6.15$ GPa. Other recent measurements show smaller values, as low as 130 GPa for polysilicon that has experienced various thermal treatments. From the results of 19 tests, the only measurements of Poisson's ratio were $\nu = 0.22 \pm 0.01$.

For initial design calculations only, one can use

$$E = 160 \text{ GPa and } \nu = 0.22$$

for vapor-deposited polysilicon. More appropriate values of E and ν would require that the material be tested using specimens that have experienced the same processing as the microdevice. Metals produced by the LIGA method for MEMS must be tested also and again preferably using specimens that are similar in size to the microdevice. One would expect the elastic properties to be similar to those obtained with bulk specimens, providing the specimen has grains fine enough to justify the assumptions of continuum mechanics. Mazza *et al.*[41] have measured the tensile stress-strain curves of LIGA nickel specimens that are 300 μm long, 20 μm wide, and 120–200 μm thick. They measured strain directly on the specimen with a microscope and an image analysis system. The results from four specimens gave $E = 202$ GPa with excellent repeatability. Sharpe *et al.*[42] have tested nine LIGA nickel specimens, yielding $E = 176 \pm 30$ GPa. The bulk value from the *Metals Handbook*[43] is 207 GPa. Electroplated pure nickel has a very low proportional limit (deviation from linearity), so it is difficult to measure Young's modulus. This is an example of a material with thin-film properties that are not the same as its bulk properties.

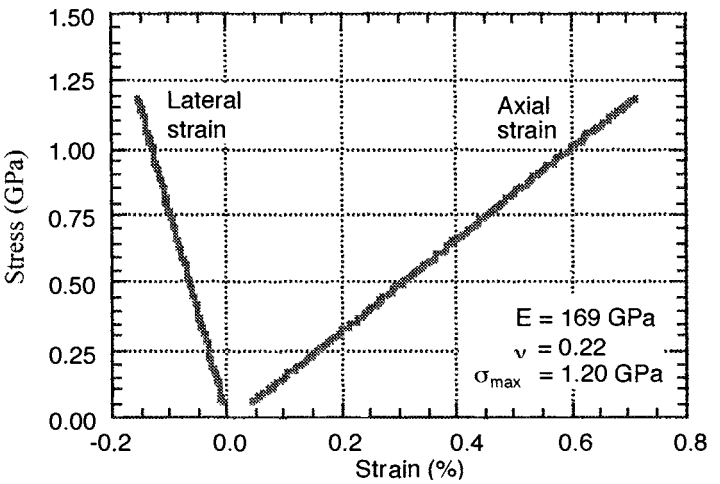


Fig. 3.8. Stress vs biaxial strain for a polysilicon film 3.5 μm thick. (Biaxial test #18.)

It would be wise to test metals manufactured by the LIGA method and also to examine the microstructure using common metallographic techniques. However, for preliminary design calculations only, one can use the elastic properties, E and ν, of the bulk material. However, measurements of properties of the as-manufactured material are required for more accurate predictions.

3.7.2 Strength Properties

The standards for measuring the strength of materials also require that a uniform stress field be imposed upon the specimen;[44] both the elastic and the strength properties can be obtained from a single stress-strain curve, as shown in Fig. 3.2. This of course assumes that the material is isotropic and homogeneous. Strengths measured by other tests such as bending or torsion subject the material to an inhomogeneous stress field and do not provide material properties.

Koskinen[35] measured the tensile strengths of polysilicon filaments that had three different grain sizes and got values ranging between 2.69 and 3.37 GPa. Biebl *et al.*[45] used residual stresses in the polysilicon film to generate forces on a smaller tensile section and measure its strength; their values ranged from 2.11–2.84 GPa. Tsuchiya[46] has used electrostatic force to grasp the free end tensile specimens of various sizes and measured strengths of 2.0–2.7 GPa. Sharpe *et al.*[22] measured strength also (see Fig. 3.8), but their specimens were considerably larger than others and the strengths were even lower at 1.2 GPa. Jones[47] used a bending configuration to measure strength of polysilicon and obtained 1.9 GPa.

A size effect is evident in Fig. 3.9, where the strengths are plotted versus the surface area. It can be argued that a larger specimen has a greater probability of having a fatal surface flaw and the behavior in the figure thereby rationalized. Tsuchiya's work is the only one thus far that undertakes a systematic investigation, and obviously more research is needed. For preliminary design and initial geometry,

$$\text{Tensile strength} = 2 \text{ GPa}$$

can be used for vapor-deposited polysilicon. More reliable predictions would require that the material be tested using specimens that have experienced the same processing and are similar in size to the microdevice.

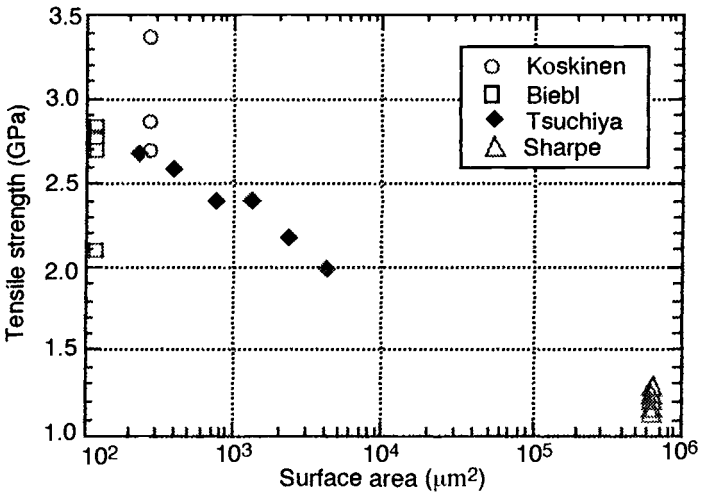


Fig. 3.9. Tensile strengths vs specimen surface area.

LIGA fabricated materials can be expected to have strengths quite different from those of the bulk material; this is because the electrodeposition process forms finer grains. Figure 3.10 is a stress-strain curve of LIGA nickel measured on a specimen with a tensile cross section 200 μm square.[42]

The yield stress for ductile materials is determined by drawing a line parallel to the initial linear region but offset along the strain axis by 0.2%, as described in Sec. 3.3.3. The intersection of this line with the stress-strain curve defines the yield stress. The value for this single test is 315 MPa; the average for nine tests is 323 ± 34 MPa. This is in marked contrast to the handbook value[48] of 59 MPa for bulk pure nickel. Mazza *et al.*[41] obtained an average value of 405 MPa from four tensile tests on specimens that ranged between 120 μm and 200 μm in thickness. The ultimate strength (the maximum stress that the material can support) is correspondingly higher; Mazza reports 782 MPa in contrast to 317 MPa for the bulk material.

Jacobson and Sliwa[49] measured the yield stress of electroplated films to be 410 MPa, with an ultimate strength of 600 MPa. Their films were 150- and 210-μm thick; the gauge length of the tensile specimens was 25 mm.

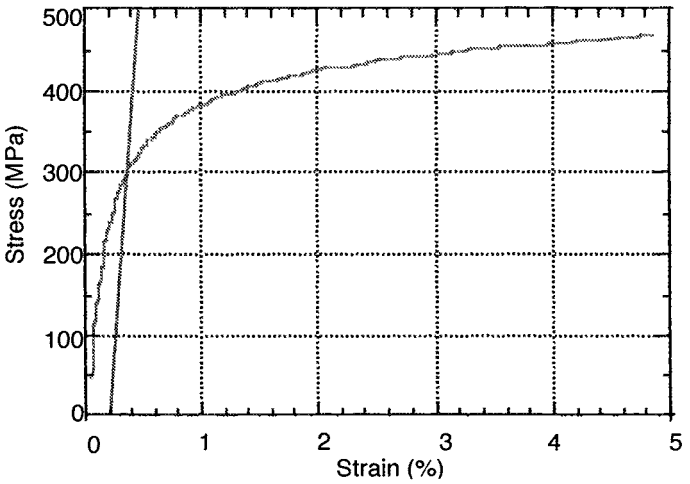


Fig. 3.10. Stress-strain curve of a nickel specimen 200 μm thick fabricated by the LIGA process. The 0.2%-yield stress is 315 MPa.

3.7.3 Fracture of Polysilicon

Given that polysilicon is a brittle material, it is very important to know its fracture toughness. Ductile materials can absorb some deformation at the tip of a crack and therefore allow some margin of error in a design that does not include fracture analysis. Brittle materials do not have this characteristic and can experience sudden and unexpected failure.

Fracture toughness testing is accomplished by preparing a specimen with a geometry that is carefully analyzed to relate the applied forces to the local stress-intensity factor K. The simplest geometry is a wide, long plate loaded on its ends and containing a small center crack that is perpendicular to the loading direction (Table 3.3, center crack, open mode). However, that geometry is too large for most applications, and other geometries along with very carefully stated test procedures have been developed over the years.[50] The critical stress-intensity factor K_{max} is obtained by recording the applied force at which the specimen breaks or at which a plot of force versus crack opening displacement is nonlinear.

An important consideration (at least for metals) is the thickness of the specimen. Thin center-cracked plates show a variation in K_{max} with thickness; whereas thicker plates do not. There is a certain amount of shear at the surface of the plate at the crack tip, at least for materials with some ductility. This tends to be independent of the plate thickness, so if the plate thickness satisfies Eq. (3.22), it has little influence. Next the plane strain-fracture toughness is obtained, which is indeed a material property that is independent of the specimen shape or size. Standard test procedures guarantee that appropriate testing conditions are met. Plain strain fracture is not achievable for thin films, but this may not be an issue for polysilicon, which is brittle.

A difficulty in fracture testing of thin films is generating a crack. Metal specimens are prepared by machining a sharp notch and then precracking to produce an even sharper tip of the crack. Theoretical fracture mechanics assumes an infinitely thin crack, and this condition is achieved in practice; that is, the dimensions of the crack tip are much smaller than any other dimension of the specimen. That condition is difficult to achieve in a thin film; however, Connally and Brown[51] have been able to grow sharp fatigue cracks in polysilicon.

In spite of the difficulties in achieving plane strain conditions and a sharp crack, there are some attempts to measure fracture toughness of MEMS materials. Fan *et al.*[52] estimated fracture toughness of thin silicon-nitride films using an array of surface micromachined structures. Each component of the array had the general shape of an edge-cracked fracture specimen, but each was a different size. When the center portion of the array was released from the substrate, tensile forces were applied to each component because of the residual stresses in the as-deposited film. By estimating the residual stress and observing which cracked structures failed, they were able to bracket a value of fracture toughness. Kahn *et al.*[30] used a sharp probe to pry apart the ends of a long double-cantilever fracture specimen of polysilicon. By measuring displacement of the ends when fracture occurred, they were able to measure fracture toughnesses averaging 2.3 MPa-$m^{1/2}$.

A similar approach is under development, and preliminary results have been obtained. The specimen geometry is shown schematically in Fig. 3.11. A narrow, thin slit is patterned into a polysilicon tensile specimen. The tensile specimen is released by etching away the single crystal substrate, and it is pulled in a small test machine using a linear air bearing to eliminate friction.[22] Two gold pads are deposited across the crack, and the relative displacement between them is measured

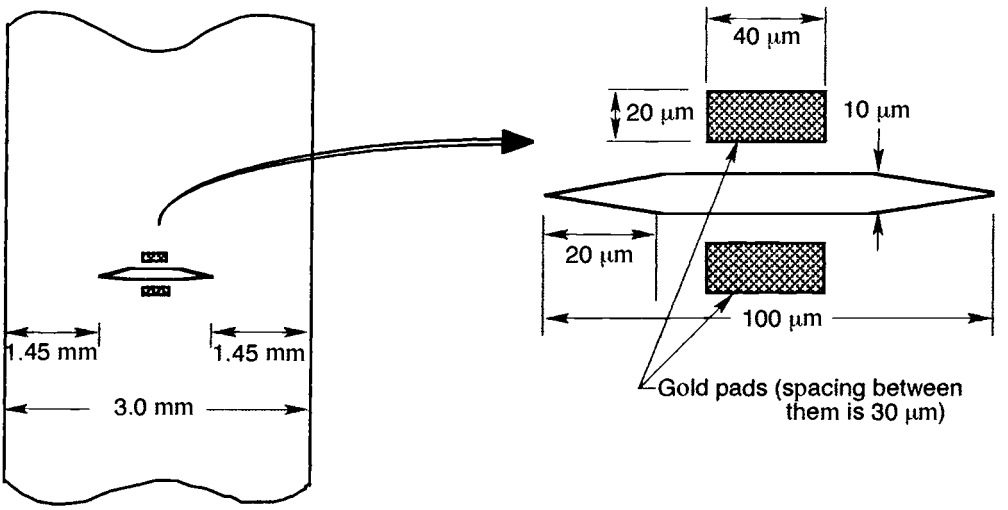


Fig. 3.11. Schematic of center-cracked polysilicon fracture specimen

using laser interferometry. With this approach, one can obtain a plot of force versus crack opening displacement as is done in traditional fracture testing.

The results of 10 tests of polysilicon manufactured at the Microelectronics Center of North Carolina using its MUMPs process are shown in Fig. 3.12. The "theory" line in the figure is the response predicted from linear elastic fracture mechanics (Table 3.3). The agreement between measurement and prediction is remarkable given the fact that the tip of the slit (crack) is not infinitely sharp; it has a radius of curvature of approximately 1 μm. However, this is apparently small enough compared with the overall length of the slit to satisfy the assumptions of the theory.

There is considerable variation in the response of these 10 specimens. Part of that arises from the testing technique that is being developed, which requires very careful alignment of the specimen. The average fracture toughness, computed from the maximum stress attained and the geometry of the specimen, is 1.4 ± 0.15 MPa-m$^{1/2}$.

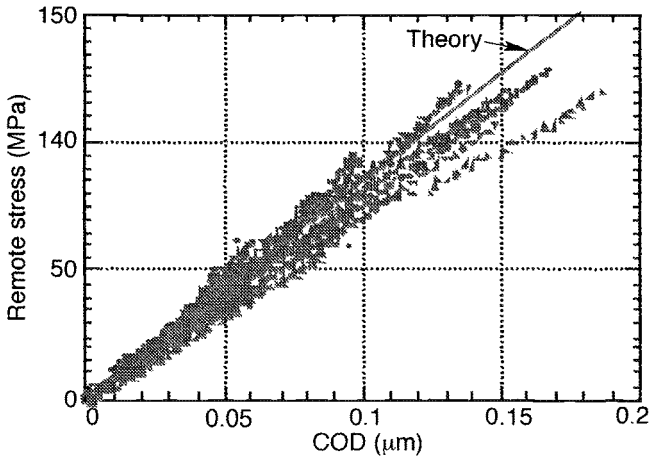


Fig. 3.12. Ten plots of the applied stress vs the opening of the crack in the center of a polysilicon fracture specimen.

3.7.4 Closing Comments

Figure 3.13 shows stress-strain curves for three materials tested; they are presented to compare MEMS materials with a commonly used material. The polysilicon data are from Fig. 3.8, and the LIGA nickel data are from Fig. 3.10. The steel, A533B, is a common pressure vessel material, and the microspecimens used in the tests were the same size as the LIGA nickel ones.[53]

In general, the following are the mechanical properties of materials used in MEMS:

- Vapor-deposited films such as polysilicon have no bulk material counterpart. In other words, one must test the material in the thin-film form to get even an initial estimate of its properties. This requires new test techniques and procedures, which are beginning to emerge.
- Metals, whether deposited by vapor deposition or by the LIGA method, can be expected to have elastic properties similar to the bulk material, but their strengths may be greatly enhanced, as is the case for nickel.

In either case, one should test materials that are produced by the same methods and are similar in size to the microdevices in which they are used. The processes used to manufacture MEMS materials are straightforward, but the dependence of results upon details is not yet established. One can be certain of the values only if the test material is identical to the microdevice material.

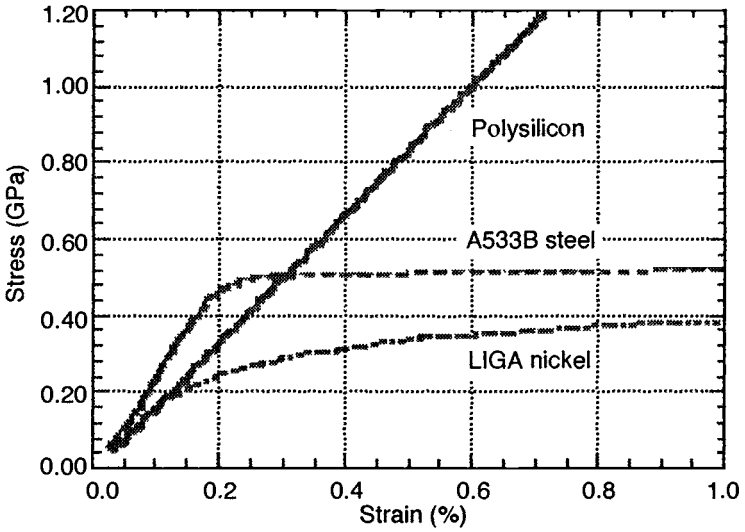


Fig. 3.13. Stress-strain curves for microspecimens of three materials.

3.8 Fatigue

Fatigue loading refers to repetitive applications of external forces to structures or components, and fatigue failure refers to sudden breakage under ordinary use conditions. Sudden failure that occurs for no apparent reason is what makes fatigue so dangerous, but suitable material properties can be measured and design procedures applied to avoid the problem.

A completely different approach is based on fatigue crack growth; a flaw is either assumed or detected, and its growth to a critical size can be predicted. Fatigue loading may be random, periodic, or sinusoidal, as illustrated in Fig. 3.14, and the forces may generate tensile or compressive stresses. In many situations, the stresses vary symmetrically between tension and compression in a sinusoidal fashion; rotating shafts typically experience this behavior. This is called fully reversed loading. The relative amount of maximum force versus minimum force is designated by R, where $R = P_{min}/P_{max}$, and tension is defined as positive with compression as negative. P_{min} and P_{max} are the minimum and maximum loads or forces that are applied. A fully reversed loading has $R = -1$ (the minimum compressive force is equal to the positive tensile force). Tension-tension testing typically has an R value on the order of 0.1; that is, the minimum force is 10% of the maximum force applied. That minimum is usually not zero because the specimen may shift in the grips under zero load.

There are two approaches toward characterizing a material and predicting the fatigue life of a structure or component. The oldest is the stress-life approach, where the fatigue resistance of a material is determined by subjecting samples to harmonic loading with $R = -1$ at various stress levels (from low elastic stresses up to the ultimate tensile strength) and measuring the number of cycles to failure. The strain-life approach is similar to stress-life approach, but the strain range is specified. As applied to metals, the stress-life approach is used to design components that last a very long (infinite) time, while strain-life is used when the number of loading cycles is expected to be only a few thousands.

Fatigue of metals is a relatively mature subject, and there are a number of good references. Dowling's textbook[54] is an excellent introduction not only to fatigue, but also to inelastic deformation and fracture of materials. Fuchs and Stephens wrote an early text[55] on the subject of

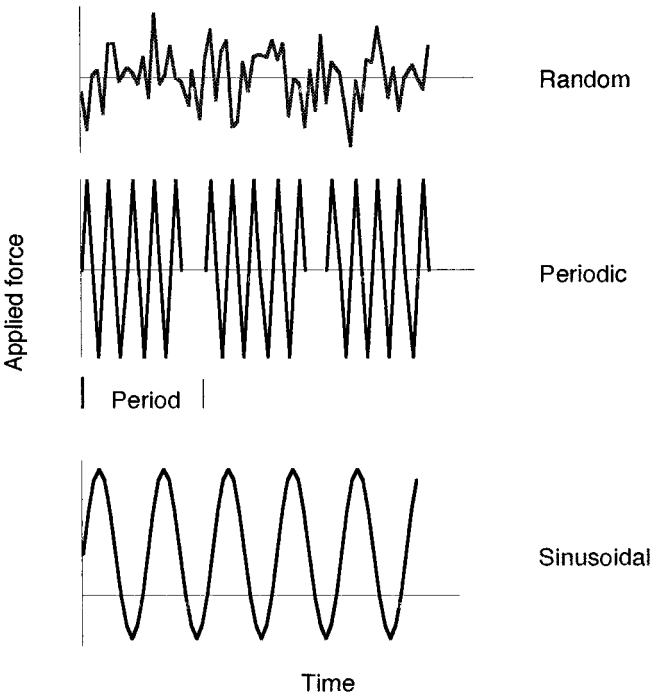


Fig. 3.14. Schematic of different load vs time relations.

fatigue. The monograph by Suresh[56] is a more up-to-date comprehensive view. If one is interested in the fatigue behavior of various metals, the atlas by the American Society of Metals[57] is a good starting place. The fatigue design approach taken by the automobile industry is presented in a handbook published by the Society of Automobile Engineers.[58]

3.8.1 Stress-Life Testing

Just as the simple tension test is used to determine the material properties for quasistatic loading, a simple sinusoidal loading is used to obtain the response of materials to fatigue loading. In early fatigue testing, the specimen had a circular cross section and rotated while subjected to an applied moment. The stress on the specimen surface alternated between tension and compression in a sinusoidal fashion. Such a specimen is subjected to stress gradients across its cross section, and these can influence the results. A "cleaner" stress state is uniaxial tension, and with the development of modern servohydraulic test machines, one can conduct sinusoidal tests at various R-ratios.

The fatigue behavior of a ductile material is shown schematically in Fig. 3.15. The ordinate is the peak stress applied to the specimen, and it is assumed that the loading is fully reversed. The abscissa is the number of loading cycles until a specimen breaks; this is commonly called the "life" or "lifetime" of the specimen. One simply sets the maximum load and frequency of cycling and lets the machine run until failure signals the cycle counter to stop. Each test produces one data point contributing to a so-called S-N curve (stress versus number of cycles).

The ultimate strength of a material is shown at the first cycle. (Actually, this is the first quarter-cycle for $R = -1$ and half cycle for $R = 0$, because the specimen breaks under the peak tensile load.) If the specimen continues to cycle at applied stresses slightly less than the ultimate strength, it fails within 1000 cycles. This Region I is referred to as the low-cycle fatigue regime. If the

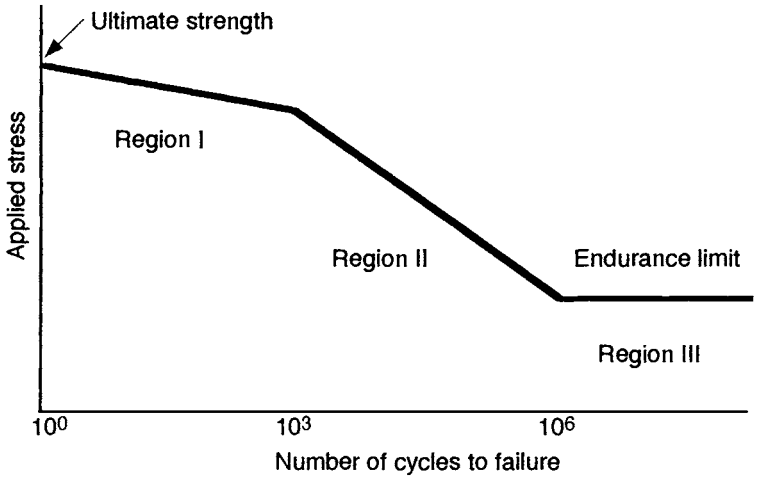


Fig. 3.15. Schematic of the stress-life (S-N) curve for a metal.

maximum stress is decreased further, the specimens last longer in Region II. If the applied stress is reduced below the endurance limit, the material will never fail, as shown in Region III. In general, the material in Region I is undergoing considerable plastic deformation during each cycle, while there is very little such deformation in Region II. Certainly the material in Region III is experiencing only elastic deformation. Fatigue behavior is often divided into low-cycle and high-cycle regimes, with the division set at 10^3 or 10^4.

Fatigue testing is expensive, because many tests must be run to yield a good S-N curve. Fatigue failure is by its nature rather unpredictable, and it is easily possible to get variations of two to five lifetimes of supposedly identical specimens subjected to the same applied stress. Further, a test that runs for many cycles may take a long time on an expensive test machine.

Note that the above description is based on the behavior of metals. There are schemes for constructing the S-N curve of a material using only its static ultimate strength; that permits a very easy and quick design guideline. More elaborate schemes are also used, which bring in other static properties; see Dowling, Chapter 10.[54]

3.8.2 Strain-Life Testing

The advent of servo-hydraulic test machines and strain transducers mounted on the specimen permitted testing using strain as the controlling parameter. That has a certain appeal because in typical components made of ductile metals, failures originate at stress concentrations. If the material at the sharpest point of the stress concentration is loaded into the plastic region, that local volume is subjected to controlled displacements because the overall displacement of the component is determined by elastic behavior of the material surrounding the plastic region. It therefore makes more sense to measure the material behavior as it is exposed to various strain levels.

A schematic of a strain-life plot is given in Fig. 3.16. The curved response can be represented by the addition of two straight lines on the plot, and this leads to a relatively simple equation relating the strain range $\Delta\varepsilon$ to the number of cycles to failure, N_f:[54]

$$\Delta\varepsilon = \frac{\sigma_f}{E}(2N_f)^b + \varepsilon_f(2N_f)^c \tag{3.28}$$

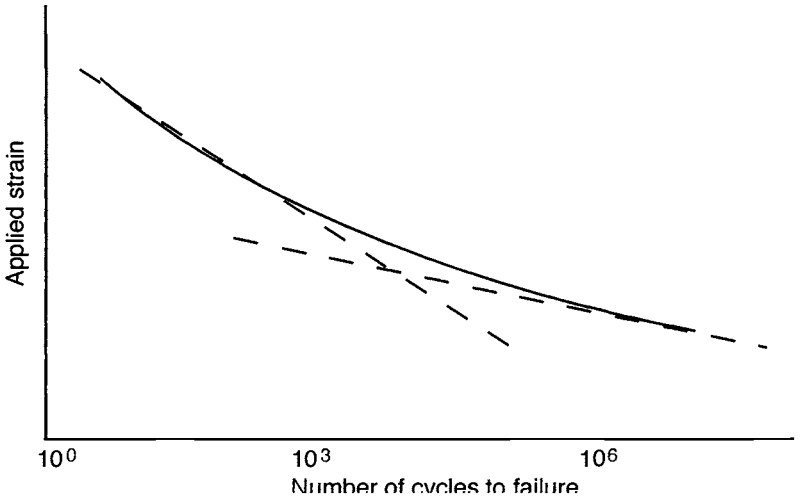


Fig. 3.16. Schematic of strain-life response.

The quantities σ, b, ε_f, c are material dependent. The first term in the equation describes the behavior at short life (where the material responds inelastically), and the second term is the behavior at longer life (where the response is elastic). Considerable testing of metals has produced tables that list the coefficients of Eq. (3.28); see Dowling, Chapter 14.[54]

3.8.3 Fatigue Crack Growth

Figure 3.17 is a schematic of the history of a fatigue crack, plotted as the crack length versus the number of cycles of loading. If a simple specimen such as a thin sheet with a central hole were subjected to alternating tensile loads (e.g., $R = 0.1$), eventually small cracks would appear at the sides of the hole. When these cracks became visible (to the unaided eye or through a microscope), they would have reached a "detection limit" at which one could expect to find cracks through nondestructive inspection. The phase of crack growth up until that time is called "initiation."

Obviously, the history of the crack growth cannot be known before detection, but it can be measured afterward as the crack becomes long enough to cause failure. This growth phase, which can be expressed in da/dn, is the slope of the crack growth curve at a particular number of cycles. The range of applied loads ($\Delta P = P_{max} - P_{min}$) enters through ΔK via the fracture mechanics formulas of Sec. 3.6.4.

The initiation phase is a subject of continuing research. The material local to the stress concentration is undergoing low-cycle fatigue, but the Coffin-Manson equation (Eq. 3.28) predicts the number of cycles to failure, not to a certain crack-detectable length. However, the growth phase has been studied quite thoroughly, and one can find data on many materials, usually in the form of plots of da/dn versus ΔK.

The damage-tolerant approach to structural life prediction assumes that either (a) one can detect a crack in a component in service once it reaches the detection limit or (b) cracks no larger than the detection limit exist in a new component. One then predicts the growth of that crack using material data and the appropriate stress intensity factor for the geometry and loading of the component. Inspections are then scheduled at suitable intervals to monitor the growth of the crack, and the component is replaced before failure occurs. In Fig. 3.17, the crack may or may not be detected upon inspection after the second interval. Even if it were missed, it would be easily found at the end of the third interval in plenty of time for a safe replacement.

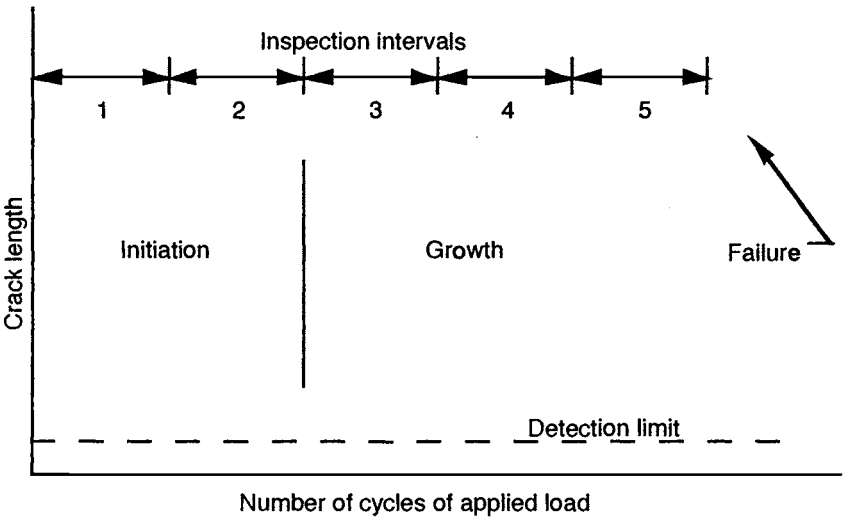


Fig. 3.17. Schematic of fatigue crack growth.

3.8.4 Fatigue of MEMS Materials

There exists a good data base for common structural materials along with established design procedures for construction of components or structures that are subject to fatigue loading. The same cannot be said for materials used in MEMS, since there has been little experimental work. This is basically an open area of research. Given the current and projected applications of MEMS, the focus should probably be on the high-cycle regime.

Connaly and Brown[51] have developed a test structure that is a cantilever beam of single crystal silicon deflected perpendicular to the plane of the die by electrostatic means. A precrack is initiated from nanoindentations, and the crack can actually be observed to grow across the beam to failure. They measure a growth rate of 2×10^{-12} m/cycle.[59] This is slower than the accepted definition of crack arrest in metals, which is 1×10^{-10} m/cycle, but of course a crack in a microdevice does not have to grow as far to become fatal. More recently, Brown and Jansen[60] have developed a similar test structure for in-plane bending of a polysilicon beam.

Mohr and Strohrmann[61] have developed a test technique in which a cantilever beam of LIGA nickel is deflected back and forth by a magnetic field. Their preliminary results (of tests on four specimens) show that LIGA nickel has an S-N curve higher than that obtained from macroscopic specimens. Specimens subjected to maximum stresses of 300 MPa survived over 1 million cycles of loading without failure. Just as LIGA nickel has higher strength properties, it has better fatigue resistance.

Krulevitch *et al.*[62] have subjected Ni-Ti shape memory films to temperature cycling and computed the stress from substrate curvature measurements. They ran up to 2000 cycles and conclude that the applied stress should be kept below 350 MPa for consistent response; this is in the regime of low-cycle fatigue. Extrapolating their data led them to conclude that the applied stress should be limited to 250 MPa for a life of one million cycles. It is interesting to note that this value of 250 MPa is one-half the maximum value of 500 MPa at one cycle; this is the same as the traditional machine design approach,[54] where the endurance limit for steels (stress below which fatigue failure will not occur) is one-half the material's ultimate strength.

3.9 Microstructure of MEMS Materials

As with any material, the microstructure is important in determining the mechanical properties. If the material is isotropic, microstructure details have little effect on the elastic properties but may lead to wide variations in inelastic behavior and strength. Aluminum is a good example; its Young's modulus is approximately 70 GPa regardless of the alloy content or processing, but other parameters, such as constituents and crystallography, have huge effects on its strength. There have been few studies on the microstructure of materials in MEMS.

3.9.1 Microstructure of Polysilicon

Polysilicon produced by low-pressure chemical vapor deposition is by its very nature thin—only a few microns thick. This precludes optical microscopy and requires examination by either a scanning electron microscope (SEM) or a transmission electron microscope (TEM). There have been a few studies to-date (see Kamins[63]), but this is a rich area for research.

Legros *et al.*[64] have studied the microstructure of polysilicon produced by the MCNC MUMPs process. Grain morphology and distribution, texture, dislocation substructure, and microtwinning were examined in undeformed films. Figure 3.18 is a TEM micrograph of the cross section of a polysilicon film. The MUMPs process deposits polysilicon in two layers; intervening layers enable one to manufacture movable microdevices. When the specimens for tensile tests such as shown in Fig. 3.8 were manufactured, the intervening layers were omitted, and the second polysilicon layer deposited onto the first. Figure 3.18 clearly shows these two layers; the first one is 2 μm thick, and the second one is 1.5 μm thick.

A general observation from Fig. 3.18 is that the grains tend to be elongated in a direction perpendicular to the film. In materials science terms, this is referred to as a nonequiaxed (meaning the grains do not have the orientation of the substrate) columnar grain structure. Columnar grains perpendicular to a thin film are common because of the nature of the grain growth process as the material is deposited. When the film is viewed perpendicular to its surface, the grains do not have an elongated shape, but have aspect ratios on the order of one, with dimensions on the order of 0.2–0.4 μm.

3.9.2 Microstructure of Nickel Film

Electrodeposited nickel has a columnar grain structure that is similar in nature to polysilicon, even though it is much thicker. Figure 3.19 is an optical micrograph of a nickel film cross section; in this case, many of the columnar grains extend all the way through the thickness of the specimen.

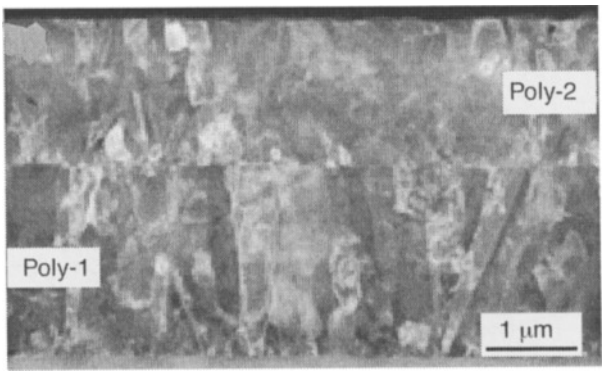


Fig. 3.18. TEM micrograph of cross section of polysilicon films.

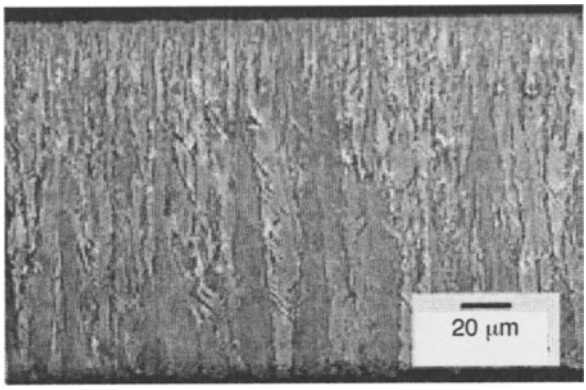


Fig. 3.19. Optical micrograph of cross section of nickel film. The film is 200 μm thick.

When viewed perpendicular to the film surface, the nickel grains have a low aspect ratio (i.e., mostly circular) with a size on the order of microns. This is shown in Fig. 3.20, which is a top view of the same specimen shown in Fig. 3.19.

Although the nickel film pictured here was produced in molds made by the LIGA process, the grain structure shown is typical of nickel films. Betteridge[65] shows a cross section of nickel that is similar to Fig. 3.19. Sard and Weil[66] show a cross section of electroplated copper that is also similar.

3.9.3 Closing Comments

There is an obvious need for more extensive studies of the microstructure of materials now used in MEMS and those under development. By the very nature of the manufacturing process, the microstructure is likely to be quite different than expected for bulk materials. Since the specimens for study will be small and thin, special techniques and procedures will be required in some cases.

Studies thus far show that MEMS materials are not isotropic. In fact, these thin materials are transversely isotropic; that is, they have the same properties in any direction in the plane of the film, but different properties measured perpendicular to the plane. This occurs because fine grains nucleate on the surface of the substrate when the deposition begins (whether vapor deposition or electrodeposition). As material is continuously added, grains with a preferential orientation will grow faster and larger, which leads to the columnar structure.

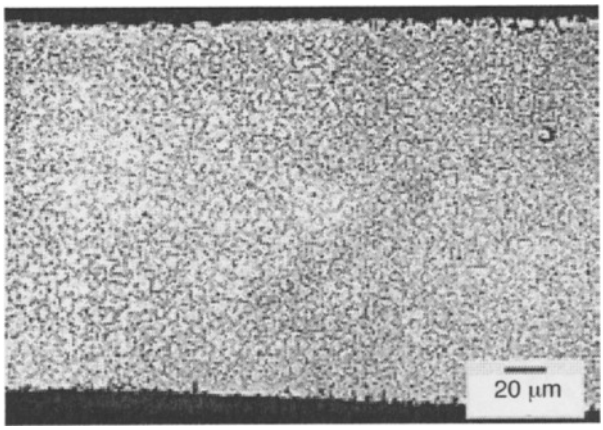


Fig. 3.20. Top view of a nickel specimen.

This transverse isotropy can be important; the stiffness of a beam bending in the plane of the film can be different from that of a beam bending out of the plane. Components of MEMS accelerometers bend in the plane of the film, while components of pressure transducers bend out of the plane. A more complete characterization of the mechanical properties of the material would account for this difference and would involve measuring more than just two elastic constants. Such a study has not been conducted.

3.10 Other MEMS-Related Subjects

The basic principles of some MEMS devices are briefly described, and determination of thin-film residual stress is discussed. The mechanics of thin films is part of the MEMS system, but since information on that subject is widely available through the open literature, only the residual stress aspect is presented here.

3.10.1 Accelerometers

Accelerometers are probably the most popular MEMS devices. They measure the acceleration at a particular location in a structural system. They can be used to characterize the acceleration, motion, and level of shock of a system. Currently, the largest application is for the automobile industry. For example, accelerometers are used to control the activation of air bags. There are many designs for accelerometers, but the basic principle is a spring-mass system. Under an acceleration a, the mass m exerts a force $F = ma$, which deforms the spring system. There are two ways to measure the deformation. The first is the measurement of deformation; while the second is to measure the strain associated with the stress generated by the force F.

For example, consider a cantilever beam of length L, width b, and thickness h. A proof mass M is attached at the tip of the cantilever. The material has a Young's modulus E. Under an acceleration a, the lateral deformation at the tip of the cantilever is given by $(Ma)L^3/(3EI)$, where I is the area moment of inertia of the cross section. This deformation can be measured using the capacitance technique.

Using the strain gauge concept, one can measure the strain values at the fixed end when an acceleration is applied. The theoretical strain at the tension side of the cantilever is $6(Ma)L/E(bh^2)$. There are many different accelerometer designs to achieve high g, high sensitivity. Currently, 100-g and 100,000-g accelerometer devices are being designed by the Charles Draper Laboratory.[67]

3.10.2 Pressure Transducer

The pressure transducer is used to measure either the absolute or the gauge pressure across a surface. In many cases, a thin plate of circular shape is used. When one side of the plate is pressurized, it deforms. For an isotropic material, the relation of the center deformation of the thin plate and the hydrostatic pressure is expressed as

$$P_0 = \frac{64D}{R^3}\left(\frac{w_{\max}}{R}\right) + \frac{8}{3}\frac{E}{1-\nu}\frac{t}{R}\left(\frac{w_{\max}}{R}\right)^3 \tag{3.29}$$

where $D = Et^3/[12(1-\nu^2)]$, E is the Young's modulus, ν is the Poisson's ratio, t is the plate thickness, R is the plate radius, and $w_{\max}$ is the displacement at the center of the plate. The amount of the $w_{\max}$ can be measured by the capacitance technique.

3.10.3 Residual Stress in Thin Films

Almost all MEMS devices use thin-film coating in some way. Therefore, the determination of residual stresses in thin films as a result of deposition is important. During the course of film deposition, the substrate undergoes two types of deformation: thermally induced and nonthermally induced. The former is caused by the temperature rise and the temperature gradient associated with the film deposition; while the latter is caused by the atomic interaction between the film and substrate materials. In general, the film-thickness-to-substrate-thickness ratio and the film-thickness-to-lateral-dimension ratio are so small that:

$$\frac{t_f}{t_s} << 1, \frac{t_f}{l} << 1 \tag{3.30}$$

where t is the thickness; subscripts f and s refer to film and substrate, respectively; l is the lateral dimension.

The intrinsic stress, S, in a thin film for a cantilever of length l, a simply supported beam of length l, or a simply supported disk of diameter l, is expressed by the famous modified Stoney equation:

$$S = \frac{Et_s^2\delta}{3(1-\nu)t_f l^2} \tag{3.31}$$

where E and ν are the Young's modulus and Poisson's ratio of the substrate material, and δ is the deflection for the following cases:

1. The end deflection δ_c of a cantilever ($\delta = \delta_c$)
2. Four times the deflection δ_s between the center and the support of a simply supported beam ($\delta = 4\delta_s$)
3. Four times the deflection δ_{sd} between the center and the support of a simply supported disk ($\delta = 4\delta_{sd}$)

For example, consider a 0.1-μm-thick gold film deposited to a 0.279-mm-thick silicon cantilever substrate. Let the length of the substrate be 10 mm. The Young's modulus and Poisson's ratio for the substrate material are 170 GPa and 0.22, respectively. If the measured tip deflection of the cantilever is 1 μm between precoated and postcoated substrate, then using Eq. (3.31) the calculated residual stress in the gold film would be 565 MPa (82 ksi).

3.11 Examples

The following examples of MEMS devices illustrate various states of stress. Most of these microdevices are planar in shape, so the calculation of stresses is actually not very difficult. This is not intended to be a survey, but instructional.

3.11.1 In-Plane Bending

One of the most widely used MEMS devices is the accelerometer based on the in-plane motion of a polysilicon mass with flexural arms whose motion is sensed capacitively. Figure 3.21 is an SEM photo of a portion of the sensing element of an accelerometer made by Analog Devices. The entire polysilicon element is fabricated by surface micromachining, and the mass in the center is supported at the top and bottom by long, narrow beams. Motion along the vertical axis is sensed by the symmetric series of interdigitated fingers; the capacitance changes as the distance between fingers changes.

Stress in the supporting elements can be calculated from simple beam theory. An estimate of the maximum stress in the flexural element can be made by considering the largest beam-tip displacement allowed. The element is 2 μm thick and 1.8 μm wide. The maximum deflection of one of the long supporting elements shown across the bottom of the photo is estimated to be 0.5 μm.

Using the simple formula for a cantilever beam, $\delta = PL^3/3\ E\ I$, where $L = 130$ μm, $E = 165$ Gpa, and $I = 9.7 \times 10^{-25}$ m^4, one can estimate $P = 1.1 \times 10^{-7}$ N. From $\sigma = PL\ c\ /\ I$, one estimates the maximum stress as 13 MPa.

3.11.2 Out-of-Plane Bending

The Scratch Drive Actuator[68] is a clever application illustrating out-of-plane bending. The polysilicon plate and bushing, which are conductive, of Fig. 3.22 are fabricated by surface micromachining. When an electrostatic force is applied between the plate and the substrate, the plate bends and the bushing moves forward slightly. When the voltage is released, the friction at the tip of the bushing pulls the component and its attachments forward in an "inchworm" fashion.

One can estimate the maximum stress in the plate by guessing the minimum radius of curvature of the plate when the maximum voltage is applied. Using elementary plate theory, the maximum stress can be calculated as

$$M = \frac{EIh}{6(1-\nu^2)(l-\lambda)^2}\ \sigma_{max} = \frac{Eh^2}{12(1-\nu^2)(l-\lambda)^2}, \tag{3.32}$$

where M and σ_{max} are the plate maximum bending moment and maximum bending stress, respectively; E and ν are the material modulus and Poisson's ratio, and λ is the portion of length l that is pulled down by the electrostatic force. For the case in Fig. 3.22 with a length of $(l - \lambda)$ equal to 10 μm, the calculated σ_{max} is about 212 MPa (31 ksi).

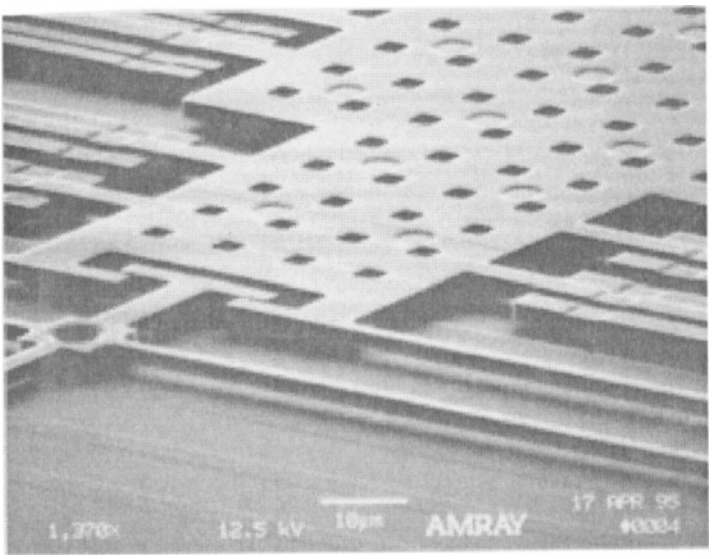


Fig. 3.21. Accelerometer sensing element. (Courtesy of John Yasaitis, Analog Devices)

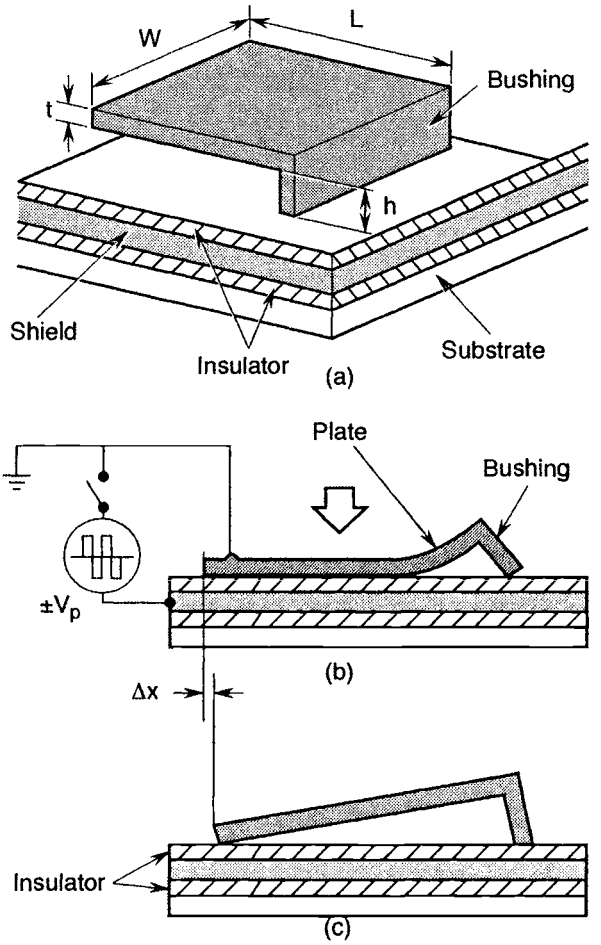


Fig. 3.22. Schematic view of a scratch drive actuator. (a) Dimensions for the self-assembling are L = 50 μm, W = 75 μm, t = 1 μm, and h = 1.5 μm; (b) and (c) model for the step motion of the SDA showing the evolution of the plate deformation according to an applied pulse. (© 1993 IEEE)

3.11.3 Torsion

An example of a polysilicon element subjected to torsion is shown in Fig. 3.23 (from Judy and Muller[69]). A nickel micromirror is deposited onto a polysilicon component; it is lifted off the substrate by a magnetic field interacting with the ferromagnetic nickel. When a voltage is applied between the polysilicon component and the ground plane, the mirror snaps down flat. The polysilicon torsion rod supports the mirror.

The determination of maximum shear stress of the polysilicon rod element is briefly described. Consider a rod with uniform rectangular cross section, width b, thickness a, and length l, where l is much bigger than either a or b. The rod is fixed at one end and free of constraint at the other end. A torsion of magnitude T is applied at the free end in the plane normal to the geometrical axis of the rod. Suppose $b \geq a$ and the torsion-applied axis coincides with the geometrical axis of the) rod; then the relations between torsion, the geometrical dimensions of the rod, and the material parameters of the material are expressed in Eq. (3.33).

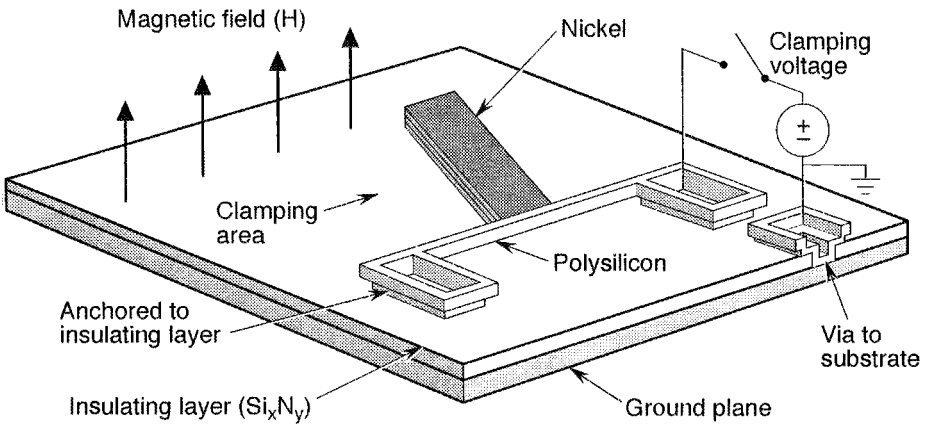


Fig. 3.23. Schematic of a magnetically and electrostatically actuated micromirror. (© IEEE 1997).

$$T \cong \frac{G\theta}{l}\left[\frac{ba^3}{3} - \frac{64a^4}{\pi^5}\tanh\left(\frac{\pi b}{2a}\right)\right]$$

$$\tau_{max} \cong \frac{G\theta a}{l}\left[1 - \frac{8}{\pi^2}\operatorname{sech}\left(\frac{\pi b}{2a}\right)\right] \qquad \text{or}$$

$$\tau_{max} \cong T\frac{1 - \frac{8}{\pi^2}\operatorname{sech}\left(\frac{\pi b}{2a}\right)}{\frac{ba^3}{3} - \frac{64a^4}{\pi^5}\tanh\left(\frac{\pi b}{2a}\right)} \tag{3.33}$$

where G is the material shear modulus, θ/l is the angular twist per unit length, sech and tanh are the hyperbolic functions. The τ_{max} occurs at the midpoint of the longer side. For large b/a ratio, the τ_{max} is simply $3T/ba^2$. It should be noted that Eq. (3.33) is much more complicated than that shown in Sec. 3.3.5.4 because the cross section here is rectangular rather than circular.

For a rod of length l with both ends fixed and a torsion T applied at a distance x from one end, then different magnitudes of torsion are resisted by two different segments of the rod. The magnitude of the torsion in the segment with length x is $T(l-x)/l$, and the magnitude of the torsion in the other segment is Tx/l.

3.11.4 Biaxial Tension

Membranes experience biaxial tension, and an example of a microdevice using a silicone-rubber membrane as an actuator is shown in Fig. 3.24 (from Yang *et al.*[70]). The relatively thick silicone membrane expands upward to close the valve inlet and outlet when the working fluid below it is heated with a resistive heating element.

The rectangular silicone membrane is 1.5 × 2.5 mm with a uniform film thickness of 50 μm. All four edges are clamped against rotation. The maximum stress, σ_{max},occurs at the clamped edge with a magnitude of[71]

$$\sigma_{max} = \beta E\frac{w_o h}{a^2} \tag{3.34}$$

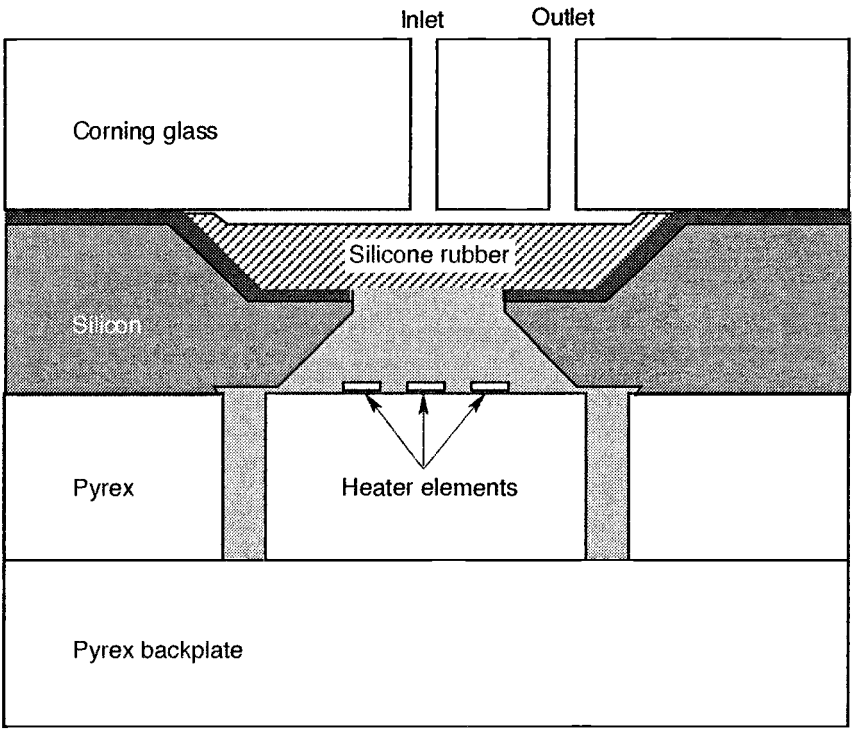


Fig. 3.24. Schematic of a membrane-actuated valve. (© IEEE1 1998)

where w_o is the maximum normal deflection of the film, h is the film thickness, and a is the smaller of the two rectangular edges. Assume the applied pressure is 0.79 MPa (100 psi); then the magnitude of the maximum stress is 830 MPa (120 ksi).

3.11.5 Three-Dimensional Stress States

Not all components of microdevices are as simple as the four examples above. In order to transmit larger forces and torques, microdevices must be thicker, and this is the basis of the intense interest in LIGA (or its equivalent) fabricated structures. An example of such a thicker component is the polymer gear shown in Fig. 3.25 (from Lorenz *et al.*[72]).

Typically one would use the traditional approaches of machine design to estimate the torque that could be transmitted by such a gear. If the component is not a standard one or has a more complicated structure, then one would need to use finite element methods to determine its strength and stiffness.

3.12 Future Challenges

MEMS is a new and revolutionary field that is rapidly changing the way we perceive and sense the world. There are also many challenges ahead. Although numerous devices are being developed and used for certain applications, the reliability of many MEMS systems has not been addressed. For example, solid mechanics is yet to be introduced and applied in the MEMS community to enhance integrity of the MEMS devices.

Many mechanics-related issues face the MEMS community. First is an urgent need to understand the materials, including fabrication technique, material homogeneity, and material properties. Different fabrication techniques will control the material homogeneity and material

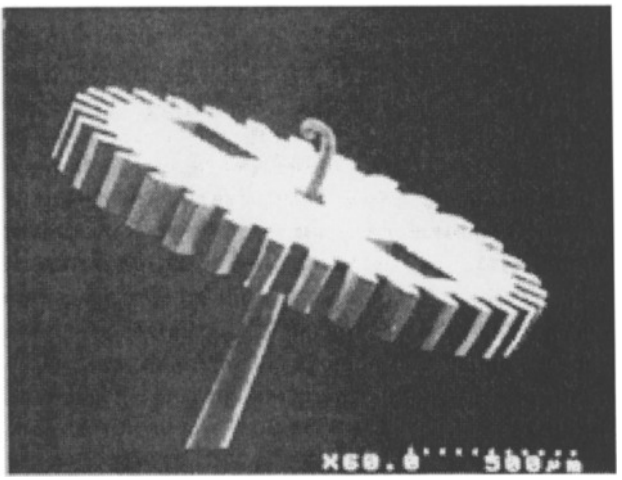


Fig. 3.25. SEM image of a polymer gear. (Courtesy of IEEE, © 1998)

properties. In turn, material homogeneity and properties control the stress field under external environment and affect service life.

A second issue is the residual stresses generated from fabrication processes. Depending upon fabricating techniques, residual stresses in a MEMS layer can be on the order of several hundred mega pascals. In microelectronic interconnects, high residual stress up to 1000 MPa in 300-nm-thick aluminum film was reported.[73] Such a high residual tensile stress is definitely detrimental to the desirable service life of devices.

The third and most important issue involves acquiring a better understanding of the mechanics at the micro- and nanodimension levels. Classical continuum mechanics is based on the assumption that materials are homogeneous. When the grain size is much smaller than the dimension of the structures, continuum mechanics gives very accurate answers. When localized defects or cracks are present, fracture mechanics is sufficient to describe the stress behavior at the neighborhood of the crack tip. But when the grain size becomes comparable to the dimensions of the devices, as occurs with MEMS structures, the validity or applicability of classical solid mechanics and fracture mechanics becomes questionable. Therefore, the current theories on material constitutive relations, failure mechanisms, and fatigue behaviors need to be revisited.

Many areas of materials research are needed in MEMS/nanotechnology. Following is a list of subjects for near- and long-term research.

Areas for near-term research:

- Investigate material property measurement methods for micro-sized MEMS components and coatings.
- Develop nondestructive stress measurement techniques at the device levels.
- Generate models for stress migration in metal lines and around contact/vias.
- Develop models for better understanding of electromigration as a function of line sizes and current level.

Areas for long-term research:

- Address the validity of current solid mechanics and modeling at the MEMS level.
- Address the validity of current fracture mechanics at the MEMS level.
- Address the failure mechanisms and failure theories of thin metal lines at the micro level.
- Develop reliability improvement models and algorithms.

At a Mechanics of MEMS Workshop convened by the Air Force Office of Scientific Research, representatives from government agencies, universities, and industry agreed that in the next 5 years, the mechanics-related work should emphasize the following subjects.

- Property measurements
- Crack propagation/initiation
- Stress modeling
- Fracture toughness measurements
- Processing techniques development and standardization
- Interface investigation

To bring mechanics into MEMS technology and ensure reliability of MEMS components, considerable basic research and database generation must be done. Thus it seems only appropriate that the Government take the initiative to fund the investigations of these subjects. Only with a better understanding of the principles of the mechanics and sufficient design data can MEMS achieve the desired objectives.

3.13 References

1. H. Helvajian, ed., *Microengineering Technology for Space Systems*, monograph 97-02, (El Segundo, Calif., The Aerospace Press, 1997), pp. 143–201. First published as The Aerospace Corp., Report no. ATR-95 (8168)-2 (1995).
2. A. E. H. Love, *A Treatise on the Mathematical Theory of Elasticity* (Dover, New York, 1944).
3. I. S. Sokolnikoff, *Mathematical Theory of Elasticity* (McGraw-Hill, New York, 1956).
4. S. Timishenko and J. N. Goodier, *Theory of Elasticity*, 2nd ed. (McGraw-Hill, New York, 1951).
5. R. D. Mindlin and F. F. Tiersten, "Effects of Couple-Stresses in Linear Elasticity," *Archive for Rational Mechanics and Analysis* **11** (5), 415–448 (1962).
6. A. C. Erigen, "Theory of Micropolar Elasticity," in *Fracture—An Advanced Treatise*, vol. II, edited by H. Liebowitz (Academic Press, New York, 1968).
7. B. A. Boley and J. H. Weiner, *Theory of Thermal Stresses* (Wiley, New York, 1970), p. 430.
8. R. C Hibbeler, *Mechanics of Materials*, 2nd ed. (Prentice Hall, Englewood Cliffs, N.J., 1994).
9. F. P. Beer, *Mechanics of Materials*, 2nd ed. (McGraw-Hill, New York, 1992).
10. R. J. Roark, *Roark's Formulas for Stress and Strain*, 6th ed. (McGraw-Hill, New York, 1989).
11. "Standards on Piezoelectric Crystals, 1949," *Proceedings of the IRE, Standards Committee*, 1949–1950: *Piezoelectric Crystals Committee*, 1947–1950.
12. N. M. Kocharyan, Kh. B. Pachadzhyam, and Zh. Tivriktsyam, *Dokl. Akad. Nauk. Ann. SSR* 44 (3), 111 (1967).
13. B. K. Mukherjee and S. Sherrit, "Characterization of Piezoelectric and Electrostrictive Materials for Acoustic Transducers: II. Quasistatic Methods," *Fifth International Congress on Sound and Vibration* (Adelaide, Australia, 15–18 December 1997).
14. S. Sherrit *et al.*, "Domain Wall Motion in Piezoelectric Materials Under Stress," *Proceedings of the Fifth IEEE International Symposium on Application of Ferroelectrics*, (Greenville, SC, 30 August–2 September 1992).
15. T. H. Lin, *Theory of Inelastic Structures* (Wiley, New York, 1968), p. 109.
16. W. Weibull, "A Statistical Theory of the Strength of Materials," *Ing. Vetenskaps Akad.*, 151 (1939).
17. S. B. Batdorf and J. G. Crose, *A Statistical Theory for the Fracture of Brittle Structures Subjected to Nonuniform Polyaxial Stresses*, The Aerospace Corp. Report no. TR-0073(3450-76)-2 (1973).
18. S. B. Batdorf and D. J. Chang, "On the Relation Between the Fracture Statistics of Volume Distributed and Surface Distributed Cracks," *Int. J. of Fracture* **15** (2) (1979).
19. A. A. Griffith, "The Phenomena of Rupture and Flow in Solids," *Phil. Trans. Roy. Soc. of London* **A221**, 1163–1167 (1921).

20. A. A. Griffith, "The Theory of Rupture," *Proceedings of First International Congress on Applied Mechanics* (1924), pp. 55–63.
21. D. Broek, *Elementary Engineering Fracture Mechanics*, 4th rev. ed. (Martinus Nijhoff Publishers, 1986).
22. W. N. Sharpe, Jr., B. Yuan, R Vaidyanathan, and R. L Edwards, "New Test Structures and Techniques for Measurement of Mechanical Properties of MEMS Materials," *Proceedings of the SPIE Symposium on Microlithography and Metrology in Micromachining II* (Austin, TX, 1996), pp. 78–91.
23. W. N. Sharpe, Jr., B.Yuan, and R. L. Edwards, "Variations in Mechanical Properties of Polysilicon," *Proceedings of the 43rd International. Symposium of the Instrumentation Society of America* (Orlando, FL, 1997), pp. 179–188.
24. T. P. Weihs, S. Hong, J.C. Bravman, and W.D. Nix, "Mechanical Deflection of Cantilever Microbeams: A New Technique for Testing the Mechanical Properties of Thin Films," *J. Mater. Res.*, **3** (5), 931–942 (1988).
25. P. Ruther, W. Balcher, K. Feit, D. Maas, W. Menz, 1995, "Microtesting System Made by the LIGA Process to Measure the Young's Modulus in Cantilever Microbeams," *Proceedings of the ASME Dynamic Systems and Control Division*, DSC-Vol 57-2, pp. 963–967.
26. M. Biebl, T. Scheiter, C. Hierold, H. V. Philipsborn, and H. Klose, "Micromechanics Compatible with an 0.8 mm CMOS Process," *Sensors and Actuators A* **46–47**, 593–597 (1995).
27. M. Biebl, G. Brandl, and R. T. Howe, "Young's Modulus of In Situ Phosphorus-Doped Polysilicon," *Transducers '95-Eurosensors IX, Proceedings of the 8th International Conference. on Solid-State Sensors and Actuators. and Eurosensors IX* (Stockholm, Sweden, June 1995), pp. 80–83.
28. L. Kiesewetter, J.M. Zhang, D. Houdeau, and A. Steckenborn, "Determination of Young's Moduli of Micromechanical Thin Films Using the Resonance Method," *Sensors and Actuators A* **35**, 153–159 (1992).
29. S. Roy, S. Furukawa, H. Miyajima, and M. Mehregany, "In situ Measurements of Young's Modulus and Residual Stress of Thin Electroless Nickel Films for MEMS Applications," *Materials Research Society Symposium Proceedings* (1995), vol. 356, pp. 573–578.
30. H. Kahn, S. Stemmer, K. Nandakumar, A. H. Heuer, R. L. Mullen, R. Ballarini and M. A. Huff, "Mechanical Properties of Thick, Surface Micromachined Polysilicon Films," *Proceedings of the IEEE 9th Annual International Workshop on Micro Electro Mechanical Systems* (San Diego, CA, 1995), pp. 343–348 (1996).
31. S. Hong, T. P. Weihs, J. C. Bravman, and W. D. Nix, "Measuring Stiffnesses and Residual Stresses of Silicon Nitride Thin Films," *J. Electron. Mater.* **19**, 903–909 (1990).
32. V. M. Paviot, J. J. Vlassak, and W. D. Nix, "Measuring the Mechanical Properties of Thin Metal Films by Means of Bulge Testing of Micromachined Windows," *Materials Research. Society Symposium Proceedings* (1995), vol. 356, pp. 579–584.
33. L. Tong, M. Mehregany, and L. G. Matus, "Mechanical Properties of 3C Silicon Carbide," *Appl. Phys. Lett.* **60** (24), 2992–2994 (1992).
34. ASTM Standard E 111-82, "Standard Test Method for Young's Modulus, Tangent Modulus, and Chord Modulus," *1995 Annual Book of ASTM Standards, Sec. 3, Metal Test Methods and Analytical Procedures*, vol. 03.01 (American Society for Testing and Materials, West Conshohocken, Penn., 1995).
35. J. Koskinen, J. E. Steinwall, R. Soave, and H. H. Johnson, "Microtensile Testing of Free-Standing Polysilicon Fibers of Various Grain Sizes," *J. Micromechanics and Microengineering* **35**, 13–17 (1993).
36. D. T. Read and J. W. Dally, "Mechanical Behavior of Aluminum and Copper Thin Films," *Mechanics and Materials for Electronic Packaging: Volume 2 - Thermal and Mechanical Behavior and Modeling, ASME Proceedings* (1994) AMD, vol. 87, pp. 41–49.
37. J. J. Vlassak and W. D. Nix, "A new bulge test technique for the determination of Young's modulus and Poisson's ratio of thin films," *J. Mater. Res.* **7** (12) 3242–3249 (1992).

38. W. N. Sharpe, Jr., B. Yuan, R. L. Edwards, and R. Vaidyanathan, "Measurements of Young's Modulus, Poisson's Ratio, and Tensile Strength of Polysilicon," *Proceedings of the Tenth IEEE International Workshop on Microelectromechanical Systems* (Nagoya, Japan, 1997), pp. 529–534.
39. ASTM Standard E 132 - 86, "Standard Test Method for Poisson's Ratio at Room Temperature," *1995 Annual Book of ASTM Standards, Sec. 3, Metal Test Methods and Analytical Procedures*, vol. 03.01 (American Society for Testing and Materials, West Conshohocken, Penn., 1995).
40. W. N. Sharpe, Jr., B. Yuan, and R. L. Edwards, "Variations in Mechanical Properties of Polysilicon," *Proceedings of the 43rd International. Symposium of the Instrumentation Society of America* (Orlando, FL, 1997), pp. 179–188.
41. E. Mazza, S. Abel, and J. Dual, "Experimental Determination of Mechanical Properties of Ni and Ni-Fe Microbars," *Microsystem Technologies* **2** (4), 197–202 (1997).
42. W. N. Sharpe, Jr., D. A. LaVan, and R. L. Edwards, "Mechanical Properties of LIGA-Deposited Nickel for MEMS Transducers," *Proceedings Transducers '97* (Chicago, IL, 1997), pp. 607–610.
43. *Metals Handbook*, 10th ed., vol. 2, (ASM International, 1990), p. 1143.
44. ASTM Standard E8 - 89, "Standard Test of Tension Testing of Metallic Materials," *1995 1995 Annual Book of ASTM Standards, Sec. 3, Metal Test Methods and Analytical Procedures*, vol. 03.01 (West Conshohocken, Penn.: American Society for Testing and Materials, 1995).
45. M. Biebl and H. V. Philipsborn, "Fracture Strength of Doped and Undoped Polysilicon," *Proceedings of the 8th International. Conference on Solid-State Sensors and Actuators, and Eurosensors IX* (Stockholm, Sweden, June 1995), pp. 72–75.
46. T. Tsuchiya, O. Tabata, J. Sakata, and Y. Taga, "Specimen Size Effect on Tensile Strength of Surface Micromachined Polycrystallin Silicon Thin Films," *Proceedings of the Tenth IEEE International Workshop on Microelectromechanical Systems* (Nagoya, Japan, 1997), pp. 529–534.
47. P. T. Jones and G. C. Johnson, "Micromechanical Structures for Fracture Testing of Brittle Thin Films," *Micro-Electro-Mechanical Systems, ASME International Mechanical Engineering Congress and Exposition*, 1996, DSC-vol. 59, pp. 325–330.
48. H. E. Boyer, ed., *Atlas of Stress-Strain Curves*, (Metals Park, Ohio, ASM International, 1987), p. 551.
49. B. E. Jacobson and J. W. Sliwa, "Structure and Mechanical Properties of Electrodeposited Nickel," *Plating and Surface Finishing* (September 1979), 42–47.
50. ASTM Standard E399 - 83, "Plane-Strain Fracture Toughness Testing of Metallic Materials," *1995 Annual Book of ASTM Standards, Sec. 3, Metal Test Methods and Analytical Procedures*, vol. 03.01 (West Conshohocken, Penn.:American Society for Testing and Materials, 1995).
51. J. A. Connally and S. B. Brown, "Micromechanical Fatigue Testing," *Experimental Mechanics*, **33**, 81–90 (1993).
52. L. S. Fan, R. T. Howe, and R. S. Muller, "Fracture Toughness Characterization of Brittle Thin Films," *Sensors and Actuators A* **21–23**, 872–874 (1990).
53. W. N. Sharpe, Jr., D. Danley, and D. LaVan, "Microspecimen Tests of A533-B Steel," in *Small Specimen Test Techniques*, ASTM STP 1329, edited by W. R. Corwin, S. T. Rosinski, and E. Van Walle (American Society for Testing and Materials), in press.
54. Norman E. Dowling, *Mechanical Behavior of Materials* (Prentice Hall, Englewood Cliffs, NJ, 1993).
55. H. O. Fuchs and F. I. Stephens, *Metal Fatigue in Engineering* (Wiley, New York, a Wiley-Interscience Publication, 1980).
56. S. Suresh, *Fatigue of Materials* (Cambridge University Press, Cambridge, NY, 1991).
57. Howard E. Boyer, *Atlas of Fatigue of Curves* (American Society for Metals, 1986).
58. R. C. Rice, B. N. Leis, D. V. Nelson, H. D. Berns, D. Lingenfelser, and M. R. Mitchell, *Fatigue Design Handbook*, 2nd ed. (Society of Automotive Engineers, Inc., Warrendale, PA, 1988).
59. S. B. Brown, G. Povirk, and J. Connally, "Measurement of Slow Crack Growth in Silicon and Nickel Micromechanical Devices," *Proceedings IEEE, Micro Electro Mechanical Systems. An Investigation of Micro Structures, Sensors, Actuators, Machines and Systems* (1993), pp. 99–104.
60. S. B. Brown and E. W. Jansen, "Reliability and Long Term Stability of MEMS," *Digest. IEEE/LEOS 1996 Summer Topical Meetings. Advanced Applications of Lasers in Materials Processing; Broad-*

band Optical Networks – Enabling Technologies and Applications; Smart Pixels; Optical MEMS and Their Applications (1996), pp. 9–10.

61. J. Mohr and M. Strohrmann, "Examination of Long-Term Stability of Metallic LIGA Microstructures by Electromagnetic Activation," *J. Micromechanics and Microengineering,* **2** (3) Germany, 193–195 (1992).
62. P. Krulevitch, A. P. Lee, P. B. Ramsey, J. C. Trevino, J. Hamilton, and M. A. Northrup, "Thin Film Shape Memory Alloy Microactuators," *J. Microelectromechanical Systems* **5** (4), 270–282 (1997).
63. T. I. Kamins, "Structure and Properties of LPCVD Silicon Films," *J. Electrochem. Soc.: Solid-State Science and Technology* **17** (3) 686–690 (1980).
64. M. Legros, S. Kumar, S. Jayaraman, K. J. Hemker, and W. N. Sharpe, "Microstructural Observations of LPCVD Double Layer Polysilicon Thin Film Tensile Specimens," *Polycrystalline Thin Films, MRS Symposium Proceedings,* vol. 472 (San Francisco, 1997), pp. 275–280.
65. W. Betteridge, former chief scientist, International Nickel Limited; *Nickels and Its Alloys* (Ellis Horwood Limited, Chichester), p. 128
66. R. Sard, H. Leidheiser, Jr., and F. Ogburn, *Properties of Electrodeposits; Their Measurement and Significance* (Electrochemical Society, Inc., Princeton, NJ, 1975).
67. *High-Dynamic-Range Microdynamic Accelerometer Technology*, Semiannual Progress Report, The Charles Stark Draper Laboratory (July 1966).
68. T. Akiyama and K. Shono, "Controlled Stepwise Motion in Polysilicon Microstructures," *J. Microelectromechanical Systems* **2**, 106–110 (1993).
69. Jack W. Judy and Richard S. Muller, "Magnetically Actuated Addressable Microstructures;" *J. Microelectromechanical Systems* **6** (3) 249–256 (1997).
70. Xing Yang, Charles Grosjean, Yu-Chong Tai, and Chih-Ming Ho, "A MEMS Thermopneumatic Silicone Membrane Valve," *Proceedings of the Tenth Annual IEEE International Workshop on Micro Electro Mechanical Systems* (Nagoya, Japan, 1997), pp. 114–118.
71. S. Timoshenko and S. Woinowsky-Krieger, *Theory of Plates and Shells*, 2nd ed. (McGraw-Hill, New York, 1959), p. 410.
72. H. Lorenz, *et al.*, "High-Aspect-Ratio, Ultrathick, Negative-Tone Near-UV Photoresist for MEMS Applications," *Sens. Actuators A, Phys.* **64** (1), 33–39 (1998).
73. M. A. Koehonen, P. Borgesen, and Che-Yu Li, "Stress-Induced Voiding and Stress Relaxation in Passivated Aluminum Line Metallization," *Stress-Induced Phenomena in Metallization, First International Workshop, American Vacuum Society Series 13* (American Institute of Physics, 1992), pp. 29–43.

4

MEMS for Harsh Application Environments

M. Mehregany* and C. A. Zorman*

4.1 Overview

Use of silicon (Si) as a mechanical material has enabled the development of a broad range of solid-state sensors and actuators that are well suited for many aerospace applications. Unfortunately, the high-temperature operating limit for these devices is about 250°C, due in large part to degradation in electrical performance above 250°C, a significant decrease in the elastic modulus above 600°C, and temperature limitations of metal contacts. Therefore, many application areas, such as engine instrumentation, cannot benefit from microelectromechanical systems (MEMS) without expensive and bulky packaging schemes to keep Si-based MEMS devices below their high-temperature limit. Wide band-gap semiconductors offer promise for the development of high-temperature MEMS, because these materials have stable electronic properties at high temperatures. In addition, wide band-gap semiconductors such as silicon carbide (SiC) and diamond have outstanding mechanical properties, excellent chemical inertness, and high radiation resistance, which are attractive properties for aerospace applications.

This chapter provides the reader with an overview of SiC as a semiconductor for MEMS in harsh environments. The focus is on SiC because it is the leading material for high-temperature MEMS. The chapter will open with a presentation of the material properties and microstructure of SiC, followed by sections on thin-film growth, processing techniques, micromachining of SiC, and SiC-on-insulator technologies. The chapter will conclude with a review of SiC-based MEMS devices that are suitable for aerospace applications.

4.2 Material Properties of SiC

4.2.1 Introduction

Semiconductors to be used in high-temperature MEMS should have a wide band gap, high thermal conductivity, and excellent mechanical stability at elevated temperatures. Based on these properties, the most attractive materials are diamond and SiC. Prototype MEMS devices have been fabricated from both materials.[1,2] However, SiC has a number of distinct advantages over diamond, which has made SiC the leading material for MEMS in harsh environments. Economic issues require MEMS materials to be compatible with batch fabrication, preferably, batch fabrication processes used in the Si integrated circuit (IC) industry, which necessitates the use of large-area substrates. Single and polycrystalline SiC films can be grown on large-area Si wafers; whereas diamond can only be deposited in polycrystalline form on Si substrates. Another advantage of SiC is that numerous Si-compatible plasma etch processes have been developed to pattern SiC films; whereas diamond films are not readily patterned. Despite these advantages of SiC, diamond is still an attractive high-temperature material for MEMS applications. Details concerning film growth, patterning, device fabrication, and performance can be found in the literature.[3–5] The remainder of this chapter will concentrate on the development and implementation of SiC as a MEMS material for harsh environments.

---

*Department of Electrical Engineering and Applied Physics, Case Western Reserve University, Cleveland, Ohio.

4.2.2 Properties of SiC

SiC has long been recognized as a semiconductor with outstanding physical and chemical characteristics. Compared to Si, SiC exhibits a larger band gap, a higher breakdown field, a higher thermal conductivity, and a higher saturation velocity. These properties make SiC a very attractive material for the fabrication of high-temperature, high-power, and high-frequency electronic devices. Moreover, a high elastic modulus and high hardness make SiC an excellent material for the mechanical components in high-temperature microsensors and microactuators. High-temperature microsensors and microactuators can be used for pressure sensing, temperature sensing, and chemical sensing in gas turbine engines. SiC also has a higher chemical inertness and radiation resistance than Si, which expands its potential as a material for sensors and actuators in satellite and other space systems.

4.2.3 Crystal Structure

Any discussion of SiC as a material for microdevices requires an understanding of the crystalline structure of SiC. SiC exhibits a one-dimensional polymorphism called polytypism. All polytypes of SiC have a common planar arrangement of Si and C atoms, but each polytype is distinguished by a unique stacking sequence of the identical planes. Disorder in the stacking periodicity of the similar planes results in a material that has numerous crystal structures (polytypes), all with the same atomic composition. The magnitude of the disorder is such that over 250 SiC polytypes have been identified to date.[6] Despite the large number of polytypes, only three crystalline symmetries exist: cubic, hexagonal, and rhombohedral. Historically, the cubic phase of SiC has been referred to as ß-SiC, and the hexagonal and rhombohedral phases have been called α-SiC. Recently, a more descriptive nomenclature that identifies both the crystalline symmetry and stacking periodicity has been adopted. Using this system, cubic SiC is called 3C-SiC, which is the only cubic polytype known to exist. The most common α-SiC polytypes have hexagonal symmetries and are called 6H-SiC, 4H-SiC, and 2H-SiC.

4.2.4 Physical Characteristics

Even though polytypes have the same atomic composition, the electrical properties of each polytype are different. For instance, the band gap for SiC ranges from 2.3 eV for 3C-SiC to 3.4 eV for 2H-SiC. Due, in part, to its cubic crystalline symmetry, 3C-SiC has the highest electron mobility (1000 $cm^2/V \cdot s$) and saturation drift velocity ($10^7 cm/s$).

SiC has long been noted for its hardness, wear resistance, and chemical inertness. SiC has a Mohs hardness of 9, which compares favorably with values for other hard materials, such as diamond (10) and topaz (8). In terms of wear resistance, SiC has a value of 9.15, as compared with 10 for diamond and 9 for Al_2O_3. SiC can be etched by alkaline hydroxide bases (i.e., potassium hydroxide [KOH]), but only at very high temperatures (~600°C), and is not etched by acids. SiC does not melt, but sublimes above 1800°C. The surface of SiC can be passivated by the formation of a thin thermal SiO_2 layer, but the oxidation rate is very low when compared with that of Si. A summary that compares the important semiconductor properties of 3C-SiC and 6H-SiC with those of other noted semiconductors is presented in Table 4.1.

4.3 Thin Film Growth

4.3.1 Homoepitaxy of 6H-SiC

A major impediment to the commercialization of SiC as a high-temperature semiconductor is the availability of large-area, high-quality, defect-free SiC substrates suitable for epitaxial growth. Although 3C-SiC has the least complex crystal structure of all the SiC polytypes, bulk crystal growth of high-quality 3C-SiC is very difficult, and no wafer-grade substrates have been

Table 4.1. Important Properties of High-Temperature Semiconductors

Property	3C-SiC (6H-SiC)	GaAs	Si	Diamond
Melting point (°C)	> 1800[a]	1238	1415	1400[b]
Thermal conductivity (W/cm°C)	5	0.5	1.5	20
Coeff. thermal expan. ($°C^{-1}\times10^{-6}$)	4.2	6.86	2.6	1.0
Young's modulus (GPa)	448	75	190	1035
Physical stability	Excellent	Fair	Good	Fair
Bandgap (eV)	2.2 (2.9)	1.424	1.12	5.5
Electron mobility ($cm^2/V\cdot s$)	1000 (600)	8500	1500	2200
Hole mobility ($cm^2/V\cdot s$)	40	400	600	1600
Breakdown voltage (10^6V/cm)	4	0.4	0.3	10
Dielectric constant	9.72	13.1	11.9	5.5

[a] Sublimation temperature.
[b] Phase change temperature.

fabricated. Several companies have developed successful processes to grow high-quality bulk 6H-SiC crystals, and 2-in.-diam electronic-grade wafers are commercially available. Unfortunately, these wafers are relatively expensive, which limits their use to low-volume, high-cost applications.

Development of SiC has focused on 6H-SiC for high-temperature, high-power, and high-frequency microelectronics, and a detailed review has been recently published.[7] For high-temperature electronic devices, high quality n- and p-doped single-crystal SiC films are required. A common method for growing homoepitaxial 6H-SiC films on 6H-SiC substrates uses atmospheric-pressure chemical vapor deposition (APCVD) with Si- and C-containing gases and high substrate temperatures (1500°C–1700°C). Commonly used precursor gases are silane (SiH_4) and propane (C_3H_8), with H_2 as a carrier gas. Typical dopant gases are trimethyl-aluminum [$Al(CH_3)_3$] and diborane (B_2H_6) for p-type films, and N_2, ammonia (NH_3), and phosphene (PH_3) for n-type films.

4.3.2 Heteroepitaxy of 3C-SiC on Si

Unlike the other polytypes, single-crystal 3C-SiC, hereafter simply called 3C-SiC, can be heteroepitaxially grown on Si substrates by both APCVD and low-pressure chemical vapor deposition (LPCVD). The most common APCVD process uses H_2 as the carrier gas, SiH_4 as the Si source gas, and C_3H_8 as the C source gas.[8] Other processes use dichlorosilane as a Si source, and acetylene as a C source.[9] Single C- and Si-containing sources, such as methyltrichlorosilane (CH_3SiCl_3) and methylsilane (CH_3SiH_3), have also been used to grow 3C-SiC by LPCVD.[10,11] Heteroepitaxy is possible because 3C-SiC and Si have similar cubic crystal structures: 3C-SiC has a zinc-blend structure with a lattice constant of 0.436 nm, while Si has a diamond structure with a lattice constant of 0.543 nm, resulting in a lattice mismatch of approximately 20%. A process called carbonization is often used to initiate heteroepitaxial growth by forming a thin 3C-SiC film directly from the Si substrate. Carbonization converts the near surface region of the Si substrate to 3C-SiC by exposing a heated substrate to a carbon-containing gas. Carbonization temperatures range from 1250°C to 1360°C, depending on the process. The carbon-containing gas dissociates

into hydrocarbon reactants, which react with Si on the wafer surface, forming a thin, heteroepitaxial 3C-SiC film. Because SiC films are excellent diffusion barriers, the carbonization process is self-limiting. Film growth is continued by introducing a Si-containing gas to the flow, which initiates homoepitaxial growth of 3C-SiC on the heteroepitaxial 3C-SiC carbonization layer. 3C-SiC films as thick as 40 μm have been grown on small Si substrates.[12] Moreover, 3C-SiC films have been grown on large-area Si wafers (4-in.-diam),[13] enabling the batch fabrication of 3C-SiC MEMS devices.

Despite the obvious advantage of growing 3C-SiC films on inexpensive, large-area Si wafers, heteroepitaxial 3C-SiC films suffer from a large density of crystalline defects. The large defect density results, in part, from the 20% lattice mismatch, but also from an 8% difference in thermal expansion coefficients between 3C-SiC and Si. The defect density is highest at the SiC/Si interface and decreases with increasing film thickness. Unfortunately, the defect density in heteroepitaxial 3C-SiC films is not low enough to make the performance of 3C-SiC electronics comparable with 6H-SiC and Si devices. However, the crystal quality of 3C-SiC may be good enough for high-temperature microsensor applications.

Another troubling problem associated with 3C-SiC heteroepitaxy on Si substrates is the formation of voids at the SiC/Si interface. These voids are sealed microcavities and are common to samples grown by most APCVD and LPCVD processes. The void density can be quite high, with values as high as 10^5 voids/cm^2 reported in the literature.[12] Voids compromise the contact between the 3C-SiC film and the Si substrate, and may contribute to the large defect density in 3C-SiC films. The mechanism for void formation is not clear. However, it has been suggested that voids result from Si out-diffusion from uncarbonized regions of the Si surface during the carbonization process.[14] Void formation also appears to be dependent on the temperature ramp-up rate during the carbonization step. Typically, high ramp-up rates of about 50°C/s are used. Void densities as low as 2 voids/cm^2 have been achieved when low ramp-up rates (~3°C/s) are used.[13] Others[15,16] have since developed growth processes with similar results.

As mentioned previously, the main advantage of 3C-SiC from a MEMS perspective is that 3C-SiC can be heteroepitaxially grown on Si substrates, which enables 3C-SiC growth on inexpensive, large-area wafers. Using an APCVD process, uniform heteroepitaxy of 3C-SiC across 4-in. Si wafers has been demonstrated.[13] Additionally, a LPCVD reactor has been used to grow 3C-SiC films on multiple 4- and 6-in. Si wafers.[17]

4.3.3 Polycrystalline SiC

For many MEMS applications, polycrystalline SiC can be used. Unlike 3C-SiC and 6H-SiC, polycrystalline 3C-SiC, hereafter called poly-SiC, can be deposited on a wide variety of substrate types, including suitable sacrificial layers such as SiO_2. Poly-SiC has been deposited by plasma-enhanced chemical vapor deposition (PECVD), sputtering, and electron beam evaporation at substrate temperatures ranging from 200°C to 1000°C.[18–20] These films are either amorphous, as in the case with low-temperature PECVD, or polycrystalline, with a texture dependent on deposition temperature. APCVD and LPCVD processes have been used to deposit poly-SiC on Si substrates, resulting in films with microstructures much like the aforementioned films.[21,22]

Traditionally, APCVD has been used to grow 3C-SiC films on Si substrates and 6H-SiC films on 6H-SiC substrates. Recently, however, APCVD has been used to deposit SiC films on polysilicon substrates.[23] Polysilicon was chosen because of its potential as a sacrificial layer in a SiC-based surface micromachining process. As mentioned previously, SiC etching in KOH is not practical, except at temperatures above 600°C; whereas Si is readily etched in KOH at temperatures below 70°C. Thus, a 2- to 3-μm-thick polysilicon layer deposited on a thermally oxidized

Si wafer provides an excellent substrate for a SiC surface micromachining process that uses KOH as a release agent. In this study, poly-SiC films were grown on polysilicon using the traditional 3C-SiC APCVD growth process. Two types of polysilicon films were used as substrates: (1) as-deposited polysilicon, and (2) annealed polysilicon. The annealing time was such that the as-deposited polysilicon, which is highly oriented in the [110] direction, was completely recrystallized during the process. The annealed polysilicon is a mixture of (220) and (111) crystallites. X-ray diffraction (XRD) and transmission electron microscopy (TEM) were used to characterize the films. Polycrystalline SiC grown on as-deposited polysilicon has a texture that closely resembles the as-deposited polysilicon substrate, despite the fact that during the growth process, the underlying polysilicon is fully recrystallized. The recrystallization process does not modify the crystallinity or adversely affect the adhesion of the poly-SiC films. Poly-SiC films grown on fully recrystallized (annealed) polysilicon exhibit a grain-to-grain epitaxial relationship with the substrate. It is well established that for polysilicon films, texture influences such physical properties as oxidation rate, thermal conductivity, elastic modulus, and film stress.[24–26] The same may also be true for poly-SiC. This process makes possible the ability to grow highly textured poly-SiC by controlling the substrate texture.

The most straightforward surface micromachining process uses an insulating film, like SiO_2, as the sacrificial substrate material. Poly-SiC films have been sputter deposited and reactively evaporated on SiO_2 substrates, but only recently has APCVD been used.[27] No carbonization of the SiO_2 layer was necessary, so a single-step film growth process was used. It was reported that films grown at 1050°C and 1160°C exhibited good adhesion to the SiO_2 substrates, while films deposited at 1280°C delaminated during or shortly after film growth. At lower growth temperatures, the poly-SiC films were randomly oriented, with grain size increasing with increasing temperature. Lower deposition temperatures resulted in the deposition of Si-rich poly-SiC films, although the excess Si concentration was less than 10 at%.

4.3.4 APCVD or LPCVD

APCVD and LPCVD are the two most versatile techniques to deposit SiC for MEMS applications, since APCVD and LPCVD can be readily used to deposit both 3C-SiC and poly-SiC. This capability enables the fabrication of hybrid sensors and actuators that consist of 3C-SiC electronics coupled with poly-SiC mechanical components. LPCVD is noted for producing films with a high degree of thickness uniformity (<5% variation), although the growth rates for 3C-SiC tend to be low (~0.1 μm/h).[22] LPCVD is particularly well suited for batch fabrication, since film growth can be performed in furnace tubes that can accommodate numerous large-area wafers. APCVD films are grown at a much higher rate (1–2 μm/h); however, thickness uniformity on large-area substrates varies by as much as 30%,[13] and only one or two 4-in. wafers can be loaded in each deposition run.

4.4 Processing Techniques

4.4.1 Introduction

As mentioned previously, SiC films can be processed using many of the standard processing tools common to Si IC fabrication. Thermal oxidation, metallization, and plasma etching can be performed in a Si fabrication facility without any major retooling. In terms of materials compatibility, it has been shown that SiC processing does not lead to contamination of Si processing equipment.[28] In fact, poly-SiC is now being used as an alternative to quartz for wafers, wafer boats, susceptors, and other components used in Si furnaces.

4.4.2 Oxidation

Unlike for diamond, stable thermal oxides can be grown on SiC. In fact, standard Si thermal oxidation processes are used to grow thermal oxides on SiC, even though the fundamental oxidation mechanism is different. For example, oxidation of Si is achieved by the direct reaction of Si with O_2 to form SiO_2; whereas SiC reacts with O_2 to form SiO_2 and CO during the SiC oxidation process.[7] Because SiC is more chemically stable than Si and thereby less likely to react with O_2, the time required to form a thermal oxide of equivalent thickness is longer for SiC. For example, a process used to form a 1.5-μm-thick thermal oxide on Si yields only a 900-Å-thick thermal oxide on 3C-SiC. As with Si oxidation, H_2 enhances the oxidation rate of SiC.

Many MEMS applications require thick (>1 μm) oxides that cannot be achieved, with reasonable feasibility, by thermal oxidation of SiC. As an alternative to direct thermal oxidation, SiO_2 layers can be "deposited" on SiC by first depositing a polysilicon film onto the SiC substrate, then converting the entire polysilicon film to SiO_2 by thermal oxidation.[29] This process takes advantage of well-established polysilicon deposition and oxidation techniques. Also, because the oxide is formed by thermal oxidation, it has favorable high-temperature properties. SiO_2 thicknesses of well over 1.5 μm can be achieved by thermal oxidation. An example that uses this technique as part of a wafer-bonding process will be presented later in this chapter.

4.4.3 Metallization

Several comprehensive reviews have recently been published that summarize the research conducted on electrical contacts to SiC.[30,31] In general, the best high-temperature contacts are metals with high melting temperatures, such as Ni and W. These metals can be used to form either ohmic or Schottky contacts to 3C-SiC and 6H-SiC. This section will focus only on metal contacts to 3C-SiC.

The best and most widely used ohmic contact to 3C-SiC is Al. Al is the preferred ohmic contact because it is easily deposited by sputtering and evaporation, it is commonly used in silicon processing, and a wire bonding/packaging technology for Al currently exists. Unfortunately, Al melts at about 600°C, making it unsuitable for high-temperature contacts. Sputter-deposited Ni, W, Mo, and other complementary metal oxide semiconductor (CMOS) compatible refractory metals make ohmic contacts to 3C-SiC after a short high-temperature (i.e., 900°C for Ni) post-deposition anneal. Binary compounds such as $TiSi_2$ and WSi_2, and alloys like Au-Ta, are also ohmic contacts to 3C-SiC.

The best Schottky contact to 3C-SiC is Au. Like Al, Au is easily deposited, and a wire bonding/packaging technology currently exists. Unfortunately, Au is not a Si-compatible metal, so care must be taken when using Au. Although Au has a relatively high melting temperature, it is not a suitable high-temperature contact material, because of electromigration at relatively low temperatures. High-melting point metals like Ti and Ni have shown Schottky behavior on 3C-SiC. Comprehensive tables summarizing the deposition conditions and evaluation data for metal contacts to n- and p-type 3C-SiC are presented in Ref. 30.

4.4.4 Plasma Etching

Selective etching is required to pattern the SiC films into the desired structural components for a MEMS device. For Si-based devices, patterning can be performed by dry (plasma) or wet chemical etching. Although wet chemical etching of SiC is not feasible, plasma etching techniques have been developed.[32–34] These techniques use nearly the same F-based plasma chemistries developed for Si, SiO_2, and Si_3N_4 films. Commonly used fluorinated compounds are CF_4, CHF_3, NF_3, and SF_6. Usually, etching is conducted in reactive ion etching (RIE) mode, meaning that pressures are kept below 200 mtorr and sputtering of the substrate is suppressed. For the most

part, these techniques are effective for device fabrication. However, the etch rates for SiC are relatively low when compared with etch rates for Si under similar conditions.

Most RIE processes for SiC require plasmas consisting of O_2 mixed with a fluorinated compound. Unfortunately, common photoresists are not resistant to these highly oxygenated plasmas and cannot be used to etch micron-thick SiC films. Therefore, other masking materials, such as Al, are used. Al is effective as a masking material, but may be responsible for a phenomenon called micromasking. Micromasking occurs when sputtered material from the chamber or etch mask is deposited in the etch field and forms small masks that shield the underlying etch field from the plasma. If the etching process has a high degree of anisotropy, micromasks will produce undesirable "grasslike" structures in the etch field. Decreasing the anisotropy of the plasma can undercut and eliminate the micromasks, but will also reduce the anisotropy of the etched features. Using graphite electrodes[35] or adding small concentrations of hydrogen to the plasmas reduces the micromasking effect.[36]

Etch selectivity is another key issue facing the SiC MEMS processing. SiC-surface micromachining processes utilize SiO_2 and polysilicon films for sacrificial layers, and SiO_2 films for dielectric isolation. In general, plasma conditions that etch SiC at high rates, etch SiO_2 and polysilicon at even higher rates. Some progress has been made in improving the etch selectivities of SiC to Si and SiO_2. The best reported selectivities for room temperature etching were 2:1 for SiC to Si, and 1:3.6 for SiC to SiO_2.[37] Because these selectivities are nowhere near ideal for micromachining, great care must be taken when etching SiC on Si and SiO_2 substrate materials. Examples of SiC-based MEMS devices that utilized plasma etching during the fabrication process will be presented later in this chapter.

4.5 Micromachining of SiC

4.5.1 Bulk Micromachining

As stated previously, bulk micromachining of SiC is very difficult, because of its outstanding chemical stability. Standard Si bulk micromachining techniques that use KOH or ethylenediamine pyrocatechol (EDP) are not effective in etching SiC. Some success in bulk micromachining of SiC using nonstandard techniques has been demonstrated. For instance, a laser-assisted photoelectrochemical etching (PEC) technique for n-type 3C-SiC has been developed.[38] Subsequently, an n-type 3C-SiC etch process that uses a p-type 3C-SiC etch stop was demonstrated.[39] PEC has since been extended to 6H-SiC and was successfully used in a pressure sensor fabrication process.[40]

For MEMS applications such as pressure sensing, hybrid structures consisting of SiC films on bulk micromachined Si substrates have been investigated. Si bulk micromachining techniques are well suited for fabricating these structures. The resistance of SiC to Si etchants is so high that SiC is an excellent etch stop for Si bulk micromachining. The most basic hybrid structure is a 3C-SiC membrane on a Si substrate. A cross section of the simple fabrication process is shown in Fig. 4.1. A 3C-SiC film of the desired thickness is first grown on a Si substrate. A thermal oxide is then grown on both sides of the sample. The back-side oxide is photolithographically patterned to form a KOH-resistant diaphragm mask, and the front-side oxide is etched off the 3C-SiC surface. The sample is then immersed in a KOH etch bath to etch away the exposed regions of Si to form the SiC diaphragms. The potential of this technique for batch fabrication was demonstrated when approximately 275 2-μm-thick 3C-SiC diaphragms were successfully fabricated on a single 4-in. Si wafer.[41]

Fundamental to the success of SiC-based MEMS is a thorough understanding of the mechanical properties of SiC films. Bulk-micromachined SiC diaphragms have been used extensively as

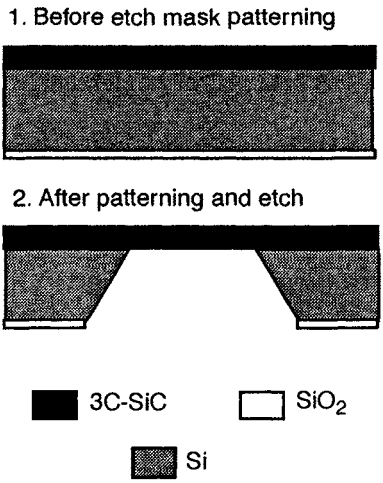


Fig. 4.1. Schematic diagram of the bulk micromachining process to fabricate 3C-SiC diaphragms.

test structures to determine the biaxial modulus and residual stress by investigating the load-deflection behavior of the diaphragms. This load-deflection technique uses an interferometer to measure the center deflection of a diaphragm experiencing an applied pressure. Deflection versus applied pressure data is acquired and fit to a model that contains terms dependent on the biaxial modulus and residual stress.[42]

The load-deflection technique has been used on both 3C-SiC and poly-SiC diaphragms. For diaphragms fabricated from 3C-SiC films grown by APCVD on small Si substrates, an average biaxial modulus of 441 GPa and average residual stress of 221 MPa were reported.[43] No dependence on film thickness was found for the biaxial modulus; whereas, residual stress decreased with increasing thickness. Biaxial modulus and residual stress for both 3C-SiC and poly-SiC films grown by LPCVD have also been studied.[44] It was reported that for 3C-SiC films of thicknesses up to 1.3 μm, the biaxial modulus was about 450 GPa, while the residual stress was about 150 MPa. For poly-SiC films of thicknesses up to 5.0 μm, the biaxial modulus was nearly that of the 3C-SiC films, averaging 465 GPa, while residual stress could be adjusted from 0 to 250 MPa by changing the deposition parameters. Another study investigated the change in biaxial modulus as a function of dopant gas concentration for LPCVD poly-SiC. This study found that for increasing boron concentrations (for B/Si ratios up to 0.02), the biaxial modulus peaked at 600 GPa, as compared to 480 GPa for undoped films.[45] Residual stress values for these films were not reported.

Other bulk micromachined structures have been used to determine the mechanical properties of 3C-SiC films. A vibrating membrane technique was used to determine the residual stress in 3C-SiC films grown on Si substrates.[46] It was reported that the residual stresses were highest in p-type films as compared with n-type and undoped films. It was also reported that the internal stresses in doped and undoped films decreased with increasing temperature up to 600°C. By extrapolating the linear portion of the stress versus temperature data to high temperatures, it was found that zero stress occurred between 1250°C and 1350°C, which corresponded to the growth temperature of the films. Thus, it was concluded that the films were grown under stress-free conditions, and therefore the internal stress measured at room temperature was due to the thermal mismatch between the 3C-SiC films and the Si substrates. In a different study, the free-standing microcantilever beams shown in Fig. 4.2 were used to investigate the bending moment caused by the residual stress gradient in 3C-SiC films.[47] The average bending moment was about

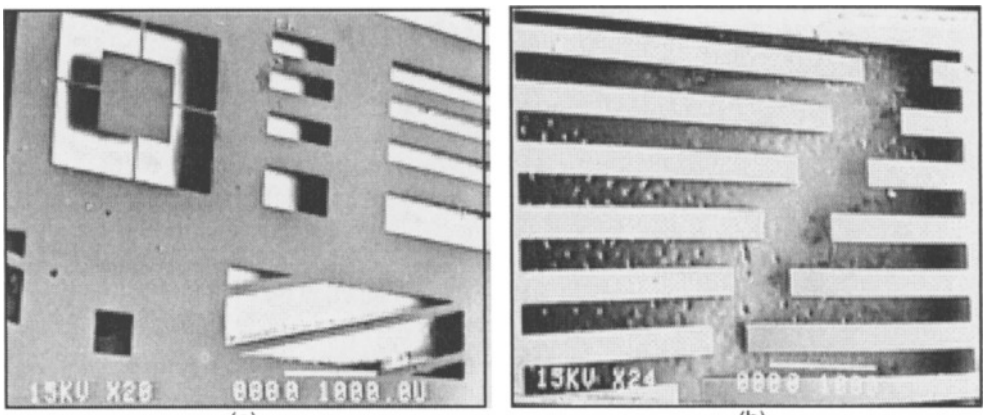


Fig. 4.2. (a) SEM micrographs of bulk-micromachined SiC suspensions[43] and (b) cantilever beams.

3.4×10^8 Nm. Using a vibrating cantilever technique, the Young's modulus (which is closely related to the biaxial modulus) for 3C-SiC films was measured to be about 694 GPa for undoped 3C-SiC and 474 GPa for p-type 3C-SiC.[48]

The spatial variation of the biaxial modulus and of the residual stress of 3C-SiC films deposited by APCVD on large-area substrates has been studied.[41] The load-deflection technique was applied to 12 diaphragms taken from well-distributed locations across each wafer. It was reported that under all deposition conditions, the spatial variation of the biaxial modulus and of the residual stress was lowest within the center 3-in. diameter of a 4-in.-diam wafer. In the same study, the variation in the biaxial modulus and in the residual stress as a function of precursor gas flow rates and deposition temperatures was investigated. In general, at a fixed deposition temperature (1280°C), high precursor gas flow rates produced 3C-SiC films with lower average residual stresses (112 MPa) and higher biaxial moduli (348 GPa) than low flow rates (340 MPa, 294 GPa). At a lower deposition temperature (1160°C), the films changed from single to polycrystalline at both high and low flow rates. However, the residual stress increased significantly (from 112 to 311 MPa) for high flow rates, while only modestly (from 340 to 360 MPa) for low flow rates.

The variability of the mechanical properties in APCVD 3C-SiC films as a function of susceptor age and susceptor supplier has also been studied.[49] Susceptor age is defined by the total number of deposition hours a susceptor has been used for SiC growth. Early reports suggested that material deposited on the surface of susceptors during SiC growth may contribute to the deteriorating electrical and morphological properties of SiC films.[12] To study the influence of susceptor age on the mechanical properties, 3C-SiC films were grown under identical conditions 50 runs apart, which was equivalent to about 100 deposition hours. During this span, considerable changes were observed on the susceptor, namely, dark gray deposits along the upstream surfaces. Similar observations have been reported for films grown using small susceptors.[12] In terms of biaxial modulus and residual stress, no significant changes between the samples were noticed. Films grown using susceptors from two different manufacturers were also studied. The susceptors were "seasoned" by using them for 25 depositions prior to growing films for study. As in the susceptor age study, no significant differences were found between the samples. In terms of the mechanical properties of SiC films, this study shows that susceptors can be used for many depositions (>100 deposition hours), and that the performance of a susceptor is not dependent on the supplier.

4.5.2 Surface Micromachining

The first reported SiC surface micromachining process was developed for poly-SiC films deposited on polysilicon sacrificial layers.[50] A single-mask lateral resonant structure was chosen as the demonstration vehicle. A schematic cross section of the fabrication process and a scanning electron microscopy (SEM) micrograph of the device are shown in Figs. 4.3(a) and 4.3(b). The substrate was prepared by first growing a 1-μm-thick thermal oxide on a silicon wafer. The thermally grown SiO_2 layer provided electrical isolation and protected the Si substrate during release. A 2-μm-thick polysilicon film was then deposited on the oxidized Si substrate. Poly-SiC was deposited on the polysilicon using the three-step SiC growth process (in-situ clean, carbonization, film growth) detailed earlier in this chapter. RIE was used in conjunction with photolithography to pattern the poly-SiC film into the desired shape. The free-standing sections of the resonator were released by using a timed etch in KOH, which does not etch the poly-SiC or the thermal oxide. Supercritical drying was used to minimize stiction problems associated with the release. To operate the devices, an actuation voltage ranging between 30 and 175 V was required. These devices had a resonant frequency of 20 kHz.

Although polysilicon is an adequate sacrificial layer material for SiC surface micromachining, SiO_2 is preferred, since it can be used as both the electrical isolation layer and sacrificial layer. A

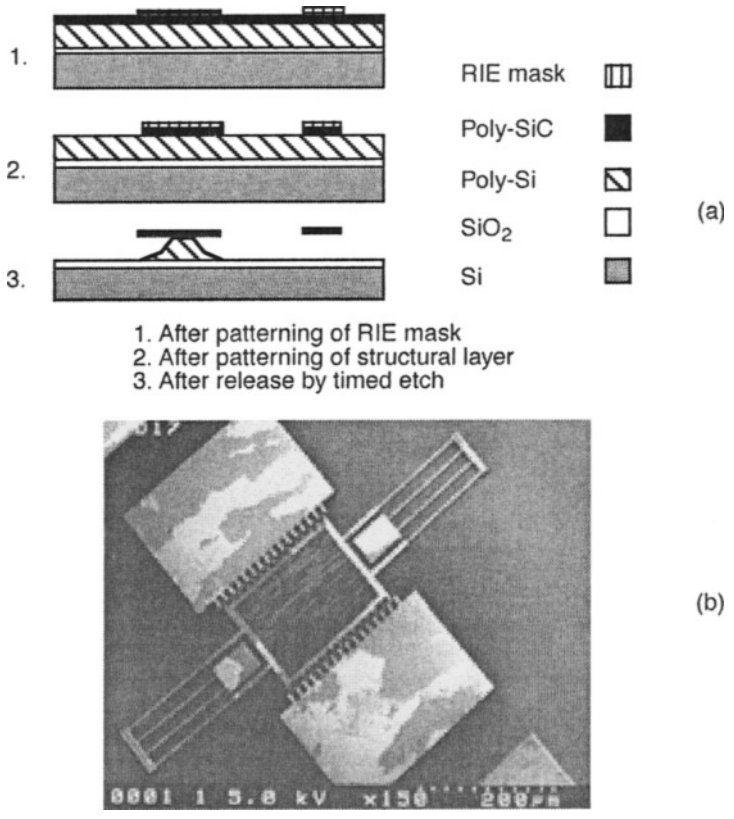


Fig. 4.3. (a) Schematic diagram of a SiC surface micromachining process using a polysilicon sacrificial layer. (b) SEM micrograph of a SiC surface micromachining lateral resonant structure fabricated using a polysilicon sacrificial layer.[50]

SiC surface micromachining process that uses a poly-SiC film deposited by APCVD on a SiO_2 sacrificial layer has been developed.[27] A schematic cross section and SEM micrograph of a lateral resonant structure are shown in Figs. 4.4(a) and 4. 4(b). These structures were fabricated from 2-μm-thick poly-SiC films grown on 3.5-μm-thick SiO_2 sacrificial layers on Si substrates. The poly-SiC films were photolithographically patterned using RIE to form the resonator structure. The devices were released by a timed etch in 49% HF solution, which etched the underlying SiO_2 to form the free-standing poly-SiC structures. The devices resonated at frequencies between 18 and 50 kHz using actuation voltages of about 50V.

To confirm the commonly held belief that SiC is a better mechanical material than Si at high temperatures, the resonant frequency response of poly-SiC and polysilicon lateral resonant structures was compared for devices heated to 900°C.[51] All devices used in this study were fabricated from 2-μm-thick films deposited on SiO_2 sacrificial layers using the same photolithographic mask and surface micromachining process. Testing was performed in an argon-filled chamber that was equipped with a heated sample stage capable of reaching temperatures above 1000°C. Figure 4.5 shows the behavior of the resonant frequency as a function of temperature for polysilicon [Fig. 4.5(a)] and poly-SiC [Fig. 4.5(b)] lateral resonant devices. For the polysilicon devices, the resonant frequency began to decrease at 350°C. The average rate of reduction in resonant frequency for the polysilicon devices between room temperature and 900°C was 1.11 Hz/°C. For temperatures above 350°C, the rate of reduction increased to 1.92 Hz/°C. For the poly-SiC devices, no detectable change in the resonant frequency was observed for temperatures below

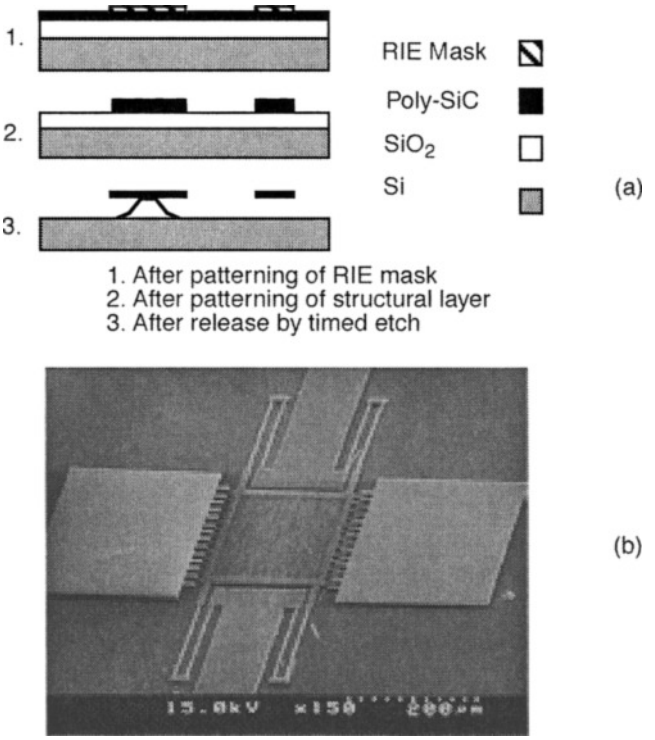


Fig. 4.4. (a) Schematic diagram of a SiC surface micromachining process using a SiO_2 sacrificial layer. (b) SEM micrograph of a SiC surface micromachining process using a SiO_2 sacrificial layer.[27]

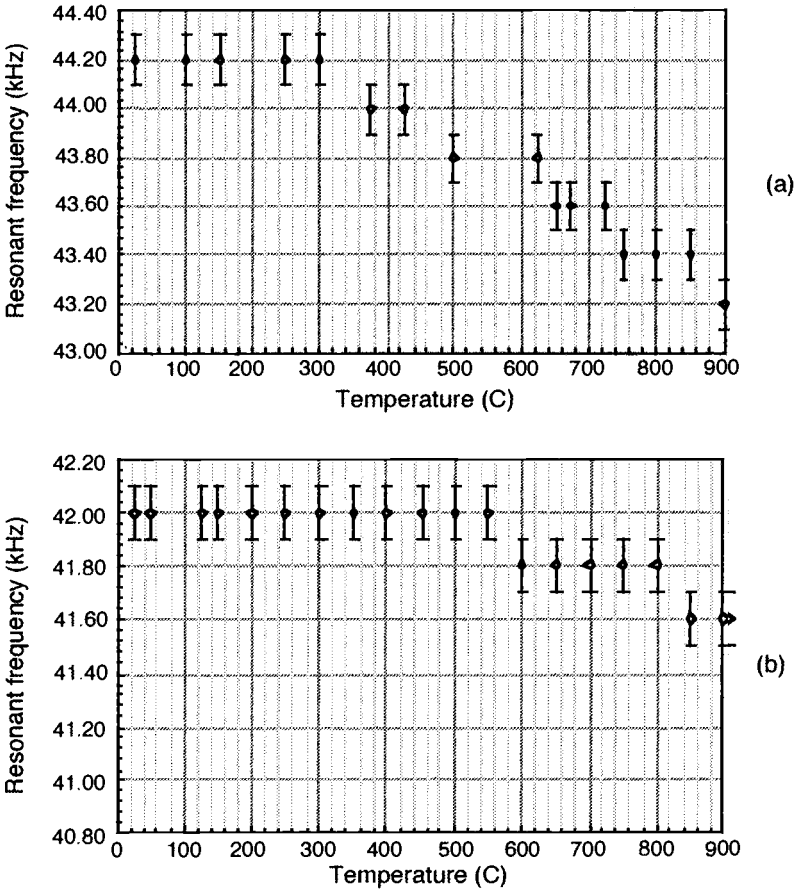


Fig. 4.5. Resonant frequency vs temperature data for lateral resonant devices (a) polysilicon and (b) poly-SiC.[51]

500°C. The average reduction in resonant frequency for the poly-SiC devices was 0.44 Hz/°C over the entire temperature range and 1.14 Hz/°C for temperatures above 500°C.

The resonant frequency of a lateral resonant structure is a function of device geometry and material properties and is described by the following equation:

$$f_r = \sqrt{\frac{hEw^3}{2\pi^2 m l^3}} \tag{4.1}$$

where h is thickness of the beams, E is the Young's modulus, w is the width of the beam, m is the mass of the device, and l is the length of the beam. If the thermal expansion properties of the device are taken into account, it can be shown that the thermal expansion of l and w will cancel out, and that the thermal expansion of h is very small and should cause an increase in the resonant frequency of the device. However, the resonant frequencies of both the poly-SiC and the polysilicon devices are reduced, implying that the Young's modulus for both materials is reduced at elevated temperatures. Poly-SiC, which exhibits a smaller variation in Young's modulus with increasing temperature, is superior to polysilicon as a structural material for high-temperature MEMS applications.

4.6 SiC-on-Insulator Technologies

Unlike poly-SiC, single crystal SiC cannot be grown directly on SiO_2 or on any other suitable nonconducting sacrificial material. Therefore, surface micromachining and electrical isolation of 3C-SiC structures and devices are very difficult. Because nitrogen is a shallow n-type donor in SiC and is easily incorporated during deposition and crystal growth processes, deposition of insulating SiC films is also very difficult. The need for electrically isolated single crystal SiC films has motivated researchers to borrow from silicon-on-insulator (SOI) fabrication techniques in order to produce SiC-on-insulator substrates. The three SiC-on-insulator fabrication processes that have been reported in the literature are: (1) growth of 3C-SiC on SOI substrates;[52–54] (2) the Smart-Cut™ process;[55] and (3) wafer bonding.[56]

The first process to fabricate electrically isolated 3C-SiC films used SOI wafers as substrates for epitaxial growth of 3C-SiC. The SOI wafers were produced either by ion implantation of O_2 into the subsurface region of an Si wafer (separation by implanted oxygen [SIMOX]), or by bonding two thermally oxidized Si wafers and removing all but a very thin layer of one of the wafers. The processing details for each of these techniques can be found elsewhere.[52–54] To create a SiC-on-insulator substrate from a SOI wafer, the thin Si layer atop the buried oxide must be fully converted to 3C-SiC; otherwise, a 3C-SiC-on-Si-on-SiO_2 structure is created. The conversion process occurs during the carbonization step, which limits the thickness of the Si layer to about 200 nm. It has been observed that during the carbonization process, out-diffusion of oxygen from the buried SiO_2 layer occurs, which creates sealed cavities at the 3C-SiC/SiO_2 interface.[53,54] These sealed cavities may adversely affect the electrical and mechanical properties of the structure. SIMOX substrates have limited high-temperature applications, because the thickness of the implanted oxide layer is limited to a few thousand angstroms, and the quality of the implanted oxide is poor when compared with thermal oxides.

The second SiC-on-insulator fabrication process, known as the Smart-Cut™ process, was first developed to produce SOI structures for silicon-based microelectronics.[55] The process combines ion implantation with wafer bonding to create a SOI substrate. Two Si wafers, hereafter called the handling wafer and the implant wafer, are thermally oxidized to form a thick (~1 μm) SiO_2 film on each wafer. The implant wafer then undergoes hydrogen ion implantation, which deposits hydrogen below the thermal oxide and below a thin layer of Si. After implantation, the SiO_2 surfaces of the two wafers are fusion bonded at high temperature. During this step, the implanted hydrogen condenses into a thin, well-defined region of voids. This region forms a seam that is used to remove the implant wafer from the bonded pair, thus transferring the thin Si layer to the handling wafer and creating an SOI substrate. The high-temperature bonding step also serves to anneal any ion-induced lattice damage in the thin Si layer. After bonding, the Si surface is polished and readied for processing.

By using 6H-SiC wafers as the implant wafers, the Smart-Cut™ process has been adapted to fabricate 6H-SiC-on-insulator substrates.[57] The 6H-SiC implant wafers have been successfully bonded to 6H-SiC, poly-SiC, and Si-handling wafers. These substrates may not be suitable for MEMS, since the overall area of a 6H-SiC-on-insulator substrate is limited by the size of the 6H-SiC implant wafer, which is currently only 2 in. in diameter. An additional concern is that implantation damage in the 6H-SiC layer is not annealed during the bonding step.[58] Annealing these defects may require temperatures that may damage the underlying SiO_2 layer.

The third SiC-on-insulator fabrication process uses wafer bonding to create SiC-on-insulator structures. A schematic of the process is shown in Fig. 4.6. The first wafer bonding process used SiO_2-to-SiO_2 bonding to create 3C-SiC-on-SiO_2 structures.[56] The process begins with epitaxial growth of 3C-SiC on a Si transfer wafer. The handling wafer is prepared by thermally oxidizing

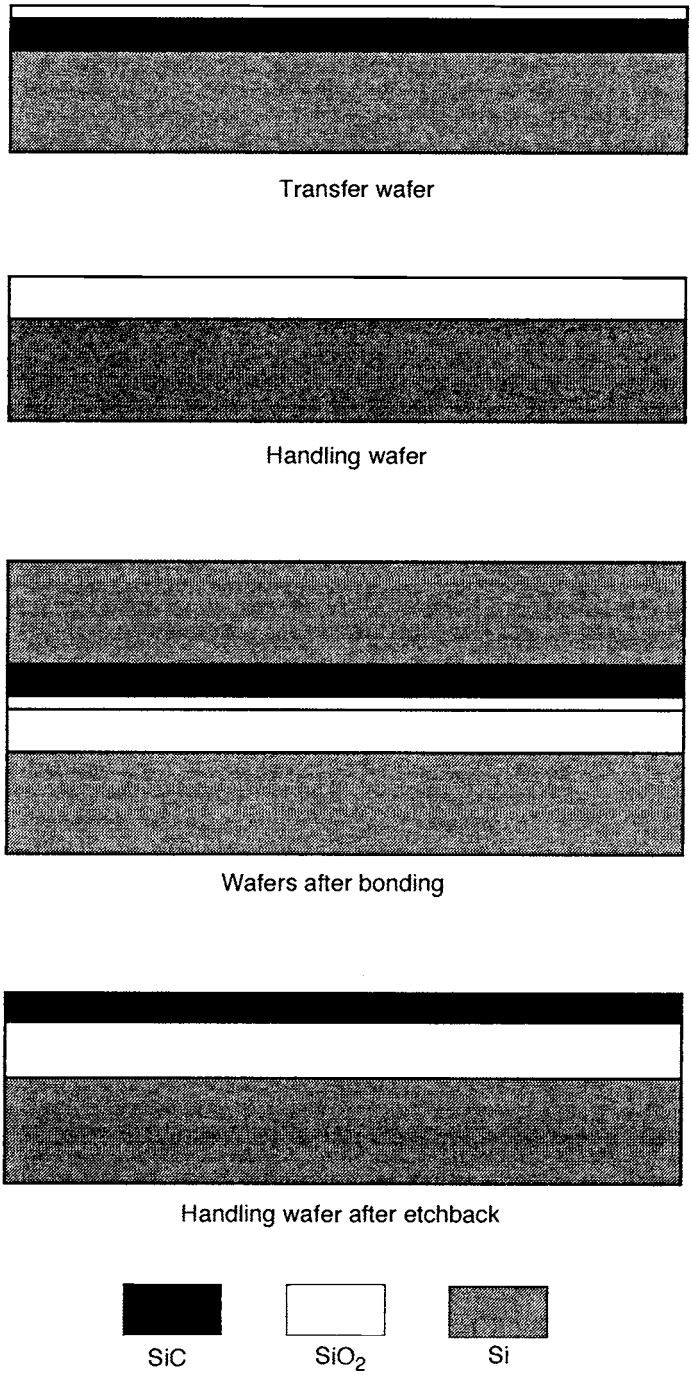


Fig. 4.6. .Schematic diagram of the 3C-SiC-on-SiO_2 wafer bonding process.

a Si wafer. A thermal oxide is also grown on the 3C-SiC film. The two SiO_2 surfaces are chemically treated and bonded. KOH etching is used to remove the SiC-coated transfer wafer, leaving a 3C-SiC-on-SiO_2 structure on the handling wafer. Unfortunately, only 30% of the SiC film area remains bonded after KOH etching, which is too low for batch processing.

A fourth improved wafer-bonding technique has been developed to produce 3C-SiC-on-SiO_2 structures on 4-in. Si wafers with areal transfer yields of up to 80%.[59] Unlike the third process mentioned above, the improved process utilizes a polysilicon-to-polysilicon bond to create the desired structure. A schematic of the process is shown in Fig. 4.7. The process begins with epitaxial growth of a 0.5-μm-thick 3C-SiC film on a 4-in. Si wafer, hereafter called the transfer wafer. The 3C-SiC film is chemically cleaned and thermally oxidized to produce a 0.3-μm-thick SiO_2 film on the 3C-SiC. A 0.5-μm-thick polysilicon film is then grown on the SiO_2 layer. The polysilicon is completely oxidized, which produces a total oxide thickness of 1.44 μm on the 3C-SiC film. Following the second thermal oxidation, a 2.0-μm-thick polysilicon film is deposited on the wafer and polished to a mirror finish by chemical-mechanical polishing. A 1.5-μm-thick thermal oxide is grown on a second wafer, hereafter called the handling wafer. A 2.0-μm-thick polysilicon film is deposited on top of the thermal oxide, and the polysilicon surface is polished to a mirror finish.

The polysilicon-to-polysilicon bonding process consists of three steps: (1) a prebond surface treatment, (2) room temperature bonding, and (3) high-temperature annealing. Step 1 is a standard wet chemical cleaning of the polished polysilicon surfaces. Step 2 places the polysilicon surfaces firmly together and uses van der Waals attraction to keep the surfaces in contact. Step 3 is an anneal at 1100°C in nitrogen for at least 5 h.

Following the anneal, the bonded pair is submerged in EDP, which removes the transfer wafer, leaving a 3C-SiC-on-SiO_2 structure atop the handling wafer. A SEM micrograph of the structure is shown in Fig. 4.8. The 3C-SiC film is too thin for most MEMS applications, so a thick (<1.0 μm) 3C-SiC film is homoepitaxially grown on the 3C-SiC-on-SiO_2 substrate. The substrate is prepared for homoepitaxial growth by a four-step process: (1) mechanical polishing, (2) thermal oxidation, (3) chemical etching, and (4) in-situ high-temperature hydrogen etching. Recall from previous discussions that the region of highest defect density in 3C-SiC films grown on Si substrates is near the SiC/Si interface. By transferring the heteroepitaxial film from the transfer wafer to the handling wafer, the bonding process brings the region of highest defect density to the surface of the bonded structure. Mechanical polishing with a SiC-based slurry is used to remove this region. Any defects created during the polishing step are removed by the thermal oxidation and chemical etching steps. Any residue on the 3C-SiC surface is removed immediately before homoepitaxial growth by the in-situ hydrogen etch, which is performed at 1000°C.

Figure 4.9 shows a TEM micrograph of the homoepitaxial 3C-SiC film grown on the 3C-SiC-on-SiO_2 structure. The TEM shows a clear reduction of crystalline defects in the homoepitaxial film, as compared with the underlying 3C-SiC, which was grown by the conventional 3C-SiC-on-Si heteroepitaxial process. The TEM observations were confirmed by rocking curve XRD. Although the physical mechanism responsible for the reduction of defects has not yet been identified, it is believed that the surface treatment prior to homoepitaxial growth plays a key role.

For the first time, large-area 3C-SiC can be produced atop insulating sacrificial substrates, making electrically isolated 3C-SiC surface-micromachined devices possible. The reduced defect density will improve the performance of both mechanical and electronic structures in 3C-SiC-based MEMS devices. The biggest impact, however, may be in the high-temperature and high-power microelectronics field, where low defect-density SiC films on large-area substrates are required. This technology may rekindle interest in 3C-SiC as a high-temperature semiconductor for microelectronics.

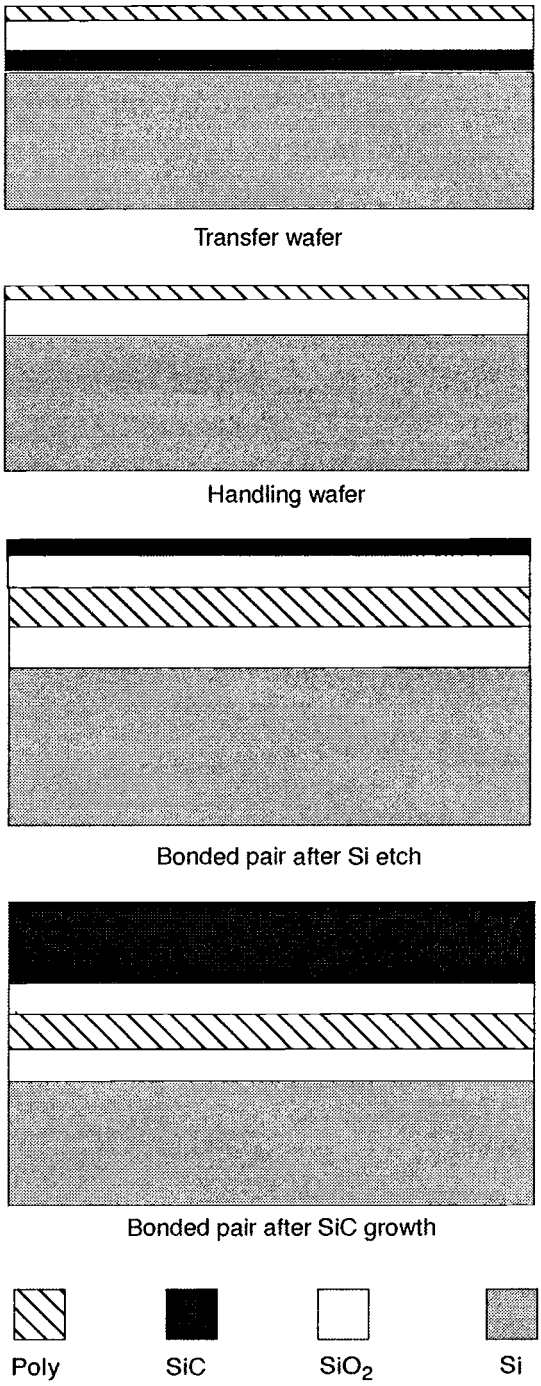


Fig. 4.7. Schematic diagram of the improved 3C-SiC-on-insulator fabrication process.

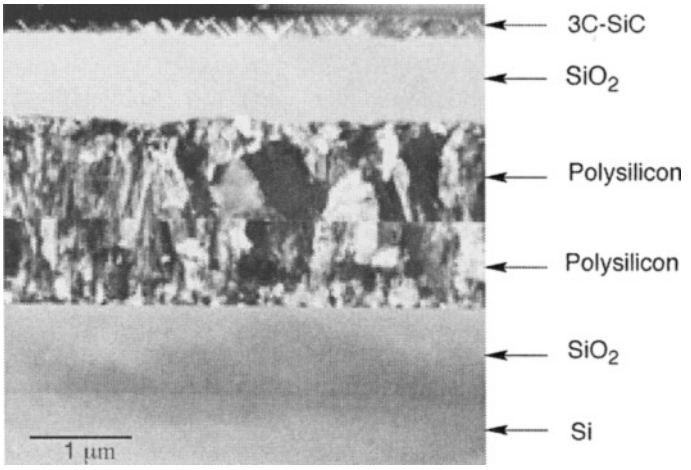


Fig. 4.8. TEM micrograph of a 3C-SiC-on-insulator structure, highlighting the polysilicon-to-polysilicon bonding interface.[59]

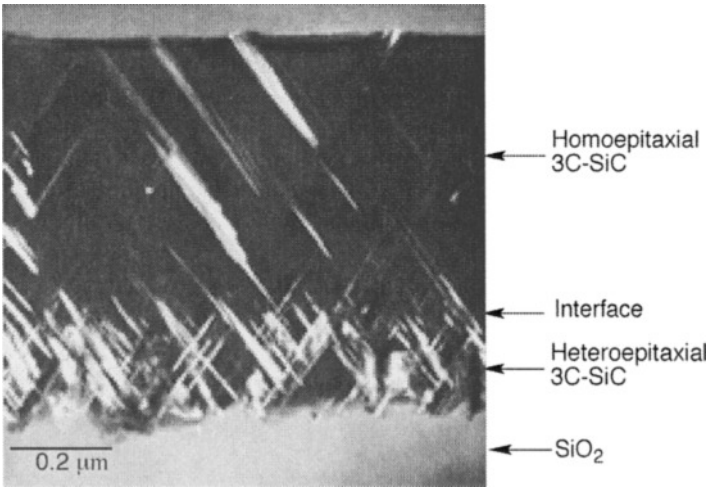


Fig. 4.9. TEM micrograph of a homoepitaxial 3C-SiC film grown on a 3C-SiC-on-insulator substrate, showing the decreased defect density in the homoepitaxial film.[59]

4.7 SiC Devices and Applications

4.7.1 Introduction

SiC research has concentrated on understanding the material properties and processing techniques required to use SiC as a high-temperature material for electronics. Only recently have the processing techniques become available to fabricate basic SiC-MEMS devices. This section will present descriptions of a representative collection of devices that use SiC as the key material for electronic and mechanical components. The devices have at least one common characteristic, high-temperature functionality, which makes them well suited for aerospace applications.

One of the principal focus areas for SiC-MEMS has been high-temperature SiC sensor systems for gas turbine engines. To improve the efficiency, power-to-weight ratios, emissions, cost, and

safety of gas turbine engines, new generation designs will require improved measurement, control, and sensor systems. These systems will often be located in or near the hot-gas flow path, which reaches temperatures above 350°C. Sensors are needed to measure combustor temperature, rotor and stator temperatures, internal cooling temperatures, cooling flow and temperature, pressure, hot gas path leakage, and coolant leakage. Sensors are also needed to control the engine near its combustion limit in order to maximize fuel combustion and minimize NO_x emissions.

Increasing the efficiency of the combustion process, through conversion of electronic control and sensor systems from Si- to SiC-based devices, will result in increased fuel economy and reduced emissions. In addition, the overall weight of an aircraft will be reduced by the elimination of the packaging, wiring, and connectors necessary to link sensor systems with control electronics. This weight reduction will directly translate to increased range and lower fuel costs.

SiC-based MEMS devices will also benefit manned and unmanned spacecraft. Currently, unmanned spacecraft require thermal radiators to dissipate heat generated by onboard Si-based electronics. Implementation of SiC-based electronic systems that can operate at temperatures above 400°C will eliminate the need for thermal radiators, thus reducing the overall weight of the spacecraft. Additionally, SiC electronics and sensors, being much less susceptible to radiation damage than their Si counterparts, will not require as much radiation shielding, thus reducing spacecraft weight even further. Reduced spacecraft weight and increased operating temperature and radiation resistance of onboard electronics and sensor systems will dramatically increase the functionality of unmanned spacecraft. Exploratory spacecraft will be able to more aggressively probe harsh planetary environments. Communication satellites will have extended operating lifetimes and will be able to carry larger numbers of sensitive onboard electronic systems.

In terms of propulsion systems, many high-temperature sensor systems will find applications in rocket motors. High-temperature sensors to monitor combustion temperature, pressure, and by-products are needed. Gas sensors for hydrogen and HCl are required for chemical detection around the launch pad and as part of in-flight safety systems. All of these systems can be fabricated from SiC.

The remainder of this section presents examples of SiC-based sensors for high-temperature MEMS applications. These sensors constitute the basic units of MEMS for the aeronautic and space applications mentioned previously. These sensor prototypes have demonstrated high-temperature functionality but require the development of high-temperature wire bonding and packaging technologies, and integration with on-chip SiC electronics before deployment in aerospace systems can occur.

4.7.2 Temperature Sensors

Some of the first SiC-based sensors were fabricated from poly-SiC because of the many methods available for deposition and the ability to deposit poly-SiC on many different substrates. One of the first sensors was a SiC thin-film thermistor.[60] The sensor was fabricated from radio frequency (RF) sputter-deposited SiC, using an Ar sputtered SiC target and a deposition pressure of 20 mtorr. The SiC film was deposited on an alumina substrate that was maintained at a deposition temperature of 650°C. Au-Pt comb-shaped thick films were used as electrodes. Various packages for the thermistor were developed and tested at temperatures between 0°C and 500°C. It was reported that the thermistor constant (B) increased linearly with temperature over the entire temperature range. Also, when compared with conventional metal-oxide thermistors, the temperature coefficient of resistance for the SiC thermistors decreased more slowly with increasing temperature. Additionally, the sensor exhibited good thermal stability and a rapid thermal response. A resistance change of only 5% was observed for a continuous 1000 h test at 500°C.

The fabrication of a SiC-based resistive temperature sensor has also been reported.[61] SiC films for this sensor were prepared by plasma-assisted chemical vapor deposition (PACVD), using dichlorosilane and methane as the Si and C source gases. Stoichiometric poly-SiC films were obtained with a dichlorosilane flow rate of 15 sccm, a methane flow rate of 5 to10 sccm, RF power of 70 W, and a substrate temperature of 750°C. The poly-SiC films were deposited on thermally oxidized Si substrates and patterned into resistors by RIE. To reduce thermal conduction to the substrate, the resistors were made into cantilever structures by bulk micromachining the Si substrate in KOH. The contact pads remained anchored to the substrate. The resistors were tested at temperatures between 0°C and 300°C, and exhibited a temperature coefficient of resistance of −1800 ppm/°C. The sensor was placed into a nitrogen flow stream to evaluate its performance as a gas mass flow sensor. The output characteristics of the SiC sensor showed a square-root dependence, which is characteristic of micromachined flow sensors.[62] The output sensitivity of the SiC mass flow sensor was 0.05 mV/sccm.

4.7.3 Gas Sensors

Interest in high-temperature semiconductors for solid-state gas and chemical sensors stems from the need for closed-loop control of the combustion process in gasoline and gas turbine engines. This control is needed to increase fuel and combustion efficiency, thereby decreasing emissions of incompletely combusted hydrocarbons, nitrous oxides (NO_x), and CO_2. Currently, the dominant polytype used for SiC-based gas sensors is 6H-SiC, because 6H-SiC has the highest crystal quality of all the commercially available polytypes. The 6H-SiC gas sensors are based on simple Schottky diode and metal-oxide-semiconductor (MOS) structures.

A common method for fabricating SiC-based gas sensors uses 6H-SiC MOS capacitors.[63] Called MOSiC (metal-oxide-SiC) structures, these sensors use catalytic metals such as Pt and Pd as the gate metals. The sensors work on the principle that hydrogen atoms or hydrogen-containing radicals diffusing through the gate will collect at the metal-oxide interface and form a dipole layer that lowers the flat band voltage of the MOS capacitor. Hydrogen sensors have been fabricated using Si MOS structures, but the operating temperature is limited to below 250°C. However, many hydrocarbons dissociate at temperatures between 350°C and 500°C, making SiC the better material for solid-state MOS hydrocarbon gas sensors.

The MOSiC gas sensor is made using a 6H-SiC substrate, on which a 40-nm-thick SiO_2 layer is thermally grown.[64] Approximately 50 nm of Pt is then deposited by either e-beam evaporation or magnetron sputtering on the SiO_2 surface, and patterned to form the MOSiC structure. This sensor has been tested at temperatures as high as 800°C without failing. At 450°C, the sensor was able to detect the presence of saturated hydrocarbons such as methane, propane, ethane, and butane at concentrations below 0.6 vol%. Moreover, the sensor can be used at elevated temperatures in vacuum or in air. The response of the sensor increases with an increasing number of carbon atoms in the detected molecule. With an operating temperature well above 500°C, this sensor is well suited for deployment in exhaust streams near the combustion chamber.

A second type of SiC gas sensor is based on the 4H-SiC Schottky diode.[65] The sensor is constructed from 4- to 5-μm-thick 4H-SiC films epitaxially grown on 4H-SiC substrates. Approximately 400 Å of Pd is sputter deposited and patterned to form circular contacts to the 4H-SiC. The sensor has been tested over a temperature range of 0°C to 400°C in a hydrocarbon environment and exhibits a sensitivity of 300 ppm to hydrogen and propylene.

A third type of SiC gas sensor uses porous SiC.[66] The sensor consists of a chromium (Cr) grid that is evaporated and patterned onto a layer of porous SiC. The porous layer is formed by PEC etching of a n-type 6H-SiC wafer. The thickness of the porous layer was not reported. A nickel

contact is deposited on the nonporous backside of the 6H-SiC wafer. A voltage is applied to the Cr grid, which sets up a voltage in the porous SiC region. Each hydrocarbon species has a characteristic dissociation voltage, so by varying the grid voltage, a specific hydrocarbon species can be forced to dissociate. The concentration of each species is determined by measuring the magnitude of current flow across the device for a given grid voltage. This sensor is an improvement over the previously mentioned sensors because it does not require high temperatures to operate, yet it can operate at high temperatures. Tests using methane and propane conducted at temperatures between 200°C and 500°C verified these capabilities.

4.7.4 Pressure Sensors

A 6H-SiC-based pressure sensor that exhibits stable operation at 500°C has recently been developed.[67] The sensor uses n-type 6H-SiC piezoresistors on a p-type 6H-SiC diaphragm. The diaphragm is fabricated using the PEC etching process of 6H-SiC described previously and detailed elsewhere.[38] During development of the etching process, a p-type etch stop for n-type etching was demonstrated. The fabrication process begins with epitaxial growth of a p-type 6H-SiC film on an n-type 6H-SiC wafer, followed by epitaxial growth of an n-type 6H-SiC film on the p-type epilayer. PEC is used to bulk-micromachine the n-type 6H-SiC wafer into 50-μm-thick diaphragms, and pattern the n-type 6H-SiC into piezoresistors. Multilayer Ti/TiN/Pt/Au metal contacts are used in conjunction with Au wire bonding to package the device. A full-scale output of 40.66 mV at 1000 psi and 25°C, decreasing to 20.33 mV at 500°C, was reported. The device has a gauge factor temperature coefficient of -0.19%/°C at 100°C and -0.11%/°C at 500°C. Despite the attractiveness of an "all-SiC" high-temperature pressure sensor, small 6H-SiC wafer diameters and nonstandardized fabrication processes currently limit the commercial application of this design.

A second pressure sensor design utilizes 3C-SiC films grown on SOI substrates combined with Si bulk micromachining to produce dielectrically isolated 3C-SiC piezoresistors on a thick Si membrane.[68] The piezoresistors are realized by conventional heteroepitaxial growth of 3C-SiC on the SOI substrates, followed by photolithography and reactive ion etching. The buried oxide in the SOI substrate prohibits leakage currents between the 3C-SiC piezoresistors and the Si substrate, a problem that is magnified at high temperatures. For each sensor, a circular Si diaphragm was micromachined to a thickness of 100 μm using RIE in a SF_6/O_2 plasma, and four 3C-SiC piezoresistors were patterned using SF_6/O_2 RIE. A thermally grown capping oxide over the piezoresistors was used for electrical isolation. Sputter-deposited TiWN alloy was used as high-temperature ohmic contacts. The sensor was tested in the temperature range between 25°C and 400°C, and at pressures up to 500 kPa. The sensor showed a linear output voltage over the applied pressure range and a sensitivity of –0.16%/°C at 400°C. Recently, a hermetic pressure sensor capsule for the SiC-based pressure sensor was constructed and tested.[2] This sensor design takes advantage of well-established Si micromachining techniques. However, the hybrid SiC/Si device may suffer from SiC/Si thermal mismatch effects, which may degrade long-term device performance and lifetime.

4.7.5 Protective Coatings

Micromachining is a technology that is not limited to the production of only microsensors and microactuators, but can be used to batch fabricate component structures for macroscale systems. A fine example is the development of micromachined silicon fuel atomizers as an alternative for conventional metal atomizers in gas turbine engines. A process to batch fabricate silicon atomizers from 4-in. Si wafers using deep reactive ion etching (DRIE) has been demonstrated.[69] A schematic and SEM micrograph of a Si-micromachined atomizer are shown in Figs. 4.10(a) and 4.10(b). Comparative tests of conventional metal and Si atomizers indicated that the Si atomizers

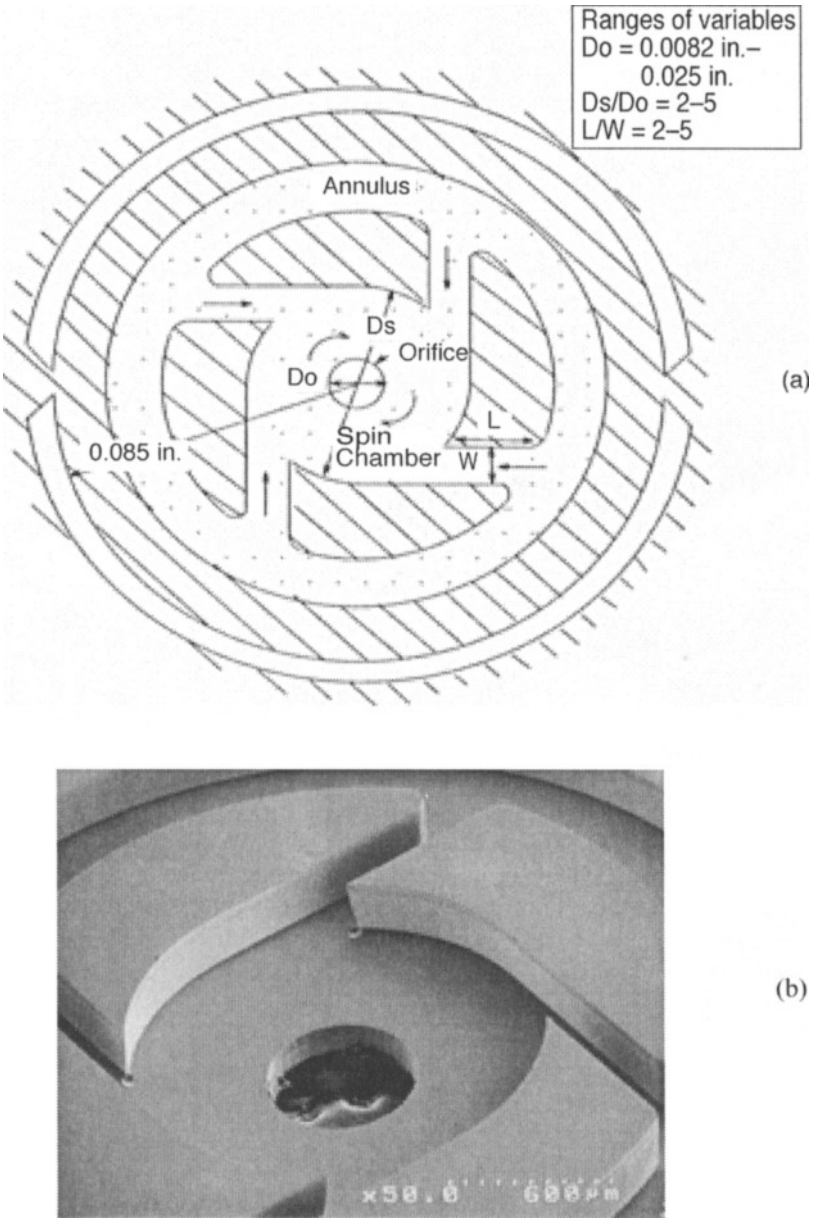


Fig. 4.10. (a) Schematic plan-view diagram of a fuel atomizer. (b) SEM micrograph of a silicon micromachined fuel atomizer.[69]

outperformed conventional atomizers, especially at low pressures. Unfortunately, the Si atomizers lacked the erosion resistance of conventional atomizers. A 3C-SiC coating was grown on the Si atomizers in hopes of increasing the erosion resistance of the structures.[70] Because the atomizers were fabricated from single crystalline Si wafers, the high-temperature deposition process with carbonization was used. This was the first attempt at growing 3C-SiC on high-aspect-ratio

Si topographies. Conformal coverage of the atomizer surfaces was achieved, with thicknesses ranging from 1.5 μm on the top surfaces to 0.5 μm on the swirl chamber floor.

The performance of the SiC-coated atomizers, in terms of flow rate, Sauder Mean Diameter (SMD), and spray angle, was compared with uncoated silicon atomizers. The test fluid was MIL-C-7024D Type II jet fuel stimulant. The flow rates for the coated atomizers were consistently 7%–11% higher than the uncoated atomizers. The SMDs were the same for both atomizer types. Higher spray angles at both low pressure (14 psi) and high pressure (100 psi) were also observed for the 3C-SiC coated atomizers. The spray angle is a rough gauge of atomizer efficiency, with wider spray angles indicating better performance.

To qualitatively determine the improvement of the erosion resistance of the SiC coated atomizers, a 30-h erosion test was performed. The test fluid consisted of jet fuel mixed with abrasives like iron oxide, quartz, and Arizona road dust. The test fluid was pressurized to 150 psi. SEM analysis of the region near the exit orifice was performed, since this region is subject to the highest erosive damage. The uncoated atomizers showed significant edge rounding near the exit orifice; whereas, the coated atomizers showed no evidence of edge rounding.

4.8 Conclusions

The field of SiC MEMS for high-temperature applications has advanced beyond material characterization and process development toward the fabrication of prototype devices that have been tested in harsh, high-temperature environments. Development of sensors and actuators will most certainly continue, and as early prototypes are successfully tested, more application areas and designs will be conceived. This will result in additional materials characterization and process development, as more sophisticated device structures are required.

Advances in two closely related areas will be required before SiC MEMS can make an impact in commercial and military applications: (1) reliable SiC electronics, and (2) viable high-temperature wire bonding and packaging technologies. Fortunately, research on SiC-based electronics for high-temperature applications has been and will continue to be a priority, and as larger area, defect-free 6H-SiC and 4H-SiC wafers become available, improvements in device performance will occur. Advances in wire bonding techniques and packaging technologies for high-temperature applications, which have been few to date, are expected to grow with advances in SiC electronics, sensor, and actuator technologies. By all indications, the future looks bright for SiC-based MEMS for applications in harsh environments.

4.9 References

1. P. Gluche, *MST News (September* 1997), p. 12.
2. G. Krotz, *MST News* (September 1997), p. 17.
3. J.C. Angus and C.C. Hayman, "Low-Pressure, Metastable Growth of Diamond and 'Diamondlike' Phases," *Science* 214, 913–921 (August 1988).
4. A. Masood, M. Aslam, M.A. Tamor, and T.J. Potter, "Techniques for Patterning of CVD Diamond Films on Non-Diamond Substrates." *J. Electrochem. Soc.* 138 (11). L67–L68 (1991).
5. I. Taher, M. Aslam, M.A. Tamor, T.J. Potter, and R.C. Elder, "Piezoresistive Microsensors Using P-Type Diamond Films," *Sensors and Actuators A* 45, 35–43 (1994).
6. G.R. Fisher and P. Barnes, "Towards a Unified View of Polytypism in Silicon Carbide," *Philosophical Mag. B* 61 (2), 217–236 (February 1990).
7. M. Capano and R. Trew, "Silicon Carbide Electronic Materials and Devices," *Mater. Res. Soc. Bull.* 22 (3), 19–22 (March 1997).
8. S. Nishino, J.A. Powell, and H.A. Will, "Production of Large-Area Single-Crystal Wafers of Cubic SiC for Semiconductor Devices, *Appl. Phys. Lett.* 42 (5), 460–462 (1 March 1983).

9. S. Nishino and J. Saraie, "Heteroepitaxial Growth of Cubic SiC on a Si Substrate Using the Si_2H_6-C_2H_2-H_2 System," in *Amorphous and Crystalline Silicon Carbide II,* edited by M.M. Rahman, C.Y. Yang, and G.L. Harris (Springer-Verlag, Berlin, 1989), pp. 8–13.
10. C.C. Chiu, S.B. Desu, G. Chen, C.Y. Tsai, and W.T. Reynolds, Jr., "Deposition of Epitaxial Beta-SiC Films on Porous Si(100) from MTS in a Hot-Wall LPCVD Reactor," *J. Mater. Res.* 10 (5), 1099–1107 (May 1995).
11. I. Golecki, F. Reidinger, and J. Marti, "Epitaxcial Monocrystalline SiC Films Grown on Si by Low-Pressure Chemical Vapor Deposition at 750°C," *Wide Band Gap Semiconductors Symposium* (Mat. Res. Soc., Pittsburgh, 1992), pp. 519–524.
12. J.A. Powell, L.G. Matus, and M.A, Kuczmarski, "Growth and Characterization of Cubic SiC Single-Crystal Films on Si," *J. Electrochem. Soc.* 134 (6), 1558–1565 (June 1987).
13. C.A. Zorman, A.J. Fleischman, A.S. Dewa, M. Mehregany, C. Jacob, S. Nishino, and P. Pirouz, "Epitaxial-Growth of 3C-SiC Films on 4 inch diam (100) Silicon-Wafers by Atmospheric-Pressure Chemical-Vapor-Deposition," *J. Appl. Phys.* 7 (8), 5136–5138 (15 October 1995).
14. J.P. Li and A.J. Steckl, "Nucleation and Void Formation Mechanisms in SiC Thin Film Growth on Si by Carbonization," *J. Electrochem. Soc.* 142 (2), 634–641 (February 1995).
15. Y.H. Seo, K.C. Kim, H.W. Shim, K.S. Nahm, E.K. Suh, H.J. Lee, Y.G. Hwang, D.K. Kim, and B.T. Lee, in *Extended Abstracts of the International Conference on Silicon Carbide, III-nitrides, and Related Materials - 1997* (Stockholm, Sweden, 31 August–5 September 1997), p. 215.
16. H.W. Shim, K.C. Kim, Y.H. Seo, K.S. Nahm, E.K. Suh, and H.J. Lee, in *Extended Abstracts of the International Conference on Silicon Carbide, III-nitrides, and Related Materials - 1997* (Stockholm, Sweden, 31 August–5 September 1997), p. 577.
17. C. Hagiwara, K.M. Itoh, J. Muto, H. Nagasawa, K. Yagi, H. Harima, K. Mizoguchi, and S. Nakashima, in *Extended Abstracts of the International Conference on Silicon Carbide, III-nitrides, and Related Materials - 1997* (Stockholm, Sweden, 31 August–5 September 1997), p. 331.
18. K. Kamimura, K. Koike, H. Ono, T. Homma, Y. Onuma, and S. Yonekubo, in *Amorphous and Crystalline Silicon Carbide IV,* edited by C.Y. Yang, M.M. Rahman, and G.L. Harris (Springer-Verlag, Berlin, 1991), p. 259.
19. J. Kobayashi, S. Yonekubo, K. Kamimura, and Y. Onuma, in *Silicon Carbide and Related Materials,* Vol. 142, edited by S. Nakashima, H. Matsunami, S. Yoshida, and H. Harima (IOP Publishing Ltd., Bristol, UK, 1995), p. 229.
20. Y. Onuma, S. Miyashita, Y. Nishibe, K. Kamimura, and K. Tezuka, "Thin Film Transistors Using Polycrystalline SiC," in *Amorphous and Crystalline Silicon Carbide II,* edited by M.M. Rahman, C.Y. Yang, and G.L. Harris (Springer-Verlag, Berlin, 1989), pp. 212–216.
21. S. Nishino and J. Saraie, "Heteroepitaxial Growth of Cubic SiC on a Si Substrate Using the Si_2H_6-C/H_2-H_2 System," in *Amorphous and Crystalline Silicon Carbide II,* edited by M.M. Rahman, C.Y. Yang, and G.L. Harris (Springer-Verlag, Berlin, 1989), pp. 8–13.
22. H. Nagasawa and Y. Yamaguchi, "Atomic Level Epitaxy of 3C-SiC by Low Pressure Vapour Deposition with Alternating Gas Supply," *Thin Solid Films* 225 (1-2), 230–234 (1993).
23. S. Roy, C.A. Zorman, C.H. Wu, A.J. Fleischman, and M. Mehregany, "XRD and XTEM Investigation of Polycrystalline Silicon Carbide on Polycrystalline Silicon," *Proceedings of the 1996 MRS Fall Meeting* (Materials Research Society, Pittsburgh, 1997), pp. 81–86.
24. T. Kamins, "Thermal Oxidation of Polycrystalline Silicon Films," *Metal. Trans. of AIME* 2 (8), 2292–2294 (1971).
25. J. Adamczewska and T. Budzynski, "Stress in Chemically Vapour-Deposited Silicon Films," *Thin Solid Films* 113 (4), 271–295 (30 March 1984).
26. L. Wei, M. Vaudin, C.S. Hwang, G. White, J. Xu, and A.J. Steckl, "Heat Conduction in Silicon Thin Films: Effect of Microstructure," *J. Mater. Res.* 10 (8), 1889–1896 (1995).
27. A.J. Fleischman X. Wei, C.A. Zorman, and M. Mehregany, in *Extended Abstracts of the International Conference on Silicon Carbide, III-nitrides, and Related Materials - 1997* (Stockholm, Sweden, 31 August–5 September 1997), p. 515.

28. W.E. Wagner III, R. Filozof, S. Gong, F. Miller, D. Young, and M. White, "Contqamination of Si Devices by SiC," in *Silicon Carbide and Related Materials 1995*, Vol. 142, edited by S. Nakashima, H. Matsunami, S. Yoshida, and H. Harima (IOP Publishing Ltd., Bristol, UK, 1996), pp. 1107–1110.
29. J. Tan, M.K. Das, J.A. Cooper, Jr., and M.R. Melloch, "Metal-Oxide-Semiconductor Capacitors Formed by Oxidation of Polycrystalline Silicon on SiC," *Appl. Phys. Lett.* 70 (17), 2280–2281 (28 April 1997).
30. L. Porter and R. Davis, "A Critical-Review of Ohmic and Rectifying Contacts for Silicon-Carbide," *Mat. Sci. and Eng. B* 34 (2-3), 83–105 (1995).
31. H. Morkoc, S. Strite, G.B. Gao, M.E. Lin, B. Svendlov, and M. Burns, "Large-Band-Gap SiC, III-V Nitride, and II-VI ZnSe-based Semiconductor Device Technologies," *J. Appl. Phys.* 76 (3), 1363–1398 (1 August 1994).
32. J. Sugiura, W.J. Lu, L.C. Cadien, and A.J. Steckl, "Reactive Ion Etching of SiC Thin Films Using Fluorinated Gases," *J. Vac. Sci. Technol. B* 4 (1), 349–354 (1 January 1986).
33. R. Padiyath, R.L. Wright, M.I. Chaudhry, and S.V. Babu, "Reactive Ion Etching of Monocrystalline, Polycrystalline, and Amorphous-Silicon Carbide in CF_4O_2 Mixtures," *Appl. Phys. Lett.* 58 (10), 1053–1055 (11 March 1991).
34. C. Richter, K. Espertshuber, C. Wagner, M. Eickhoff, and G. Krotz, "Rapid Plasma Etching of Cubic SiC Using NF_3O_2 Gas Mixtures," *Mat. Sci. and Eng. B* 46 (1-3), 160–163 (1997).
35. W. Reichert, D. Stefan, E. Obermeier, and W. Wondrak, "Fabrication of Smooth Beta-SiC Surfaces by Reactive Ion Etching Using a Graphite Electrode," *Mat. Sci. and Eng. B* 46 (1-3), 190–194 (April 1997).
36. P. Yih and A.J. Steckl, "Effects of Hydrogen Additive on Obtaining Residue-free Reactive Ion Etching of Beta-SiC in Fluorinated Plasmas, *J. Electrochem. Soc.* 140 (6), 1813–1824 (6 June 1993).
37. W.S. Pan and A.J. Steckl, "Reactive Ion Etching of SiC Thin Films by Mixtures of Fluorinated Gases and Oxygen," J. *Electrochem. Soc.* 137 (1), 212–220 (January 1990).
38. J.S. Shor, R.S. Okojie, and A.D. Kurtz, in *Silicon Carbide and Related Materials*, Vol. 137 (IOP Publishing Ltd, Bristol, UK, 1994), p. 523.
39. J.S. Shor, R.M. Osgood, and A.D. Kurtz, "Photoelectrochemical Conductivity Selective Etch Stops for SiC," *Appl. Phys. Lett.* 60 (8), 1001–1003 (24 February 1992).
40. R. Okojie, A. Ned, A. Kurtz, and W. Carr, "Alpha (6H)-SiC Pressure Sensors for High Temperature Applications," *Proceedings of the 1995 9th Annual International Workshop on Microelectromechanical Systems*, edited by M. Allen and M. Reed (San Diego, CA, 11–15 February 1996), pp. 146–149.
41. K. Chandra, C.A. Zorman, and M. Mehregany, in *Extended Abstracts of the International Conference on Silicon Carbide, III-nitrides, and Related Materials - 1997* (Stockholm, Sweden, 31 August –5 September 1997), p. 333.
42. M.G. Allen, M. Mehregany, R.T. Howe, and S.D. Senturia, "Microfabricated Structures for the In-situ Measurement of Residual Stress, Young's Modulus, and Ultimate Strain of Thin Films," *Appl. Phys. Lett.* 5 (4), 241–243 (27 July 1987).
43. M. Mehregany, L. Tong, L. Matus, and D. Larkin, "Internal Stress and Elastic Modulus Measurements on Micromachined 3C-SiC Thin Films," *IEEE Trans. Elect. Dev.* 44 (1), 74–79 (January 1997).
44. Y. Yamaguchi, H. Nagasawa, T. Shoki, and N. Annaka, "Properties of Heteroepitaxial 3C-SiC Films Grown by LPCVD," *Proceedings of the 1995 8th International Conference on Solid-State Sensors and Actuators, and Eurosensors IX* (Stockholm, Sweden, 25–29 June 1995), pp. 190–193.
45. K. Murooka, I. Higashikawa, and Y. Gomei, "Improvement of the Young Modulus of SiC Film by Low-Pressure Chemical-Vapor-Deposition with B_2H_6 Gas," *Appl. Phys. Lett.* 69 (1), 37–39 (1 July 1996).
46. C. Su, A. Fekade, M. Spencer, and M. Wuttig, "Stresses in Chemical Vapor Deposited Epitaxial 3C-SiC Membranes," *J. Appl. Phys.* 77 (3), 1280–1283 (1 February 1995).
47. L. Tong, M. Mehregany, and L.G. Matus, "Mechanical Properties of 3C Silicon Carbide," *Appl. Phys. Lett.* 60 (24), 2992–2994 (15 June 1992).

48. C. Su, M. Wuttig, M. Fekade, and M. Spencer, "Elastic and Anelastic Properties of Chemical Vapor Deposited Epitaxial 3C-SiC," *J. Appl. Phys.* 77 (11), 5611–1615 (1 June 1995).
49. K. Chandra, *Characterization of Residual Stress and Elastic Modulus of Silicon Carbide Films Grown on Silicon by APCVD,* M.S. thesis, Case Western Reserve University, May 1997.
50. A.J. Fleischman, S. Roy, C.A. Zorman, M. Mehregany, and L. G. Matus, "Polycrystalline Silicon Carbide for Surface Micromachining," *Proceedings of the 9th Annual International Workshop on Microelectromechanical Systems,* edited by M. Allen and M. Reed (San Diego, CA, 11–15 February 1996), pp. 234–238.
51. A.J. Fleischman, S. Roy, C.A. Zorman, and M. Mehregany, in *Extended Abstracts of the International Conference on Silicon Carbide, III-nitrides, and Related Materials - 1997* (Stockholm, Sweden, 31 August –5 September 1997), p. 643.
52. J. Camassel, C. Dezauzier, L. DiCioccio, J. Stoemenos, J. Bluet, S. Contreras, J. Robert, and T. Brillon, "Investigation of Structural, Optical and Electrical Properties of Cubic 3C-SiC Deposited on SOI," in *Silicon Carbide and Related Materials 1995, Vol. 142,* edited by S. Nakashima, H. Matsunami, S. Yoshida, and H. Harima (IOP Publishing Ltd., Bristol, UK, 1996), pp. 453–456.
53. W. Reichert, E. Obermeier, and J. Stoemenos, "Beta-SiC Films on SOI Substrates for High Temperature Applications," *Diamond and Related Materials* 6 (10), 1448–1450 (August 1997).
54. W. Reichert, R. Lossy, J.M. Gonzalez Sirgo, E. Obermeier, and J. Stoemenos, "Beta-SiC Deposited on SIMOX Substrates: Characterization of the SiC/SOI System at Elevated Temperatures," in *Silicon Carbide and Related Materials 1995, Vol. 142,* edited by S. Nakashima, H. Matsunami, S. Yoshida, and H. Harima (IOP Publishing Ltd., Bristol, UK, 1996), pp. 129–132.
55. L. Di Cioccio, Y. Le Tiec, F. Letertre, C. Jaussaud, and M. Bruel, "Silicon-Carbide on Insulator Formation Using the Smart Cut Process," *Electron. Lett.* 32 (12), 1144–1145 (6 June 1996).
56. Q. Tong, U. Gosele, C. Yuan, A.J. Steckl, and M. Reiche, "Silicon Carbide Wafer Bonding," *J. Electrochem. Soc.* 142 (1), 232–236 (1 January 1995).
57. L. Di Cioccio, F. Letertre, Y. Le Tiec, A.M. Papon, C. Jaussaud, and M. Bruel, "Silicon Carbide on Insulator Formation by the Smart-Cut$^{(R)}$ Process," *Mat. Sci. and Eng. B* 46 (1), 349–356 (April 1997).
58. L. Di Cioccio, C. Jassaud, Y. Le Tiec, and M. Bruel, in *Extended Abstracts of the International Conference on Silicon Carbide, III-nitrides, and Related Materials - 1997* (Stockholm, Sweden, 31 August–5 September 1997), p. 536; and private communications.
59. K. Vinod, C.A. Zorman, and M. Mehregany, "Novel SiC on Insulator Technology Using Wafer Bonding," *Proceedings of the 1997 International Conference on Solid State Sensors and Actuators,* edited by K. Wise and S. Senturia (Chicago, Il, 16–19 June 1997), pp. 653–656.
60. T. Nagai and M. Itoh, "SiC Thin-Film Thermistors," *IEEE Trans. on Ind. Applications* 26(6), 1139–1143 (November–December 1990).
61. K. Kamimura, T. Miwa, T. Sugiyama, T. Ogawa, N. Nakao, and Y. Onuma, "Preparation of Polycrystalline SiC Thin Films and Its Application to Resistive Sensors," in *Silicon Carbide and Related Materials,* Vol. 142, edited by S. Nakashima, H. Matsunami, S. Yoshida, and H. Harima (IOP Publishing Ltd., Bristol, UK, 1996), pp. 825–828.
62. M. Esashi, H. Kawai, and K. Yoshimi, in *The Trans. of IEIC Japan, C-2* 11, 738 (1992).
63. A. Arbab, A. Spetz, and I. Lundstrum, "Gas Sensors for High Temperature Operation Based on Metal Oxide Silicon Carbide (MOSiC) Devices," *Sensors and Actuators* B15 (1-3) 19–23 (August 1993).
64. A. Arbab, A. Spetz, Q. Wahab, M. Willander, and I. Lundstrom, "Chemical Sensors for High Temperatures Based on Silicon Carbide," *Sensors Mater.* 4 (4), 173–185 (1993).
65. G. Hunter, P. Neudeck, C. Liang-Yu, D. Knight, C.C. Liu, and Q.H. Wu, "Silicon Carbide-based Detection of Hydrogen and Hydrocarbons," in *Silicon Carbide and Related Materials,* Vol. 142, edited by S. Nakashima, H. Matsunami, S. Yoshida, and H. Harima (IOP Publishing Ltd., Bristol, UK, 1996), pp. 817–820.
66. V.B. Shields, M.A. Ryan, R.M. Williams, M.G. Spencer, D.M. Collins, and D. Zhang, "A Variable Potential Porous Silicon Carbide Hydrocarbon Gas Sensor," in *Silicon Carbide and Related Materials,*

Vol. 142, edited by S. Nakashima, H. Matsunami, S. Yoshida, and H. Harima (IOP Publishing Ltd., Bristol, UK, 1996), pp. 1067–1070.

67. R. Okojie, A. Ned, and A. Kurtz, "Operation of Alpha (6H)-SiC Pressure Sensor at 500°C," in *Technical Digest - 1997 International Conference on Solid State Sensors and Actuators*, edited by K. Wise and S. Senturia (Chicago, Il, 16–19 June 1997), pp. 1407–1409.
68. R. Zeirmann, J. von Berg, W Reichert, E. Obermeier, M. Eickhoff, and G. Kroetz, "High Temperature Pressure Sensor with Beta-SiC Piezoresistors on SOI, Substrates," in *Technical Digest - 1997 International Conference on Solid State Sensors and Actuators*, edited by K. Wise and S. Senturia (Chicago, Il, 16–19 June 1997), pp. 1411–1414.
69. A. Singh, M. Mehregany, S. Phillips, R. Harvey, and M. Benjamin, "Micromachined Silicon Fuel Atomizers for Gas Turbine Engines," in *Proceedings of the 9th Annual International Workshop on Microelectromechanical Systems*, edited by M. Allen and M. Reed (San Diego, CA, 11–15 February 1996), pp. 473–478.
70. N. Rajan, C.A. Zorman, M. Mehregany, R. DeAnna, and R. Harvey, "3C-SiC Coating of Silicon Micromachined Atomizers," in *Proceedings of the 10th Annual International Workshop on Microelectromechanical Systems, edited* by S. Kazuo and S. Shoji (Nagoya, Japan, 26–30 January 1997), pp. 165–168.

5

Laser Processing for Microengineering Applications

J. Brannon,* J. Greer,† and H. Helvajian‡

5.1 Introduction

Laser material processing is a technique by which materials can be fashioned in a nonintrusive manner with overall precision approaching the wavelength of the laser light. This processing is accomplished by exploiting the unique optical properties of the light to selectively remove or deposit material in a controllable manner. Materials thus processed to date include metals, ceramics, polymers, and semiconductors.

Microengineering is a discipline dealing with the design, materials synthesis, micromachining, assembly, integration, and packaging of miniature two-dimensional (2D) and three-dimensional (3D) sensors, microelectronics, and microelectromechanical systems (MEMS).[1] The physical structures and components have nominal dimensions, from the nanoscale to the microscale, and up to the millimeter scale. Microengineering has received worldwide attention because it promises to enable "intelligent" microinstruments with wide-ranging applications to be used in medicine, transportation, communications, and housing. Implicit in microengineering technology is the need to process materials with high dimensional accuracy, to process selective areas of the materials without incurring collateral damage to adjoining areas, and to prototype designs quickly without resorting to large-scale foundry operations. Laser material processing is uniquely qualified for microengineering because it can process materials without adhering to surface and crystallographic planes and can create millimeter-to-micron-scale structures in a broad range of materials. In addition, laser processing offers a number of capabilities that are complementary to both traditional material processing and semiconductor processing approaches. For example, lasers can process a large variety of materials, they can operate over large areas (meters squared) while maintaining high precision (less than microns), they can fashion materials by either selective removal or deposition, and they can alter materials through nonequilibrium chemical processes. These capabilities have profound consequences for both aeronautical and space applications, where specially "engineered" materials are often required and microengineering components using these novel materials may be necessary.

It is predicted that microengineering concepts will play an important role in the development of future aerospace systems. This prediction has bearing because the concepts make intelligent use of available volume and mass, and because microengineered components inherently use little energy. In more advanced applications, microengineering technology will enable the incorporation of localized "intelligence" and will provide the capability for exercising local autonomous action. These capabilities should be of benefit to any aerospace system design problem if component reliability can be assured. Traditional aerospace dogma is to favor reliability over new innovations, primarily because of the limited access for repair and the need for operation in extreme

*IBM Almaden Research Center, San Jose, California.

†Epion Corporation, Bedford, Massachusetts.

‡Center for Microtechnology, The Aerospace Corporation, El Segundo, California.

environments. As currently envisioned, future aeronautical and space systems will include passenger transports traveling at hypersonic speeds near suborbital altitudes and space missions that will be administered by using compact, fully integrated spacecraft "packages." These packages will roam the cold recesses of the solar system (i.e., NASA's Pluto-Kuiper Express Mission) or the "dusty" tail of a comet (i.e., NASA's Stardust Mission), or will orbit in large-number constellations around Earth or other planets, serving as communication or observation outposts. Implementing these new missions will require the development of novel materials and systems integration approaches that are specifically "engineered" to withstand harsh environments.

The development of novel materials will necessitate the development of material processing tools that can fabricate these new materials and package them alongside electronics. The laser is one such processing tool with unique advantages for materials modification. First, the laser is a nonintrusive, *in-situ* processing tool that can simultaneously perform several tasks, including serving as its own process monitor. Second, the laser is easily amenable to automation and is commonly used in processing situations where site-specific action is necessary. Third, delicate operations can be done with lasers, including atomic layer-by-layer removal by etching and controlled ablation techniques, site-specific surface oxidation, semiconductor dopant deposition and dopant drive-in, surface annealing, embedded interface processing, and pulsed laser deposition (PLD) of single-unit crystal films.[2] In essence, lasers have the capability for establishing a nonequilibrium chemical environment for materials processing. As a consequence, novel materials and microstructures can be fashioned.

In this chapter, we explore the laser material processing applications for microengineering technology. We present the processing steps that might be required for developing miniaturized systems fashioned of numerous materials (e.g., semiconductors, insulators, ceramics, polymers, diamond, metals). We also detail the micromachining processes that might be implemented for developing microstructures for the following:

- Fluid delivery channels
- Resonant high-Q structures
- Surface corrugations and special topologies
  - To direct light
  - For acoustic waves
- Surface texturing for enhancing
  - Catalysis
  - Aerodynamic flow
  - Tribology
  - Thermal conductivity

In addition, we present the laser deposition schemes for enabling the growth of various thin-film devices (e.g., ferroelectric materials for radio frequency [RF] circulators, dielectric optical coatings, solid lubricant coatings, high-temperature superconductor ceramics). Finally, we discuss the generic use of lasers in postassembly processing:

- Embedded interface processing
- Direct-write processing (deposition and etching)
- Cutting/trimming

The selected processing techniques that are explained in this chapter use the laser wavelength and unique optical properties to advantage rather than as a mere heating source. Conventional laser processing techniques, such as those used for macroscale cutting and welding applications, are not covered in this chapter. Also not covered are the techniques in which the laser is used as a

light source, namely, in photochemical curing of plastics or as a general exposure "tool." However, the use of the laser to "write" 3D microstructure images with a volumetric-exposure technique is discussed.

The general scope of this chapter is to present a tutorial review of the pertinent fundamental theories on laser material-interaction physics and a few detailed examples of laser processing applications. For more information, the reader is directed to the numerous excellent reviews and books that cover this information in greater detail.[3] The intended audience is an interdisciplinary group composed of the aerospace engineering community; the traditional material processing community; the thin-films growth community; and the microengineering community, which performs research in (1) microelectronics processing, (2) MEMS, (3) microoptical electromechanical systems (MOEMS), and (4) advanced packaging.

Specifically, this chapter is organized into the following sections. Following the introduction, Section 2 presents certain attributes of laser processing. Section 3 covers the physical principles of laser processing. Section 4 presents a general overview of the important subsystems found in typical laser processing stations. Section 5 is a brief overview of the utility and limitations of laser processing. Section 6 describes four microengineering application examples where laser processing is used. Section 7 outlines a specific case study for the development of a laser processing tool. Section 8 concludes with a brief overview of future trends and applications specific to aerospace systems.

5.2 Laser Processing

5.2.1 Attributes of Laser Light

The usefulness of lasers in materials processing and microengineering has its basis in the attributes of laser light compared to the light from conventional radiation sources.[4] The directed-energy nature of laser light is perhaps the most important of these attributes, because energy can be delivered to a surface in a contactless mode. This action-at-a-distance attribute eliminates both mechanical interaction with the surface and the need for maintenance or replacement of worn parts. Another attribute is the low beam divergence of laser radiation–typically less than a few milliradians in the far field. A low beam divergence permits tight focusing, which in turn allows for high intensity and increased spatial resolution. Yet another key attribute is brightness, defined as the laser power per unit area emitted into a unit solid angle. The high brightness often associated with lasers is important in providing high-intensity radiation to a surface. Finally, an attribute that may or may not be a factor in laser processing, depending on the particular application, is laser coherence. There are two types of coherence, spatial and temporal. Spatial coherence is of greater importance than temporal coherence, because it is related to the spatial mode structure of the laser and hence to the beam-focusing properties. Temporal coherence is closely related to the monochromaticity of lasers. Lasers enable high average power to be delivered in a narrow wavelength range, which permits a great deal of selectivity for surface processing. Although single-frequency operation is obtainable in certain lasers, rarely is this degree of monochromaticity required. What is more important is the capability to deliver energy in specific wavelength regions, such as the deep ultraviolet (UV) or mid infrared (IR). The importance of this capability will become clearer in later sections that describe existing laser applications.

5.2.2 Laser Parameters of Importance for Microengineering

Table 5.1 lists several key laser parameters and their general impact upon laser material processing. When using laser processing for a particular application, the first consideration is what wavelength (and thus what type of laser) to use. The wavelength determines the amount and efficiency

of laser radiation coupling to the material's surface, because a material's absorptivity and reflectivity are wavelength dependent. For example, the processing of metals is far more efficient (in terms of photon utilization) with UV light rather than with IR radiation because of the significantly reduced reflectivity in the UV. In addition, the wavelength influences the focused spot size because of the direct relationship between the minimum spot size and wavelength.

Once the wavelength is determined, the key parameter to consider in most applications is the laser intensity. Intensity, in typical units of W/cm^2, refers to the amount of energy delivered to the surface per unit area and per unit time. By increasing or decreasing the laser intensity, various types of physical processes can be made to occur. These processes include melting, vaporization, ablation, deposition, etching, and for some atomically selective processes, multiphoton excitation. For example, by increasing the laser intensity during pulsed irradiation of a metal surface, simple melting can give way to droplet ejection and ablation. There are several laser parameters that influence the laser intensity, which are also listed in Table 5.1. For instance, for continuous-wave (cw) irradiation, the average power and spot size directly affect the laser intensity. The spot size also determines the local processing area and the spatial resolution as defined by the heat affected zone (HAZ). Precisely focused laser beams are capable of irradiating micron-sized areas. Submicron-sized areas can also be selectively processed by using interferometric lithography techniques.[5] For pulsed operation, both the energy and duration of a single pulse directly affect the peak intensity of the focused beam. Often, for fixed-pulse-length applications, the fluence, rather than the intensity, becomes the figure of merit. Fluence is the delivered energy per unit area per pulse, and the units are in J/cm^2. The delivered number of pulses per second, or the pulse repetition rate, clearly influences the processing rate and the speed of an application. For volume manufacturing, the pulse repetition rate is the one parameter that determines the throughput and time efficiency of a process. For many materials, pulsed irradiation of the surface results in a transient temperature rise. If the material's thermal diffusivity is small, then this temperature rise may be sufficient to induce melting and vaporization. For situations where the pulse repetition rate is high enough, this transient heating effect can be accompanied by a subsequent bulk-material heating. To the extent that this heat affects the processing application, bulk heating may ultimately limit the resolution to areas much larger than the focused spot size.

Table 5.1. Laser Parameters for Microengineering Applications

Parameter	Impact
Wavelength	Radiation coupling efficiency to surface; spatial resolution.
Average power	Influences intensity, background heating.
Pulse energy	Influences peak intensity.
Pulse duration	Influences peak intensity, thermal diffusion length.
Focused spot size	Influences intensity, processing area, spatial resolution.
Intensity	Influences type and extent of surface processing.
Fluence	Influences type and extent of processing for pulse applications.
Pulse repetition rate	Influences processing rate, direct current (dc) heating.
Polarization	Influences coupling of energy to surface; process anisotropy.

Finally, another key parameter that affects laser processing is the polarization vector of the laser irradiation. Polarization may influence the degree of radiation coupling to the surface, and thus the uniformity of a particular type of processing. As an example, the quality of etching deep, high-aspect-ratio features in ferrites is affected by the polarization-dependent reflectivity of the sidewalls.[6]

5.2.3 Lasers for Microengineering

Since the invention of the laser in the early 1960s, literally hundreds of different types of lasers have been developed. It is safe to say that most of these lasers have had no influence on the commercial and military sectors. For materials processing and microengineering applications, there exist a few gas, solid-state, and metal-vapor lasers that are turned to again and again because of their inherent capabilities. Liquid-state dye lasers, never important for materials work, are disappearing because of rapid advances in tunable solid-state lasers.

The important lasers for microengineering work are listed in Table 5.2. Among the gas lasers listed, the ubiquitous CO_2 laser is the industrial workhorse. This laser is capable of providing cheap and plentiful photons in either pulsed or cw modes. Limitations of the laser are that the IR wavelength cannot be focused to micron-sized spots and that the coupling efficiency to a solid surface is often poor. Nevertheless, for the processing of certain materials (i.e., glasses), this laser is an excellent choice. Other lasers useful for microengineering are the rare gas ion lasers, such as Ar^+ or Kr^+, which provide powerful cw operation in the visible and UV spectral regions. These lasers are often used for scribing and deposition work, and can provide small intense spots because of their shorter wavelengths and good beam quality. The drawbacks of these lasers are their expensive electrical requirements and limited plasma tube lifetimes. The rare gas halide excimer

Table 5.2. Lasers for Materials Processing and Microengineering

Type	Laser	λ	Significance and Use
Gas	CO_2	9–11 μm	Pulsed and cw operation; low cost of ownership.
	Ar^+	Many lines	cw operation; strong green output; UV lines.
	Excimer	0.35–0.19 μm	Highest pulsed energy output in UV; significantly more expensive than CO_2; has revolutionized UV materials processing.
Metal vapor	Copper	0.51 μm	Short pulse, high pulse repetition frequency visible output that can be frequency doubled; unknown reliability in manufacturing setting.
	He-Cd	0.44 μm 0.32 μm	Blue and UV cw output; limited ion tube lifetime
Solid state	Nd:YAG	1.06 μm	Popular workhorse laser; cw or pulse operation; harmonics can be efficiently generated to create UV and visible radiation.
	Ti:sapphire	0.80 μm	Pulsed, tunable laser that can be efficiently frequency doubled; generator of femtosecond pulses.
	Diode	0.63–1.0 μm	Electrically efficient; pulsed or cw operation; low power and limited wavelength selection.

lasers, on the other hand, have caused a small revolution in laser processing ever since their invention in the late 1970s. The excimer lasers remain the only commercial source of high average power, deep UV radiation currently available. Their high pulse energy permits efficient etching and ablation of surfaces over relatively large areas. The UV nature of the radiation, coupled with the poor spatial coherence of the lasers, allows for speckle-free high-resolution patterning. Indeed, micron and submicron patterning applications have been routinely reported.[7] Excimers are the laser of choice for PLD applications, for reasons that will be discussed later in this chapter.

High-repetition-rate (>10 KHz) copper vapor lasers, which long suffered because of the reliability issue, have made a comeback as viable lasers for the micromachining of thick metals. Precision holes for automotive fuel injection systems are now being cut by these lasers. The visible wavelength of the lasers (517 nm) limits their applications. Copper vapor lasers can be frequency doubled to the UV (~258 nm), but the conversion efficiency is poor because of the multimode nature of the laser. On the other hand, cw He-Cd lasers, operating at two wavelengths (325 and 447 nm), have not been used much in micromachining applications. The output power of these lasers is low even though the wavelength of operation is useful. These lasers have found use in exposure and soft printing applications.

Solid-state lasers for materials processing are led by the Nd:YAG laser. This laser can be operated either in cw or pulsed mode at fairly high average powers. Unlike the CO_2 laser, the Nd:YAG can be efficiently converted to its harmonic wavelengths of 0.532, 0.355, and 0.255 μm, particularly when it is operating in a pulsed Q-switched mode. This wavelength conversion feature permits more flexibility in applications. For high pulse repetition work, the Nd:YVO_4 laser can operate at up to 100 kHz with excellent pulse-to-pulse stability. Another laser that is a relative newcomer to the solid-state group is the Ti:sapphire laser. The great advantage of this laser is its tunability near the fundamental wavelength of 0.8 μm. The Ti:sapphire laser can also be efficiently frequency doubled to provide tunable near-UV output. In addition, mode locking of this laser creates femtosecond pulses (10^{-15} s). There is increased interest in these very short pulses as a new and powerful means for performing laser-based microengineering.[8] Finally, diode lasers, with their small size and high electrical efficiency, are starting to be used for some types of laser processing. Diode lasers can operate in either cw or can be directly modulated to gigahertz rates. Because these single-spatial-mode lasers have relatively low power and limited wavelength range (0.6–1.0 μm), they are not widely used. However, continued advancement in diode technology promises more powerful lasers with a greater choice of wavelength in the future.

It is also worthwhile to mention the advent of the powerful free electron lasers (FEL). There is interest in such lasers for materials processing because of their ability to provide high average power radiation that is tunable over a wide spectral range.[9] Additionally, the high spatial quality beam emits picosecond duration pulses at megahertz rates. Thus, very high throughput operations are possible, even though the amount of material processed on a per-pulse basis may be small. Currently, these lasers are still in the experimental stage, and no FEL has yet emerged that can be considered as a viable candidate for radiation processing of materials. Nevertheless, a U. S. government-sponsored FEL program at the Thomas Jefferson National Laboratory in Virginia is commissioning a kilowatt class (37-MHz repetition rate) IR FEL designed for laser processing studies. The results from this machine should go far to discern the credibility of these lasers for industrial use.

5.2.4 Laser Material Processing Tools

Lasers as material processing tools are typically used to alter a material by exposure, ablation, or etching (the latter, by the addition of chemical reagents). Also, laser tools can grow a material by

redeposition of the ablated material, by the addition of chemical precursors to induce chemical vapor deposition (CVD), by the sintering of nanophase powders/slurries, or by the curing of polymers. In general, laser tools either operate via a batch-processing scheme, where the laser beam passes through a mask and irradiates a selected large area, or via a direct-write serial-processing scheme, where the laser beam is focused onto a specific small area. The former approach typically offers high throughput, while the latter approach offers site-specific processing control. In complex tools, a combination of masking and direct-write focusing is implemented to provide arrays of focal point sources for parallel processing. As a general rule, laser tools that employ masks are mostly designed for production, while those that implement direct-write action are used for rapid prototyping. This generality does not hold for laser welding or cutting applications, which are primarily direct-write tools. These applications are not discussed in this chapter but are deemed mature technologies found in many manufacturing industries.

For specific microengineering applications, both batch-processing and direct-write approaches are used, with the direct-write approaches offering more flexibility. Direct-write laser processing instruments typically include a laser beam delivery system (BDS) with a focusing microscope objective, a computer-driven XYZ stepper and/or optical scanners, and a surface imaging system to monitor progress. Additional modifications to the instrument may include the capability to process at multiple wavelengths; to measure the surface topography with a white-light or laser interferometer; and to continuously dose the surface with gas for etching, deposition, or debris removal. By employing the appropriate laser, a direct-write processing instrument can be used for etching materials, ablating materials, annealing materials, or depositing materials and dopants, all with site-specific control. There are two direct methods for removing material: chemical etching[10] and ablation.[11] The indirect method is by photoexposure.[12] Similarly, there are at least two techniques for depositing material: laser CVD[13] and PLD.[14] The PLD technique has at least two variants. One variant is known as MALDI (matrix-assisted laser desorption/ionization), and the other is called MAPLE (matrix-assisted pulsed laser excitation). These relatively new variants are primarily used to deposit intact large organic or biologically significant molecules. Direct-write laser techniques can be used for the micromachining of ceramics,[15] glasses,[16] and diamonds,[17] and for the deposition of polysilicon and semiconductor dopants. It is also possible to "drive in" the dopant[18] via laser irradiation. The ultimate resolution for the direct-write technique is normally subject to the limitations of diffraction. However, it is feasible to fabricate patterned lines that are less than the diffraction-limited spot size. For this fabrication, the coherent properties of the laser are exploited, and surface interference effects are taken to advantage. It is then possible to fabricate patterned lines in the ~ 0.1 μm range.[19] In general, and especially for micromachining applications, spot size and depth of focus are the critical parameters. As a consequence, material processing with short depth of focus requires a precise knowledge of the objective lens-to-surface distance. If the surface topology is corrugated, a servoloop connected with an interferometric autoranging device must be used.

5.3 Physical Principles of Laser Processing

5.3.1 Beam Propagation, Energy Delivery on Target, and Coherence

Under most laser processing conditions, the criteria for optimum interaction between the laser and the material are indicated by a handful of equations. These equations describe the propagation of the laser beam energy through the beam delivery optics, the photophysical interaction of the laser beam with the surface (i.e., absorption and surface chemical interactions), and the subsequent surface modification as a result of electronic and thermal excitation. The equations are simplified for a Gaussian laser beam propagating in a diffraction-limited optical system, though for most

material processing, a top-hat or flat-top homogenized beam is used.[20] A Gaussian beam can be described by the radius function $\omega(z)$ and the wavefront curvature function $R(z)$ along the propagation direction z. The functions $\omega(z)$ and $R(z)$ are given by Eqs. (5.1)–(5.3)[21]

$$\omega^2(z) = \omega_o^2\left[1 + \left(2\frac{z}{b}\right)^2\right] \tag{5.1}$$

$$R(z) = z\left[1 + \left(\frac{b}{2z}\right)^2\right] \tag{5.2}$$

$$b = \frac{2\pi\omega_o^2}{\lambda} \tag{5.3}$$

where λ is the wavelength, ω_0 is the beam radius at the waist, and b is the confocal parameter (i.e., distance within which the diameter of a focused beam remains nearly constant, $-b/2 < z < +b/2$). The Gaussian beam contracts to a minimum diameter $2\omega_0$ at the beam waist, where the phase front is a plane wave.

Consider a collimated and strongly focusing Gaussian beam with the criteria that 99% of the energy is to be transmitted by the focusing lens of focal length f through an aperture D, and 86% of the energy is to be contained within a diameter $d_o = 2\omega_0$. The spot size d_o for this beam is given by Eq. (5.4).[22] Equation (5.4) can be recast using the more familiar $f^{\#}$ (where $f^{\#}$ is called the f-number and is defined as (f/D). Given a fixed focal length lens, a smaller d_o means shorter wavelengths and larger Gaussian beam diameters at the lens.

$$d_o \sim \frac{2f\lambda}{D} \rightarrow 2f^{\#}\lambda \tag{5.4}$$

Lenses with $f^{\#} > 2$ are relatively inexpensive, while lenses with $f^{\#} < 1$ are commonly multielement designs and very expensive. The depth of focus for this Gaussian beam is given by the confocal parameter, b, given in Eq. (5.3), but can also be approximated as $\sim 2\pi(f^{\#})^2\lambda$. With the use of a 248-nm wavelength laser beam and a focusing lens of $f^{\#} = 2$ (i.e., $f/2$ optics), the result is a minimum spot diameter of ~1 μm and a depth of focus of ~ 6 μm for processing in air. If the medium in contact with the surface is in a higher index material, the minimum spot size is further reduced by the index n. In the example given above, 86% of the incident energy is focused onto the minimum diameter spot (1 μm), which leaves 14% of the energy distributed over a diameter of ~2.3 μm. The 14% energy "spillover" may be damaging for some micromachining applications. To guarantee greater than 98% energy confinement within the design spot size, Eq. (5.4) should be multiplied by 2.3. There are practical reasons why a very small $f^{\#}$ may not be desirable for some micromachining applications. With decreasing $f^{\#}$, the minimum spot size declines by the same factor, as opposed to the depth of focus, which declines by the larger factor, $\pi(f^{\#})^2$. Material processing with a very short depth of focus requires a very flat surface, or a servoloop connected to an interferometric autoranging device to maintain the focus as the sample is moved.

As discussed previously, spot size and depth of focus are the two critical parameters for a micromachining application. In an imaging application (e.g., image projection, photolithography), these parameters are the resolution and the depth of field. Equation (5.5) defines the resolution, R, of a diffraction-limited imaging system,[23]

$$R = \frac{C\lambda}{N.A.} \tag{5.5}$$

where C is a constant (derived from the first Fraunhofer diffraction minimum for a circular aperture) and is related to the coherence of the illumination source. The value of C goes from 0.61 (incoherent illumination) to 0.77 (coherent illumination). The *N.A.* is the numerical aperture $\equiv n\sin\theta_{max}$, where n is the index of refraction and θ_{max} is 1/2 the maximum acceptance cone angle, which specifies the "light-gathering power" of an optical system. A low *N.A.* has a value of ~0.25, while a high *N.A.* is ~ 0.5 (e.g., Schwarzschild lens[24]).The depth of field for such an imaging system is given by Eq. (5.6).

$$Z = \frac{\lambda}{(N.A.)^2} \tag{5.6}$$

Again, the *N.A.* emphasizes the inverse relationship between the need for high resolution and a practical depth of field. Equations (5.4)–(5.6) show that the shorter the wavelength, the better the resolution and the smaller the minimum spot size achievable, which argues for using UV over IR light. However, materials show increasing dispersion in the index n at the shorter wavelengths ($dn/d\lambda$), especially in the deep UV. For transmissive optics, this dispersion results in chromatic aberration in the focusing/imaging, which must be taken into account for UV laser sources with large spectral bandwidths. Current UV excimer lasers without injection-locking wavelength stabilization have bandwidths that make less than half-micron processing difficult. The defocusing error df as a function of the source bandwidth is given in Eq. (5.7).[25]

$$df = \left(\frac{dn}{d\lambda}\right)\Delta\lambda\frac{f}{(1-n)} \tag{5.7}$$

Table 5.3 presents the measured dispersion for fused silica for the excimer laser wavelengths, and the maximum allowable source spectral linewidth, $\Delta\lambda_{max}$, for achieving a defocusing error of less than 1 μm, given a $f = 1$ cm lens.[25] Current excimer laser linewidths are typically 0.8 nm for wavelengths at 248 nm. Injection-locking schemes are used to narrow the emission linewidths, $\Delta\lambda$, to < 0.003 nm and thereby also to reduce the defocusing errors.

Table 5.3. Dispersion Values for Fused Silica, $dn/d\lambda$, for Achieving a Defocusing Error of Less than 1 μm, Given a $f = 1$ cm Lens, $\Delta\lambda_{max}$

	193 nm	248 nm	308 nm	351 nm
$dn/d\lambda$ ($\times 10^{-4}$ nm^{-1})	8.9	6	3.2	1.8
$\Delta\lambda_{max}$ ($\times 10^{-2}$ nm)	6	8	16	29

Fundamentally, a reduction in the laser linewidth means an increase in the coherence properties of the laser. In essence, it means maintaining the distinct time and phase relationship of the emitted photon wavetrains. In general, the coherence properties of a laser are delineated as arising from either the spatial or temporal domain. In spatial coherence, for any extended optical source, optical wavetrains can originate from spatially separated points. For lasers, the phenomenon is related to the number of laser transverse cavity modes that can extract energy from the gain curve. Reducing the number of operating cavity modes increases the spatial coherence. In temporal coherence, the coherent emission of light from atoms has finite duration, $\Delta\tau$, and consequently a finite frequency distribution spread, $\Delta\nu$ (i.e., finite bandwidth as a result of these truncated sinusoidal wavetrains). For pulsed lasers, the phenomenon is related to the pulse width of the laser (i.e., the duration of the population inversion); the atomic energy decay modes; and the average

number of times the laser energy is allowed to recirculate prior to exiting (i.e., cavity round-trip times). Of the two coherence properties, the spatial coherence influences the imaging and focusing capabilities of the laser beam, while the temporal coherence could induce novel photochemical or photophysical processes on surfaces or in bulk media. In practicality, the spatial coherence is more important than temporal coherence, and most precision processing lasers operate only in the fundamental transverse electromagnetic mode TEM_{00} (i.e., lowest order self-reproducing hermite-Gaussian cavity mode). Maintaining the spatial coherence of a laser also requires that the paraxial optical BDS be specially designed to match the confocal parameter of the laser with that of the cascading transfer optics.

In general, most pulsed lasers are best described as partially coherent sources, while cw lasers can be designed as nearly perfect coherent sources. The temporal coherence property of pulsed lasers is typically given by the coherence length, l_c. The l_c defines a distance, usually measured from the exit window of the laser, where the laser electromagnetic field amplitude and phase front changes in a consistently predictable way. In effect, the laser is considered a temporally coherent source for interactions at a distance less than l_c. Using the time-bandwidth product theorem, or Heisenberg uncertainty principle (which states that $\Delta\nu\ \Delta\tau \sim 1$, where $\Delta\tau$ is the laser pulse width and $\Delta\nu$ is the spectral bandwidth), the corresponding coherence length l_c is given by Eqs. (5.8) and (5.9). In these equations, the c is the speed of light, and λ, ν are wavelength and frequency, respectively.

$$l_c = c\Delta\nu \approx \frac{c}{\Delta\nu} \tag{5.8}$$

given that $c = \lambda\nu$ and $\Delta\nu = (c/\lambda^2)\Delta\lambda$

$$l_c \approx \lambda\left(\frac{\nu}{\Delta\nu}\right) \text{ or } \approx \frac{\lambda^2}{\Delta\lambda} \tag{5.9}$$

Current state-of-the-art solid-state lasers with injection-seeded oscillators have typical spectral linewidths of 50 to 90 MHz. For example, a Nd:Yag laser operating at a laser wavelength of 1064 nm with a spectral linewidth of 90 MHz gives a coherence length of 3.3 m. Novel nonthermal processing of materials would be possible within this coherence length. Because most of these lasers have a pulse width of 10 ns, which corresponds to an optical length of ~ 3.3 m, laser photons from the leading edge of the pulse would be coherent in relation to laser photons at the trailing edge of the pulse. Examples of nonthermal processing include the coherent "driving" of energy into excitonic particles and surface electrons.

5.3.2 Processing Speed and Process Window

There are cases in laser material processing where the laser repetition rate or the stepper/scanner speed is not the limiting factor. In such cases, the processing throughput is determined by the fundamental process speed, Ω_{sp} (μm/s). The Ω_{sp} depends on many factors, but is intrinsically dependent on the fundamental photophysical interaction (e.g., electronic, thermal, plasma); the character of the surface under irradiation; and the properties of the incident laser light. The photophysical interaction is a function of the laser fluence (J/cm^2) or the intensity (W/cm^2). In general, at very low fluences, the photophysical process is primarily induced by electronic excitations; at intermediate fluences, by thermal processes; and at high fluences, by the above-surface laser-initiated plasma. Similarly, the morphology of the surface changes with increasing laser fluence. In general, at low fluences, there is surface and near-surface defect formation; at intermediate fluences, there is surface melting and rapid recrystallization, resulting in amorphization; and at high fluences, there is plasma sputtering, spallation, and shock-induced damage. For all laser

fluences, knowledge of prior irradiation dose is critical to predicting additional surface changes. Therefore, for controlled processing, the photophysical interaction must be maintained within the domain of interest. Table 5.4 displays pertinent photophysical processes and the factors that impact laser processing. The table illustrates the many factors that characterize a process window and ultimately the processing speed, Ω_{sp}. A process window is characterized by measurement of the phenomenological processing rate, Γ (e.g., μm/s of material etched, ablated, or deposited). A simple model can be derived to relate Ω_{sp} to Γ. Assume the laser spot size diameter on the workpiece is D(μm) and the required processing thickness is ℓ(μm) (i.e., material to be etched, ablated, or deposited). Then ℓ/Γ gives the time to process one spot size. The stepper/scanner must move a distance D per ℓ/Γ unit time. The stepper/scanner speed or material processing speed is then given by Eq. (5.10):

$$\Omega_{sp} = \left(\frac{D}{\ell}\right)\Gamma \tag{5.10}$$

Lasers can be used for etching, annealing, or forming "luminescent silicon."[26] The projected processing speeds depend on the laser parameters and the experimental conditions employed, as follows:

- Type of laser (cw or pulsed)
- Laser wavelength
- Processing technique or chemistry employed (e.g., chlorine-etch or ablation)
- Laser polarization vector with respect to the scan direction (i.e., E ||: parallel or E⊥ perpendicular)

For example, a cw laser (Ar^+ ion) operating at 514 nm with 1.5 W output power, with a chlorine base etch chemistry and polarization set to E||, can cut 5-μm-deep trenches in silicon with a

Table 5.4. Photophysical Processes Used in "Direct-Write" Laser Processing, and Critical Factors

Photophysical Processes	Critical Factors
Chemical (deposition/etch)	Reaction initiator (gas phase absorption or substrate absorption) Optical absorption coefficient Heterogeneous reaction rate at gas-solid interface Diffusion of reactants/products (mass transport) Substrate thermal conductivity Nucleation rate Point and line defect densities
Ablation	Laser fluence and intensity Irradiation dose Surface morphology Bulk defect density Thermal conductivity (thermal diffusion length)
Laser-induced desorption	Wavelength Optical absorption coefficient Adsorbate binding energy Fluence and intensity Point defect density Metals (excited plasmon density)

process speed, Ω_{sp}, of several millimeters per second.[27] For the opposite polarization, the etching rate is a factor of 2 slower and results in nonuniform etching of the sidewalls. For pulsed laser ablation (excimer laser, 248 nm, 1.3 J/cm^2, 5 Hz repetition rate), a Ω_{sp} of 0.6 mm/s has been achieved with hole depths of 150 μm.[28] In comparing the laser-assisted chlorine-etch technique with the ablation technique, the former produces a smoother lined wall,[29] but the latter has a wider dynamic range of trenching depth.

The Γ for numerous materials and laser processes (e.g., semiconductors, insulators, and metals) can be found in the literature.[30] Because the process conditions influence Γ, one expects different Γ for different laser irradiation conditions (see Table 5.1). Gas dynamics also influence Γ. For example, for process windows where Γ is limited by the diffusion of reactants or products into and out of the processing zone, using smaller spot sizes can lead to an increase in Γ by factors approaching 10^4. This increase is primarily a result of diffusion geometrics. As the spot size is *reduced,* there is a transition in the reactant, and product diffusion from a 3D expansion process to a one-dimensional (1D) process. A consequence of the reduced dimensionality is that the reaction fluxes in the active zone increase, resulting in an effective larger Γ. For laser-assisted chemical processing, the transition from 1D to 3D expansion appears for spot sizes near 80 μm.[31]

Diffusion-limited processing can be deleterious to the fabrication of high-aspect-ratio structures (hole-depth/hole-width » 1), because mass transport to and from the active zone is limited. The derived simple processing speed model, as shown in Eq. (5.10), is not valid and must include parameters for diffusion. Equation (5.11)[32] relates Ω_{sp} to Γ for processing in the diffusion-limited regime. The L is a constant related to the diffusion properties of the product. Equation (5.11) reduces to Eq. (5.10) for small l/L ratios.

$$\Omega_{sp} = \frac{D\Gamma}{L}\left[\exp\left(\frac{l}{L}\right) - 1\right]^{-1} \tag{5.11}$$

Processing speed, Ω_{sp}, is also influenced by light scattering out of the active zone or waveguiding into and out of an active zone. For example, if the material is processed via a cw or long pulse laser and the products include cluster particles, then the intensity at the processing interface could be reduced as a result of the light shadowing or scattering from the ejecta. The Mie scattering theory applied to large opaque particles (with diameters $\approx\lambda$) shows that the loss of laser energy is proportional to twice the particle diameter. Another effect in the laser fabrication of high-aspect-ratio structures like deep (>50 μm) via holes is that of waveguiding[33] by total internal reflection. Figure 5.1 shows the condition and gives the threshold condition based on Brewster's angle. Note that the in-plane (p) polarized light is not reflected but is absorbed at the wall, resulting in sidewall nonuniformity. For ℓ/D ratios resulting in $\varphi > \varphi_{Br}$, total internal reflections within the trench should be expected. Under these circumstances, great care must be exercised in aligning the laser polarization angle.[34]

5.3.3 Optical Absorption, Thermophysics, and Laser-Induced Plasmas

The fundamental laser–material interaction is nominally governed by the optical absorption of the material. Leaving out certain specifics, this absorption can have bandwidths (1) with near-molecular origin (<cm^{-1}), as for certain adsorbates; (2) more akin to the band structure of solids (>>cm^{-1}), as for periodic lattice crystals; or (3) near continuum absorption, as for a free-electron gas metal or a plasma. In the first and second cases, spectroscopic measurements define the absorption properties, while in the third case, specific models and the material dielectric properties can be used to get general absorption behavior. For example, the optical properties of metals are derived from the dielectric constant. At laser wavelengths where a metal behaves like an ideal free-electron metal, its dielectric properties can be approximated by the Drude model.[35] The

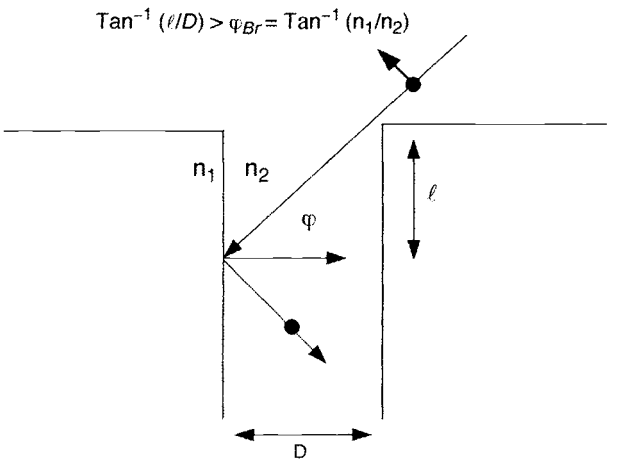


Fig. 5.1. Impact of Brewster's angle on laser "via hole" drilling. Total internal reflections occur for $\varphi > \varphi_{Br}$.

dielectric constant ε is temperature dependent and can be related to the angle-dependent reflectivity, R. Equations (5.12) and (5.13) describe the change in reflectivity for the transverse electric ([*TE*]–vector perpendicular to plane of incidence, or typically identified as s-polarized) and the transverse magnetic ([*TM*]–electric vector parallel to plane of incidence, or typically identified as p-polarized) polarized waves as a function of incident angle θ. The angle θ as defined is relative to the surface normal. At normal incidence ($\theta = 0$), the *TM* and *TE* reflectivities degenerate into one equation which, for completeness, is given in Eq. (5.14). The n and k are the real and imaginary parts of the complex index of refraction. They are related to the dielectric constant by the equation $\varepsilon^{1/2} = n + ik$. The extinction coefficient, k, is related to the optical absorption coefficient $\alpha(\text{cm}^{-1})$, as given by Eq. (5.15). Equation (5.15) also defines the optical absorption coefficient and the resulting attenuation of the laser intensity (I) over a distance (z).

$$R_{TM} = \frac{\cos(\theta) - \cos(\theta)/\varepsilon^{1/2}}{\cos(\theta) + \cos(\theta)/\varepsilon^{1/2}} \tag{5.12}$$

$$R_{TE} = \frac{\cos(\theta) - \varepsilon^{1/2}\cos(\theta)}{\cos(\theta) + \varepsilon^{1/2}\cos(\theta)} \tag{5.13}$$

$$R_{(TM)/(TE)} = \frac{(n-1)^2 + (k)^2}{(n+1)^2 + (k)^2} \tag{5.14}$$

$$\alpha = \frac{4\pi k}{\lambda} \tag{5.15}$$

where $dI(z)/dz = -\alpha I$.

Regardless of the initial absorption process, the absorbed energy quickly spreads via numerous decay channels and results in bulk heating. However, there are cases where the deposition of energy in a specific absorption feature induces a specific action without the consequences of heat. These nonthermal phenomena are generally observed in femtosecond laser pulse experiments[36] or in low-fluence laser material interaction experiments.[37] In both sets of experiments, the material is "processed" on the atomic scale with processing yields so low that, as a processing technique, is of minimal use for practical applications. However, with ever-increasing laser repetition

rates (kHz → MHz), low-yield, species-selective processes can be viable for certain applications where atomic level of control is necessary. Nevertheless, for most laser material processing, the consequences of bulk sample heating by the laser must be considered because this heating can influence the processing resolution or the fabrication throughput. The time scale and the nature of the heat flow characterizes the type of processing. For long processing times (i.e., which result with the use of cw lasers), the temperature distribution is in steady state and the heat flow problem is only tractable via a 3D solution. A practical consequence of operating in this heating domain is that the effect of high temperature on adjacent features needs to be taken into account. In contrast, for short processing times (i.e., which result with the use of short-pulse lasers), the heat flow is primarily a 1D problem, where the temperature gradient is into the bulk and normal to the surface. Under these conditions, the effect of high temperature on the surrounding area can be ignored. Because lasers can induce a wide range of heating rates–up to as high as 10^{15} K/s (femtosecond pulse excitation)–the processing thermophysics is governed by the thermal properties of the irradiated material. These properties include the thermal conductivity [(κ; W/(cm-K)], the heat capacity [c_p; J/(cm^3-K)], and the temperature-dependent optical properties. Equation (5.16) defines the thermal diffusion length (in cm), where τ is the processing duration (i.e., the laser pulse duration). The ratio (κ/c_p) is called the thermal diffusivity, D_T (cm^2/s), and can be used to calculate the time for reaching a steady-state temperature within a processing zone of size ψ. This time is given by $T \approx \psi^2/(4D_T)$.

$$\chi = \left(\frac{4\kappa\tau}{c_p}\right)^{1/2} \tag{5.16}$$

Using the metals as an example, the optical absorption depth ($1/\alpha$) for most metals, in the visible and the near-IR regions, is only a few hundred angstroms. On the other hand, for nanosecond pulsed lasers, the thermal diffusion length, χ, is on the order of 1 μm. So for most *pulsed* laser processing of metals, the optical absorption depth is much shorter than the thermal diffusion length ($1/\alpha < \chi$). Under these circumstances, the temperature at the material surface can be calculated if the laser pulse shape is known.[38] However, for the purpose of material processing, the key issue is whether the thermal diffusion length, χ, is greater or less than the feature size, ψ, to be processed; likewise, if the processing time, ψ/Γ, for the feature is greater or less than $T \approx \psi^2/(4D_T)$. If the *processing time is greater* than T, then the solution requires a 3D analysis in which the laser intensity radial distribution must be identified. The 3D analysis is complicated, but for a circular aperture and a cw irradiation zone, the affected area is a hemisphere of diameter $(\pi\psi)^{1/2}$. On the contrary, if the *processing time is less* than T, then the heat diffusion is a 1D problem, and the maximum temperature rise at the surface, ΔT_{max}, can be approximated by Eq. (5.17).[39]

$$\Delta T_{max} = \frac{F(1 - R_{sol})}{c_p \chi} \tag{5.17}$$

where F is the laser fluence (J/cm^2) and R_{sol} is the surface optical reflectance. The numerator on the right-hand side of the equation describes the laser fluence absorbed. In most cases, the thermochemistry is governed by the temperature fall time, which is given by $\Delta T_{fall} \sim \chi^2/(4D_T)$.[39] For weak absorbing materials where $1/\alpha > \chi$, $1/\alpha$ replaces χ in Eq. (5.17) and in the equation for ΔT_{fall}. In the particular case of weak absorbers (i.e., wide bandgap insulators or semiconductors) or thin films, the laser material interaction may create defects[40] in the material or may infuse stress or strain in both the irradiated and surrounding area.[41] To analyze the stress and strain distribution, the material thermoelastic equations must be solved to quantify the effect of the laser heating.[42] An understanding of the residual stress is important for implementing the laser

annealing technique[43] or any laser direct-write processing technique. The annealing irradiation dose and the scan speed have an effect on the residual stress distribution. A highly stressed material commonly quenches by atom dislocation and microcracking, while a low stressed material shows a shift in the phonon spectrum. In both cases, the material is ridden with defects, which can be used to advantage to induce particle emission from the surface via a nonthermal laser excitation scheme.[44,45]

The description of the laser material interaction, as given by the optical absorption and subsequent thermal processes, is valid as long as the photoejected species density is small. With increasing laser fluence, more material evaporates and the likelihood for photoionization and thermionic emission[46] increases. The Richardson-Smith equation estimates the thermionic ion emission current as a function of temperature.

$$J_+ = A_p T^2 \exp[-(I_p + \varphi_o - U_{ce})]/kT \tag{5.18}$$

In Eq. (5.18), A_p is a constant, T is the local temperature, I_p is the ionization potential (eV/atom), φ_0 is the electron work function (eV), and U_{ce} is the cohesive energy (eV/atom). With further increase in the laser fluence, both the photoionized and the photoemitted electrons absorb energy from the laser beam via the inverse Bremsstrahlung process. The absorption process is described as a three-body interaction with nearby ions and raises the electron to a higher electronic state. The higher kinetic energy electron ionizes additional atoms via electron impact excitation. The resulting effect is an avalanche of ionization with less light actually delivered to the target and more light into the protoplasma. The absorption coefficient for the Bremsstrahlung process can be calculated and is given in Eq. (5.19) in cgs units.[38]

$$K_\nu = \left(\frac{4}{3}\right)\left(\frac{2\pi}{3kT}\right)^{1/2} \frac{n_e n_i Z^2 e^6}{hcm^{3/2}\nu^3}\left[1 - \exp\frac{-h\nu}{kT}\right] = 3.69 \times 10^8 \left(\frac{Z^3 n_i^2}{T^{1/2}\nu^3}\right)\left[1 - \exp\left(\frac{-h\nu}{kT}\right)\right] \tag{5.19}$$

where n_i and n_e are, respectively, the ion and electron densities in a plasma of average charge Z and temperature T. The c, e, m, h, and k are, respectively, the velocity of light, the electronic charge, the electron mass, Planck's constant, and Boltzmann's constant; ν is the frequency of light that is related to the wavelength by the equation $c/(N\lambda)$ (where N is the plasma optical index). The term $1/K_\nu$ defines the light absorption pathlength (cm) into the plasma, while the term $[1-\exp(-h\nu/kT)]$ accounts for losses by stimulated emission. For specific conditions, Eq. (5.19) can be approximated. For $h\nu >> kT$ (e.g., UV wavelength laser), the $K_\nu \sim (T^{1/2}/\nu^3)^{-1}$ while for $h\nu << kT$ (e.g., high-temperature plasma), the absorption coefficient is approximated by $K_\nu \sim (T^{3/2}\nu^2)^{-1}$. All other parameters being equal, in both extreme cases, the shorter wavelength laser is preferable because it results in a smaller K_ν. If the laser fluence is such that an above-surface plasma does form, then only optical frequencies higher than the plasma frequency, $\nu_p = 8.9 \times 10^3 n_e^{1/2}$, can penetrate the plasma. Conversely, given a laser with frequency ν, the laser can penetrate the plasma for electron densities $n_e < (\nu/8.9 \times 10^3)^2$.

The plasma temperature T, which appears in Eq. (5.19), is difficult to measure for a plasma that is not in local thermodynamic equilibrium. However, where thermodynamic equilibrium can be assumed (e.g., for long pulse width lasers or for laser-induced plasma densities), the temperature can be determined by spectroscopic measurement of the emission intensities and the coupled Saha equations.[38] Regardless, a laser-induced plasma absorbs power from the laser. The absorbed power is reradiated primarily via the Bremsstrahlung or lost via the plasma thermal conductivity. For processing on the micrometer scale, the result is a reduction in resolution. On the other hand, for processing on the macroscopic scale, the plasma can be "tailored" to deliver maximum energy

transfer to the material surface.[47] The reradiated power per volume (W/cm^3) is given in Eq. (5.20), and the thermal conductivity (W/cm^{-1} K^{-1}) is given in Eq. (5.21).[38] For completeness, Eq. (5.22) shows the time for equilibrating the electron and ion temperatures. The term ($ln\ \Lambda$) is a function of plasma parameters[48] and is of the order of 10, and A is the ion atomic weight in amu.

$$P = 1.42 \times 10^{-34} Z^3 n_i^2 T^{1/2} \tag{5.20}$$

$$\kappa = \frac{1.95 \times 10^{-11} T^{5/2}}{Z \ln \Lambda} \tag{5.21}$$

$$\tau_{eq} = \frac{252 A T^{3/2}}{n_e Z^2 \ln \Lambda} \tag{5.22}$$

Assume a laser irradiates an aluminum surface with spot diameter of 1 μm. Assume also that the fluence is such that a plasma temperature of 2×10^4 K (~1.7 eV) is established with $n_i = n_e \sim 10^{17}$ cm^{-3}. (Also assume it is a weak plasma roughly 0.001% of solid-state density, or ~1% vapor density at 1 atm.) For a 10 ns pulse laser, the plasma thickness is ~2.7×10^{-3} cm (~2.7×10^{-7} cm for a 1 ps laser), which results in a volume of ~2×10^{-11} cm^3 and a total reradiated power ~1 μW. The thermal conductivity of this plasma becomes ~ 10^{-2} Wcm^{-1}K^{-1}, which is similar to the thermal conductivity of an insulator (e.g., titanium dioxide [TiO_2]). The time for establishing temperature equilibrium is $\tau_{eq} \sim 0.1$ ns. Now consider increasing the ion and electron densities by a factor of 100. The power reradiated from the small volume increases by 10^4 (10 mW), while the time for reaching equilibrium decreases by 100 (1 ps).

5.4 Supporting Systems in Laser Processing

5.4.1 Beam Delivery System

There is a great deal of information on laser BDS. This information is covered to some extent in most papers dealing with laser processing.[50] In its essence, a BDS manipulates the laser beam such that the beam is delivered to its intended target with the desired spatial, temporal, and intensity characteristics. The components of a BDS include the familiar mirrors, lenses, attenuators, beam splitters, shutters, and polarization elements. Often, gimballing, translation, rotation, and angular adjustments of optical elements are required for the BDS; therefore, the instrumentation to perform these functions is usually included on the list of components.

Prior to establishing a BDS, the user must have thorough knowledge of the laser system and its purpose. Specifically, the usual laser parameters must be known (wavelength, beam diameter or size, beam divergence, pulsed or cw operation), along with the intended spot or image size and the required surface-incident intensity. This knowledge can then be used to establish the number and type of optical components needed to accomplish the task. Correct placement and use of the optical components of course requires a working knowledge of optics. Possible damage to the optic or its coating from the laser must also be assessed, particularly for short pulse or deep UV systems.

Two types of BDSs will be discussed: image projection and focused scanning. Although the BDS comes in many forms, these two represent generic types that are commonly used. Figure 5.2 shows an image projection system, while Figure 5.3 is a focused scanning system. Note that in the focused scanning system, it is the target, rather than the laser beam, which moves.

Image projection systems are often used with excimer laser processing to take advantage of that laser's high-resolution capabilities, and because the poor mode quality of most excimer lasers does not favor small spot focusing. Indeed, the multimode operation of the excimer laser lends

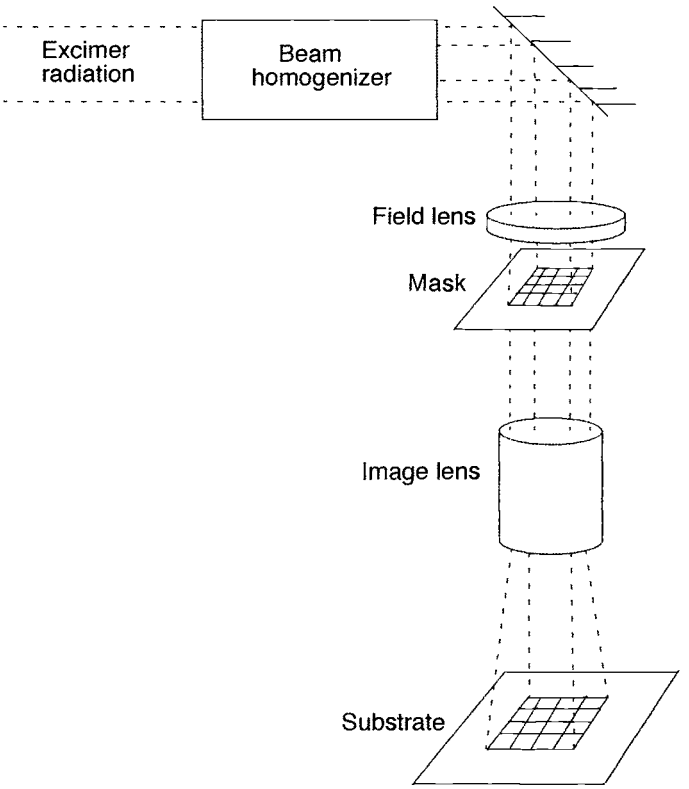


Fig. 5.2. Optical schematic of via ablation tool.

itself nicely to speckle-free imaging, which is one reason these lasers are in such demand for ultralarge silicon integration (ULSI) photolithography. The standard optical configuration for image projection is a Kohler illumination system.[51] As depicted in Fig. 5.2, the laser beam is often shaped before it enters a beam homogenizer. This ensures efficient photon utilization so that all the energy is captured by the homogenizer. In Fig. 5.2, a pair of anamorphic lenses are used in a telescope configuration to produce a square beam from an initially rectangular one. Because of the poor mode profile, most excimer beams have insufficient uniformity for projection processing. Beam homogenizers are then used to improve the spatial uniformity to the desired level.[52] Uniformity variation of less than 5% is often acceptable. Homogenizers function by producing a uniform plane of light just at their output. As the beam propagates, the spatial divergence significantly reduces the uniformity. Therefore, a field lens is used to image the uniform output plane from the homogenizer onto the aperture or mask. Note that the magnification of this lens does not have to be unity, but may either enlarge or diminish the beam size at the mask relative to the exit plane of the aperture. Once the mask is uniformly illuminated, a high-quality, low-aberration transfer lens is used to image the mask onto the intended target. The transfer lens typically possesses a magnification of unity or larger. Larger magnifications (2×, 5×, 10×) are often used in order to reduce the fabrication complexity (and thus cost) of the mask. As with all projection systems, the numerical aperture of each optical component must be matched with each of its neighbors in the optical pathway so that all energy is collected and no beam "spillover" occurs. A complete BDS can automatically locate the target in three dimensions (spatial positioning and focus).

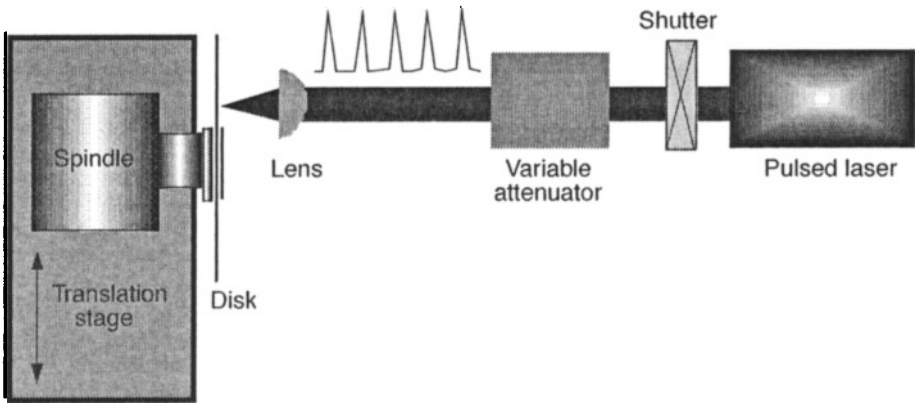


Fig. 5.3. Representative schematic of a focused laser direct-write processing system.

The BDS also includes display systems that permit real-time viewing of the part being processed. Such a camera system is often coupled with a high-quality microscope, so that micron-to-submicron patterns can be examined. The entire optical system (mask, projection lens, and target surface) must be rigidly mounted to prevent relative motion between these parts. The prevention of relative motion becomes all the more critical as the image size decreases.

For systems that possess good single-mode quality, the focused scanning BDS is a possibility. This type of BDS is possible because single-mode lasing permits a tight, well-defined focal region to interact with the surface of interest. Focused scanning systems are used in applications requiring extremely rapid beam movement or in the processing of very fine resolution patterns. Beam movement over a predetermined area is achieved by the mechanical motion of the beam under the programmed control of mirrors, prisms, or other special optics. As shown in Fig. 5.3, a single high-quality lens is used to focus the beam on the material surface. The spot size of the focused beam is controlled not only by wavelength, but also by the numerical aperture of the lens [see Eqs. (5.1)–(5.7)].[53] Because it often is easier to change the beam diameter (i.e., by varying an aperture, D) than the wavelength, a two-lens beam-expanding or beam-diminishing telescope is placed before the focusing lens to vary the focal spot size. Note that rarely are diffraction-limited conditions realized.

Intricate lines and patterns can be drawn by rastering the focused beam over the material surface. There are many ways to raster a beam. Figure 5.3 displays a scheme in which the focused beam is stationary, while the target substrate simultaneously rotates and translates beneath it. This scheme is but one of several techniques for moving the target. There exist many high-quality stages and substrate holders that can be moved and positioned with submicron accuracy under computer control. For systems in which the target cannot be moved, beam scanning can be performed by a variety of methods that fall into two general categories: those in which the focusing lens is placed before the beam-scanning optics, and those in which the beam-scanning optics are placed before the lens. The former category typically requires a long focal length lens to permit space for the scanning device between the lens and the target. Two examples illustrate these options. In the first example, two mirrors–one x-axis oriented and the other y-axis oriented–are used in an oscillating or gimballing motion to produce a "Lissajou-" type pattern on the material surface. In the second example, a rotating polygonal mirror is interposed between the lens and the target. As the mirror rotates, a beam is linearly scanned across the target. Angular adjustment of the polygonal mirror, at a rate much slower than the rotation rate, permits a series of parallel lines

to be drawn on the material surface. When a long focal length lens cannot be used, the scanning mechanism can be placed before the lens. However, the field of view of the lens must be sufficiently large to maintain focus as the beam "wanders" over the surface of the lens. Typically, the lens diameter is larger when the scanning mechanism is first than when the lens is first.

The above examples of BDSs have been for flat target surfaces. For curved and nonflat surfaces, optical components such as toric mirrors, axicons for generation of annular beams, and field-curving lenses can be used for processing.

5.4.2 Alignment, Positioning, Focusing

Alignment and positioning of a part or surface go hand in hand. Usually, it is necessary to consider the means for positioning prior to worrying about correct alignment. In general, positioning refers to the ability to move the part to meet requirements for movement in multidimensions with distance, velocity and acceleration, and incremental resolution criteria. Closed-loop control systems are often used with positioning systems to allow computer-controlled feedback and movement timing. Most positioning systems rely on x-y positioning tables, although commercial fixtures exist for rotational and angular motion. The size and weight of the workpiece dictate the size and choice of the motion system, because the mass of the part provides the inertia the positioning system must overcome. For applications of interest in this chapter, such as micromachining, semiconductor processing, and laser microchemical processing, highly precise lightweight motion tables are used. The x-y tables are composed of carriages and drivers (usually electric) that execute the motion in the desired direction. The electric drivers most commonly used are either stepping motors or dc-encoded motors. Stepping motors take a single discrete step for each voltage pulse received. This step-by-step motion can be easily computer controlled. Readout of the workpiece position is obtained by counting the drive pulses. The dc-encoded motors employ linear or synchronous motors and typically some sort of optical encoding scheme for highly accurate position readout.

Once workpiece positioning is established, alignment must be determined and calibrated (which means finding the correct workpiece position and orientation). Alignment can be a simple or complex process, depending upon the simplicity or complexity of the part to be processed. For a flat, square workpiece, alignment can entail no more than location of one corner of the square to serve as a reference point. This task can be accomplished by physically placing the part in the aligned position and calibrating the motion system. Alternatively, automated visual inspection and location can be used. For this technique, simple image recognition methods are employed to locate alignment marks purposely placed on the workpiece surface. The part is moved until the alignment marks are collocated with reference positions previously coded into the memory of the alignment system.

Automated visual systems for alignment often perform the secondary function of focusing the part with respect to the BDS. There are many kinds of auto-focus systems that have been described in the literature. A common one is based upon the Foucault "knife-edge" technique.[54] In this technique, a collimated input beam comes to a focus on the surface to be processed, and a reflected beam returns on the same path. A beam splitter sends some fraction of the return beam toward a secondary lens, which causes the return beam to pass through a focal point prior to impinging upon a split photodetector operating in a differential mode. As the beam diverges from the secondary focal region, it uniformly fills both sides of the split detector so that the difference is zero and no error signal is generated. At the secondary lens focal point, a knife or razor edge is placed perpendicular to the beam direction, and is then precisely positioned so that the edge just cuts the focal spot and diminishes the light incident upon the split detector. When the part is in

focus, the secondary focal spot coincides with the knife edge, and light remains uniformly split between both detector halves. However, if the part becomes either positively or negatively out of focus, the position of the secondary focal spot falls, either before or after the knife edge, causing more light to fall on one side of the split detector. An error signal proportional to the magnitude of the change in focus is generated and used to make a closed-loop servo system.

There are also diverse manual focusing methods that may serve well in a given situation. Often, the focusing or imaging lens system is an integral part of a microscope system that can be used for visual monitoring of the processed surface. The microscope can then be used to permit visual focusing of the part. Excimer laser systems present some challenges for focusing with microscopes, because the UV radiation of the excimer does not correspond to the visible light the user employs for sight. A surface that appears to be in focus to the eye may, in fact, be out of focus at the excimer wavelength of interest. One solution to this problem is to place a surface that fluoresces visible light under UV irradiation in the same plane as the part to be processed.[55] Irradiating this "quantum converter" surface with very low-intensity, high-pulse repetition rate excimer light induces a visible glow from the surface, which can then be used for precise focusing.

As with all focusing methods, the depth of processing must not exceed either the depth of focus of the lens, or the range over which the lens can move to remain in focus. As the spot of the image size is made smaller, this requirement becomes more of a constraint on the system.

5.4.3 Process Diagnostics

Lasers have a unique advantage over most material processing tools: they can also be used as *in situ* process diagnostics detectors. Beyond the ability for accurately measuring the distance to a surface, lasers have been used to monitor surface deflection, surface temperature, and surface contamination level. Lasers can also be used to monitor the ablation or deposition product species, either in particle form, using Mie or Rayleigh scattering techniques, or as atomic or molecular species, using spectroscopic assignment. In addition, lasers can be used to monitor the surface corrugation and topology, either by scattering or by a time-resolved high-gain optical imaging technique. In this latter technique, the pulsed laser is used to briefly illuminate the surface, and the resultant illuminated image is amplified manifold in a laser gain media. The optical properties of a thin-film deposition process can also be measured by monitoring the change in the laser electric field polarization vector upon reflection. Similarly, the development and subsequent shifting of interference fringes in a reflected beam can be used to monitor the film thickness. A laser can be used to monitor the supporting apparatus or the feed lines in a material processing station, such as measuring the concentration and flow of a particular reagent or the stability and speed of a moving workpiece.

5.5 Utility and Limitations of Laser Processing

A common, but now dated, cliché is that the laser is a solution in search of a problem. While this cliché may have been true at one time, the rapid development of laser science and technology over the past 10–15 years has led to the widespread use of lasers. In materials science, chemistry, physics, environmental analysis, medicine, biology, seismology, and engineering, the laser has developed into a virtually indispensable tool. Similarly, commercial and industrial laser development has resulted in significant military and industrial laser use: range finding, biological agent detection, guidance, packaging structures for semiconductor logic and memory chips, automotive welding, magnetic disk fabrication, video disk mastering, and many forms of drilling and cutting. There are many potential applications for the laser, which are limited only by the creative insight of the technologists involved in this work. Nevertheless, a laser solution may not be optimum for

a given situation, and attempting to force-fit the laser into a particular military, commercial, or industrial setting may be likened to forcing the proverbial square peg into a round hole. A discussion follows of the conditions that are appropriate or not appropriate for the use of a laser.

It should come as no surprise that in the majority of nonmilitary industrial settings, cost/benefit considerations almost always determine the suitability of laser processing. For military ventures, the superiority or unique quality of a product fabricated or made possible by a laser technique is often given high consideration. So, just when is laser processing an acceptable technique for a microengineering application? The many answers to this question can be distilled, as follows:

- When the laser provides a unique and desirable attribute or quality to a part or process that can be obtained by no other technique
- When the laser provides for unequaled reliability in a finished part
- When the laser process permits significantly increased throughput and efficiency, resulting in superior cost effectiveness and a high benefit-to-cost ratio
- When the laser can provide the lowest cost solution, based on many factors, to a fabrication problem

It is often a combination of these answers that justifies the use of laser processing. First, lasers can, and do, perform amazing functions that routinely tip the scales in favor of their use. Often, other techniques and procedures are hard pressed to perform the same functions. If such a function is required for a given application, then a laser should rightly be chosen. Second, laser processing should be chosen if the finished part is to possess high reliability or smooth functioning. This is particularly true if the part is to be used in remote or inaccessible environments that would make replacement or maintenance difficult or impossible. Satellite-borne instrumentation or components are clearly one of these categories where high reliability is paramount. Chemical sensors and associated electronics situated in hazardous environments are another category. Industrial and military users pay for high reliability, and if a laser process can provide that degree of security, then a laser is more likely to be employed. Third, a characteristic of lasers that may result in their use is the potential for high throughput. Often, the speed of laser microprocessing distinguishes it from other methods and techniques. Lasers with large-area beams and/or very high pulse repetition rates possess inherent capabilities for rapid processing. Even if the quality of a processed part or surface obtained through laser and nonlaser methods is similar, a laser technique with a high throughput will win out over a nonlaser method. Simple economics and cost effectiveness dictates that this will be so. Fourth, laser processing should be used if it can produce a given part with sufficient quality and reliability and is also the low-cost solution. The reader may wonder how an expensive laser could ever be the low-cost solution. However, looking at the isolated cost of a laser does not often present a clear picture of the overall cost. Because a laser technique may impart special desirable characteristics while providing for high reliability/repeatability and high production rates, it may often provide the lowest overall cost.

As prevalent as lasers are, they clearly are not used in all situations, nor should they be. Many situations exist where the use of lasers for material processing is not optimum. Even though a laser might provide a processed part of high quality, another simpler, less exotic technique may provide a slightly lower quality part that nonetheless is "good enough." This is an important point that cannot be overemphasized. High quality alone will not always justify the use of a laser technique. The laser's high quality must be coupled with an ability to fill a niche or special requirement that cannot be filled by another method. For example, consider laser wire stripping (to be discussed later in this chapter). Without a doubt, excimer laser wire stripping is of superior quality to CO_2 laser-based stripping. However, many companies use CO_2 stripping because it costs less and is of acceptable quality for the particular application.

5.6 Microengineering Applications

5.6.1 Overview

As a general technology, microengineering has enormous applications in both current and future aerospace systems because it purports to offer functionality at a reduced size and volume. Both of these reductions are valuable benefits to an aerospace system design engineer, whose nominal task is to show functional value for every unit mass to be sent aloft. Aerospace systems—whether for aeronautical or space applications—rely on "engineered" materials and structures to withstand the environment at takeoff/launch, high-speed cruise/ascent-to-orbit, and landing/deorbit and reentry. The requirements for operation can be severe and may impose a wide variation of operational tolerances in temperature (~100 to ~400 K), pressure (1 to $\sim 10^{-13}$ atm), mechanical loading (0 to ~10 g), and radiation flux (0.3 rad/yr to 10^6 rad/yr). Because of these requirements, and similar to some terrestrial applications, aerospace systems manufacturing relies on "clean" material processing techniques and employs a higher level of precision/tolerances than most other manufacturing domains. In this regard, laser-based material processing offers a capability for developing and processing engineered materials with high precision and without physical contact.

The vast number of aerospace applications for microengineering requires the fabrication of precision microholes and cuts in numerous materials, the precise fusion of "dissimilar" materials, and the controlled deposition and adhesion of an overlayer material. A closely related, but somewhat new application, is surface texturing to imbue a material with new characteristics. Finally, a major application is the development of microengineered components/devices that contain onboard "intelligence."

Precision microholes are used in the development of acoustic suppression systems within jet-engine cowlings, in the controlled metering of fluids, and in the development of fuel-efficient micropropulsion systems for future micro/nanosatellite applications. Precise fusion of dissimilar materials is used in the development of functionally gradient materials, such as for thermal control and for developing integrated component packaging. Controlled deposition of novel materials is used in the growth of optically selective films; in the deposition of specialized films for tribology (e.g., dry lubricants for space applications); and in the deposition of thick coatings for protection against the environment. Surface texturing applications include the removal of oxides from metals prior to additional processing and the ruling of fine lines in large array (m^2) polysilicon solar cells. Finally, there are also applications of laser processing to components/devices/microsystems, which are fabricated using semiconductor materials and employ integrated circuit (IC) processing techniques. For these applications, the laser-based processing techniques are used in the high-value-added processing steps, such as for via hole patterning in multichip-module (MCM) packaging, for IC circuit/device trimming/turning, or for rapid prototyping or repair operations. As a further example of high-value-added processing capable only by laser, consider the mundane application of drilling holes. Experiments show that controlling the hole shape (e.g., noncircular) and taper (e.g., noncylindrical) can result in beneficial properties for fluid and acoustic dynamics. This is also true for cutting trenches/lines. Experiments show that trenches/lines cut with noncylindrical shapes or those having rounded bottoms are less likely to fail due to stress concentration. The use of smooth tapers and minimizing the number of sharp corners is considered a viable solution.

5.6.2 3D Microfabrication

Much has been written about direct-write laser micromachining via ablative and chemical assist techniques.[56] These techniques essentially employ a 2D mass removal process with sequential rastering to fabricate true 3D objects. It is also possible to imprint a 3D pattern directly into a

material without laser rastering. This imprinting can be done in a photosensitive material that absorbs at the laser wavelength. By controlling the laser dose and using direct-write patterning, a true 3D image can be imprinted in the exposed volume of the material.[12] There are numerous materials that have these photosensitive characteristics, including certain glass/ceramic materials that also have technological applications to Aerospace systems. In particular, lithium-aluminosilicate glass is a material with ingredients that enable a photographic image to be transferred to the glass after UV exposure and subsequent heat treatment. The latent image is captured by a devitrification process in the glass. There are over 5000 compositions of this type of glass, some of which go by the trade names of Fotoceram, Pyroceram, Photsitalls, Vitrokeram, and Foturan.[57] For Foturan (manufactured by Schott Glassworks, Mainz, Germany), whose photosensitive characteristics arise from the additions of Ce_2O_3 and Ag_2O, photoexcitation and devitrification proceeds as follows. In the unexposed glass state, both the cerium (Ce) and the silver (Ag) are stabilized as ions (Ce^{3+}, Ag^{+}). Upon UV illumination within the material absorption band (Fig. 5.4),

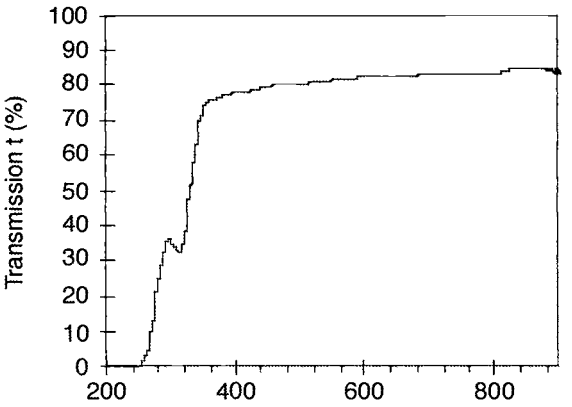


Fig. 5.4. Transmission curve of Foturan, per manufacturer's data for a 1-mm-thick sample.

there is an electron transfer process that neutralizes the Ag^{+} (reaction is shown in Eq. 5.23) and stores the latent image. The ceramization or baking step aggregates the silver nuclei and forms silver-lithium silicates (Eq. 5.24). Upon exposure to hydrofluoric acid, the silicates etch faster than the unexposed glass (Eq. 5.25). Figure 5.5 and 5.6 present data, measured at The Aerospace Corporation (Aerospace), which show an etch rate difference approaching 20:1.

$$\text{Illumination: } Ce^{3+} + Ag^{+} >> Ce^{4+} + Ag^{o} \tag{5.23}$$

$$\text{Ceramization: } nAg^{o} >> (Ag^{o})_n - Li_2SiO_3 \tag{5.24}$$

$$\text{Isotropic Etching: } Li_2SiO_3 + 3HF >> 2LiF + H_2SiF_6 + 3H_2O \tag{5.25}$$

Using the above photoexcitation process, along with some detailed understanding of the nonlinear fluence dependence properties of the material, Aerospace developed a true 3D direct-write laser micromachining technique. The technique uses a focused pulsed UV laser to expose a precise volume of the material. Under computer XYZ motion control, a pattern is "written" in the photosensitive glass. There is no application of resist material, and in general, the exposed volume has a depth dimension on the order of the confocal parameter b. Equation 5.3 can give a general idea of how large a part can be fabricated with this technique. Setting ω_o between 0.5 and 5 μm and λ between 0.2 and 0.3 μm gives a range in the confocal parameter, b, between 5 and 800 μm.

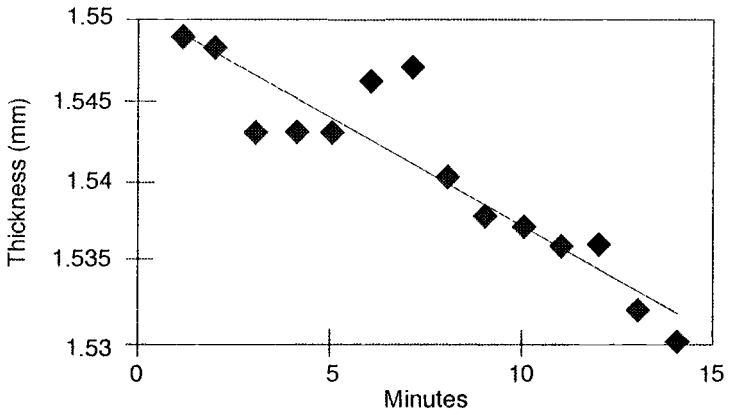


Fig. 5.5. Etch depth as a function of time for an unexposed Foturan sample. Sample was exposed to the ceramization program bake. The linear fit gives an etch rate of 1.3 μm/min.[12] Foturan unexposed—etch depth. $y = -0.0013x + 1.5504$; $R^2 = 0.846$.

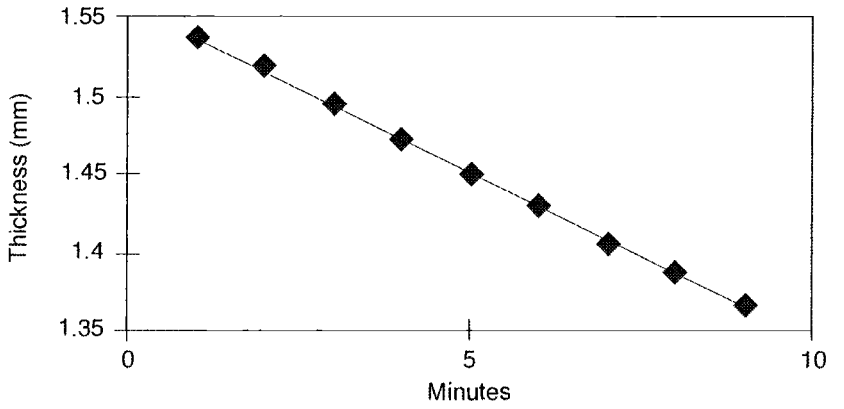


Fig. 5.6. Etch depth as a function of time following 248-nm laser irradiation at 200 Hz and 1.8-mW average power. The linear fit gives an etch rate of 21.6 μm/min.[12] Exposed Foturan—etch depth. $y = -0.0216x + 1.5591$; $R^2 = 0.9995$.

Key aspects of the process are the laser wavelength, the energy dose deposited, and the single shot fluence (J/cm^2). The laser wavelength influences the absorption depth, the total energy dose applied influences the HF etching rate, and the single-shot laser fluence defines the damage threshold. Curved 3D structures can be fabricated by also controlling the spatial contour of the laser beam near its focus. Experiments were conducted using two laser wavelengths (248 and 355 nm). Mesoscale devices were fabricated ranging from 400 to 1500 μm thick with microscale structure in the range of 10 μm. Figure 5.7 shows an optical microscope photograph of two resonant beam structures. Figures 5.8 and 5.9 show scanning electron micrographs (SEMs) of spire structures fabricated by programmed scanning of the X-Y positioning stages with a focused 248-nm laser beam, where the focal spot size is at a constant depth beneath the surface (no Z motion). Figure 5.8 shows an array of pyramidal tips formed by overlapping X and Y scans. The two SEMs show

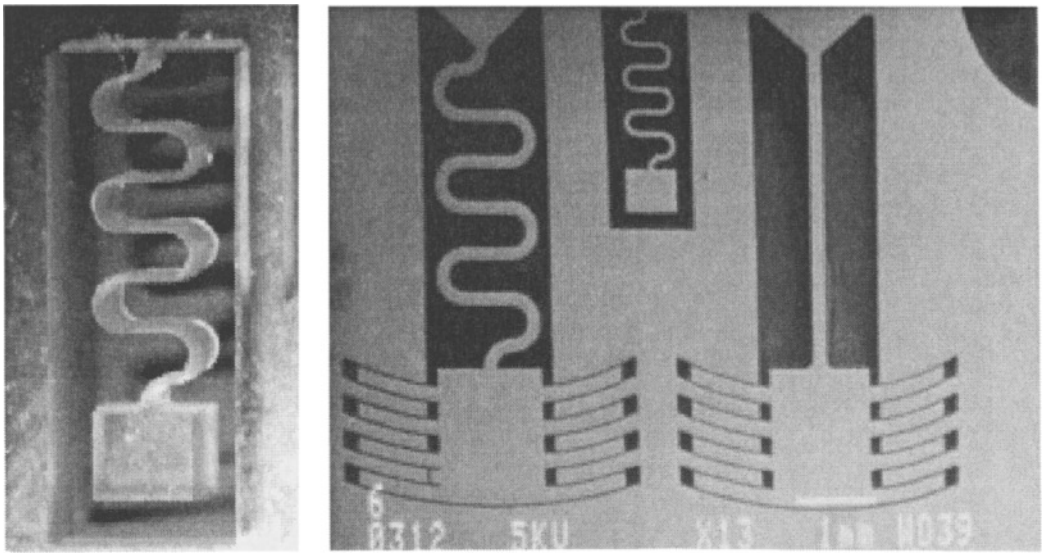


Fig. 5.7. Optical microscope photograph of a resonant beam structure: (left) the structure is approximately 5 mm long and 1 mm deep; the spring meander is 20–40 μm wide; (right) resonant structures with patterned gold.

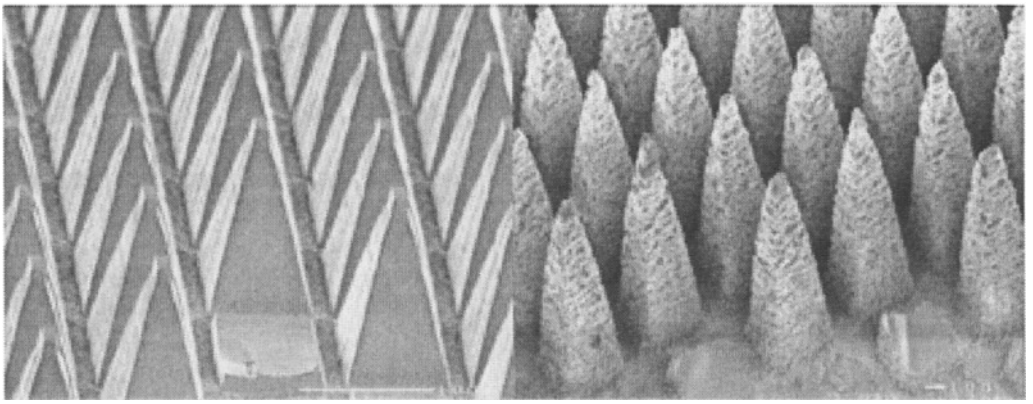


Fig. 5.8. Two SEMs of arrays of spires. The spires are ~300 μm high. By controlling the ceramization step, the spires can be made to have smooth walls or a scaly texture.

microstructures with sharp and rounded corners. Figure 5.9 also shows two SEMs: a series of concentric rings with a central spire formed by coordinated X-Y motion with nonuniform velocity, and a series of rings about a single spire that share a common tangent (1.2 mm maximum diameter by 0.3 mm deep).

The micromachining of "glasses" and ceramics by direct-write techniques enables these materials to be used for other applications besides those in future space systems. For example, these materials can be used in biological applications where glass is preferred over plastics. The applications become especially intriguing if silicon can be directly fused to these aluminosilicate "glasses" via a low-bulk temperature process. Aerospace is investigating the use of laser-based direct-write techniques for silicon-"glass" fusion. The successful development of these

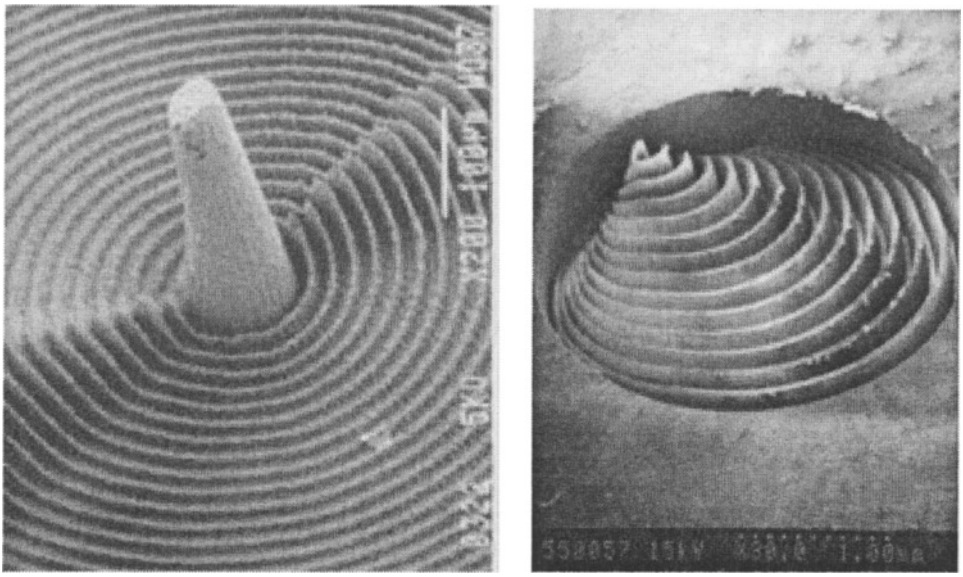


Fig. 5.9. Two SEMs of mesoscale structures fabricated in glass/ceramic material. The left SEM center spire is ~250 μm high, but thin walls are approximately 10 μm thick. The total structure is similar in dimension to the right SEM photo (~2.0 mm). In the right photo, a single spire tip is surrounded by a series of walls, where each wall is of variable height.

techniques will permit the integration of electronics, MEMS, and MOEMS (Microoptoelectromechanical systems) as fabricated in conventional silicon-based foundries with microstructures/devices fabricated in glasses/ceramics. Glass ceramic devices will be especially useful in aerospace applications where caustic propellants that can etch silicon must be used (e.g., hydrazines).

5.6.3 Laser Via Production for MCMs

In 1991, IBM introduced the first commercial mainframe computer that incorporated laser ablation technology in the manufacturing.[58] This milestone was the culmination of nearly a decade of scientific, engineering, and manufacturing effort. Extensive research and development (R&D) on 308-nm laser ablation of a polyimide dielectric resulted in the first IBM prototype ablation tool in 1987 for the production of via holes in thin-film packaging structures. This prototype, similar to a step-and-repeat photolithography exposure tool, evolved into a full-scale manufacturing tool that currently uses sophisticated beam shaping, beam homogenizing, and projection optics.

In 1982, Srinivasan and co-workers discovered the spontaneous removal of material from the surfaces of organic polymers subjected to 193-nm pulsed excimer laser radiation. This discovery led to much scientific interest in this process.[59] IBM was interested because a new method was now available for creating via holes in polymer dielectrics, and this method promised high processing speed and reduced costs. In the early 1980s before this discovery, when thin-film processes for MCM fabrication were being defined, wet etch was the only available low-cost via formation technique. Therefore, the 1983 announcement that polyimide undergoes ablation by 308-nm excimer laser radiation significantly changed the situation.[60] A process using this wavelength meant that optical problems associated with lenses and mirrors at 193 nm would be greatly reduced with operation at 308 nm. Additionally, and perhaps more important, the new 308-nm xenon chloride (XeCl) laser was of higher quality and reliability than the previously used 193-nm argon fluoride (ArF) laser.

IBM exploited the new process by launching an engineering program to develop a 308-nm ablation tool for via formation in MCMs. A chief concern was whether a commercial laser of sufficient quality for a manufacturing environment was available. The common research-type excimer lasers of those years were unsuitable for manufacturing. The component lifetime of the lasers was short, and the output power degraded far too quickly. New lasers offered by several companies, and designed with industrial use in mind, helped to resolve the problem and enable a prototype tool to be built. Further engineering decisions regarding image projection strategy, beam uniformity, projection mask technology, and beam delivery optics culminated in the first prototype ablation tool in early 1987. Much of this early work has been documented by Lankard and Wolbold.[61]

Figure 5.10 displays a schematic of a cross section of the thin-film packaging structure used in the MCMs of the IBM ES9000 computer.[62] Clearly shown are the tapered vias created in the polyimide interlevel dielectric. IBM has explored four different technologies for via production: wet etching, laser ablation, reactive ion etching (RIE), and conventional lithography using photosensitive polyimide (PSPI). Laser ablation has been shown to have fewer processing steps than the other technologies and provides higher throughput in terms of finished substrates per hour. Additionally, ablation is the only fully dry process, allowing additional cost savings. Wet etch, RIE, and PSPI are all highly process intensive, and thus expensive, relative to laser ablation.

The via production process, as currently practiced at IBM, involves four steps: application of the polymer to the substrate, curing, via hole formation by ablation, and final plasma treatment for laser-generated debris removal. The via ablation step is done by projecting the image of a laser-illuminated mask onto a polymer-laden substrate. As with modern photolithography, exposure of the polymer occurs in a step-and-repeat mode, with one chip site on the MCM being ablated in each step. Many stepping actions are required to cover the entire substrate. The details of the optics and BDS have been extensively described in the literature[63] and will be briefly reviewed here (see Fig. 5.2). Pulsed radiation from a commercial industrial excimer laser operating at 308 nm is passed through beam-shaping optics prior to impinging on a beam homogenizer. This "fly's eye" type homogenizer produces a beam uniformity that varies by no more than 3% at the

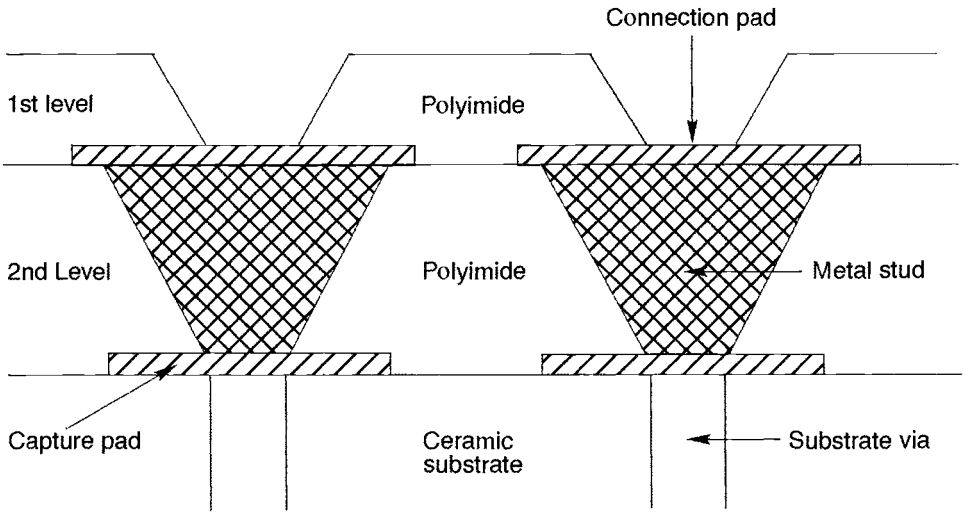


Fig. 5.10. Cross section of a thin-film packaging structure.

substrate image plane. This uniformity is achieved with only a 5% energy loss in transmission. A field lens images the homogenizer output onto a dielectric mask that contains the pattern of vias to be processed. In turn, the mask is imaged by a high-quality, UV-compatible 1:1 transfer lens onto the substrate. The positioning and alignment of the substrate, and focusing of the imaging system, are all under computer control. Figure 5.11 displays an electron microscope photograph of a laser-processed via hole in polyimide surrounded by an ablated pattern of lines and spaces of micron-sized dimensions. A higher resolution view of the lines and spaces is shown on the right of Figure 5.11.

In this via hole production application, ablation is used to generate approximately 10^5 vias on a single polymer level of the substrate. The vias are 75 μm in diameter, and are created in 18 to 20 μm thick polyimide. The two critical parameters for the laser ablation process are the laser fluence and the number of pulses per via site. Both of these parameters can be well controlled, but in practice, tight control is not necessary. Variations in laser fluence of up to 5% are acceptable because many pulses are required to completely form the via and any variation is averaged out. Depending on the particular substrate and process, a laser fluence in the range of 150 to 300 mJ/cm^2/pulse is employed. The number of laser pulses needed for a given polymer thickness is determined from standard ablation rate curves. At an ablation rate of 0.1 μm/pulse, a 20-μm-thick polymer sample would require about 200 pulses to completely form the via. At 200 pulses per second from the laser, via hole formation would take approximately 1 s. A suitable number of excess pulses are used to ensure that all vias are completely formed, and that small variations in polymer thickness and pulse-to-pulse fluence do not adversely affect the result. Since nonerodable metal pads reside at the via bottom, these excess pulses cause no further ablation at the fluences employed.

As can be partially seen in Fig. 5.11, the via wall angle is less than 90 deg. This angle results from the inability of the lens system to transfer high-contrast images from the mask to the polymer surface. Additionally, the high intrinsic absorption of the polyimide further inhibits perfect fidelity of image transfer. This wall angle turns out to be quite useful in increasing the adherence of metal that is subsequently deposited into the via hole. The via wall angle can be controlled very tightly between about 40 and 70 deg by changing the focal plane for ablation.[64]

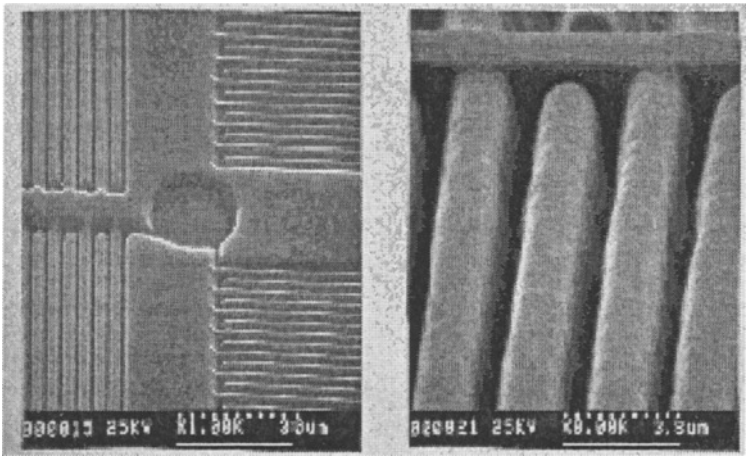


Fig. 5.11. SEM photos of polyimide via hole (left) with micron-sized lines and spaces (right). The scale bar on the left hand SEM is 30 μm; the bar on the right is 38 μm.

IBM's early laser-projection tools for via fabrication were to a large extent designed and built in house. While the tools served the needs and demands occurring from 1985 to 1987, such as low-volume production rates and flexibility, they were woefully inadequate for high-volume manufacturing. As a result, continual improvements were made in the tools. Table 5.5 gives a profile of the major tool improvements between 1987 and 1994. While not explicitly stated in the table, improvements in the beam-delivery optics, the laser, and the computer automation drove most of the advances. For example, improvements in the fabrication and assembly of the imaging lenses allowed significantly smaller vias to be created. Likewise, longer gas lifetimes, increased reliability in high-voltage components, and advanced discharge electrodes all greatly improved

Table 5.5. Improvements in the Laser Via Process

Parameter	1987	1994
Process cycle time	25 min	12 min
Via wall-angle range	50–65 deg	20–75 deg
Minimum via size	11 μm	6 μm
Exposure field size	14×6 mm^2	40×30 mm^2
Substrate alignment	Manual	Automated
Alignment accuracy	±7 μm	±1 μm
Tool availability	50%	>95%
Process yield	~ 85%	>99.99%

the laser. This improvement resulted in vastly improved tool availability, defined as the percent of time the tool is operational relative to the demand time. Computer-controlled alignment and focusing reduced the errors in substrate positioning to the micron level. As a result of these many improvements, the mean-time-between-failure (MTBF) for the ablation system grew from about 150 h to over 700 h for the current system. The MTBF for the laser itself went from a few hundred hours to well over 1000 h–a significant measure of success. The demand time for the via fabrication tool is approximately 150 h/week.

The laser process for via hole fabrication has been used at IBM for nearly a decade. During this time, several billion vias have been produced in a variety of thin-film packages, including MCMs. There have been no known field failures of vias. Clearly, this is one of the most robust, reliable, and high-yield technologies in the thin-film fabrication industry.

IBM chose the laser technology for via hole creation because it could provide highly precise and defined vias, high-speed parallel processing that was compatible with large-volume manufacturing, an environmentally sound dry process, higher than acceptable reliability, and the most cost-effective manufacturing solution.

5.6.4 Laser Wire Stripping

Advanced techniques for localized removal of plastic insulation have found significant application in microelectronic packaging and interconnect technologies. One such application is the laser via process described in the preceding subsection. Another application is the removal of the plastic insulation that typically covers electrically conducting wires. This wire stripping has been

performed by a variety of techniques in the aerospace, data storage, and electrical industries. Techniques include mechanical cutting, abrasive action, electrical arcing, chemical etching, and simple thermal methods such as burning. These thermo-chemical-mechanical techniques can be used, and often are, when the wire is of sufficient durability, and when the processing speed, precision, or cleanliness of the stripped region is of no great concern. However, for many applications in the microelectronic and computer industries, strip length, precision, and cleanliness are major concerns and thus significantly restrict the available choices for wire stripping.

In the magnetic disk drive industry, a small read/write head is part of a ceramic slider that is suspended above a spinning disk. Electrical signals are conveyed to and from the head by very fine magnet wires (20–50 μm diam) that connect to the supporting electronics. These wires must be bonded to the head in such a manner that electrical contact is made. Adequate bonding only occurs if metal-to-metal contact is made between the magnet wire and the bonding pad on the slider. This contact is accomplished by the removal of the polyurethane-based plastic insulation in the vicinity of the region to be bonded. The conventional industry technique for wire stripping has been the use of electrical arcing. While this has been an inexpensive and efficient method, it lacks the cleanliness, and particularly the strip location precision, necessary for current and future generations of head suspension systems. Imprecision in the strip length or location may result in electrical short failures. To resolve this problem, the magnetic drive industry is exploring the use of laser wire stripping.

Laser-based methods for wire stripping, usually employing pulsed or cw CO_2 lasers, have been used in many industries since the mid-1970s.[65] Laser wire stripping offers the advantage of being a noncontact method–a significant factor for the small, precision parts currently used in microelectronics. Another advantage is that light can be imaged to small, divergence-limited spot sizes, allowing very small lengths of insulation to be stripped. For example, using the UV wavelengths of an excimer laser, a strip length as short as 10 μm should be possible. However, the light absorption properties of the insulation are much more important for effective wire stripping than are the imaging capabilities of the laser wavelength employed. As will be discussed, precision laser wire stripping requires the plastic insulation to absorb the laser radiation very strongly, and requires the laser energy to be delivered in pulses short enough to limit thermal diffusion effects.

Primarily because of its high power and relatively low cost, the CO_2 laser has been used in many wire-stripping applications. The laser emits IR radiation peaked near the 10.6-μm wavelength–a spectral region where most plastic insulations absorb only moderately. Typical absorption coefficients for plastic films are in the range of 10^2 to 10^3 cm^{-1} for CO_2 radiation. Because of this limited absorption, wire stripping with the CO_2 laser requires that the light be focused for sufficient intensity to break down the insulation. Further, the moderate absorption can limit the efficiency with which the last few insulation layers are removed, resulting in thin nonconducting layers remaining on the wire. To prevent charring of the edge in the stripped region, the CO_2 laser is often run in a gain-switched pulse mode. Focusing of the pulse onto the insulation surface initiates dielectric breakdown and rapid ejection of molten and solid material. Stripping of the localized wire region can be accomplished in just 2 or 3 pulses incident from a few circumferential directions. The strip edge definition is not optimum, and debris generation can be problematic. Nevertheless, in many instances, this edge definition is of acceptable quality. When exposed, the high reflectivity of the bare (stripped) wire to CO_2 radiation protects the metal from unacceptable levels of heating and consequent mechanical fatigue.

In addition to the CO_2 laser, the excimer laser is used in many wire-stripping applications. Specifically, for high-precision, efficient removal of wire insulation, the excimer laser is the laser of choice for several reasons.[66] One reason is that most plastic insulators strongly absorb UV light.

The short (10–20 ns) pulse of the excimer laser, along with the deep UV output, provides a strong absorption event in a very short time. It is almost ideal for wire-stripping applications. As the pulsed excimer radiation interacts with the plastic insulation, the light only penetrates a short distance. For polyurethane insulation, the measured absorption coefficient at 248-nm wavelength is about 7×10^4 cm^{-1}, providing a short penetration depth of less than 150 nm. The deposited energy is some 2–3 orders of magnitude larger than that for the CO_2 laser. At the 308-nm excimer laser wavelength, the absorption coefficient is less than 3000 cm^{-1}, accounting for why this wavelength provides stripping action considerably less appealing than the 248 nm wavelength. Regardless, the absorption coefficient is larger in the UV than the IR, which results in a number of advantages that makes the excimer technique the better choice. For example,

- In the UV, the shallow penetration depth causes only a thin layer to be ablated per pulse, with the result that many pulses are required to completely strip a wire containing a 5-μm thickness of insulation. This approach to material removal promotes control and aids in defining the edge with precision.
- Strong absorption in the UV aids in the removal of the last residual layers of insulation, helping to provide clean surfaces.
- Strong absorption means that the laser light decomposes the organic matter, which significantly reduces debris.
- Strong absorption means that the laser beam does not have to be focused on the surface; rather, the incident laser intensity can just be above some threshold value for insulation removal.
- Strong absorption means that blocks of insulation can be removed by image projection rather than by scanning a focused laser beam. The result is increased efficiency and high throughput.
- UV laser short pulse limits thermal diffusion and heating of the wire in the laser irradiated region, thereby aiding edge definition and precision.

Figure 5.12 displays an optical configuration for performing excimer laser wire stripping. Pulsed 248-nm radiation illuminates a mask. The mask is imaged by a single lens onto the plane where the wire is vertically held. The image size at the wire plane is controlled by the size of the mask and the imaging system. The image height determines the wire strip length. Because the width of the image is much greater than the wire diameter, most of the light passes by the wire

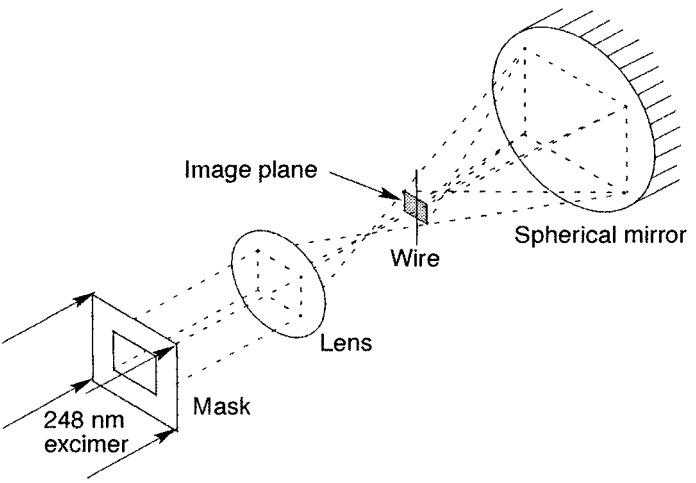


Fig. 5.12. Optical schematic for excimer laser wire stripping.

and is intercepted by the spherical mirror. At the mirror, the radiation is reflected and reimaged onto the back side of the wire for insulation removal. In this manner, excimer pulses strip the entire 360-deg circumference in a single-beam configuration, without resorting to multiple beams or rotating wires.

Excimer laser wire stripping is performed on polyurethane-based insulation using incident fluences of 300 to 400 mJ/cm^2/pulse. Approximately 100 pulses are required, and can be accomplished in less than 1 s using high-pulse repetition rates. Figure 5.13 shows electron microscope photographs of an excimer-laser-stripped magnet wire. The stripping is clean and well defined, with no evidence of debris.

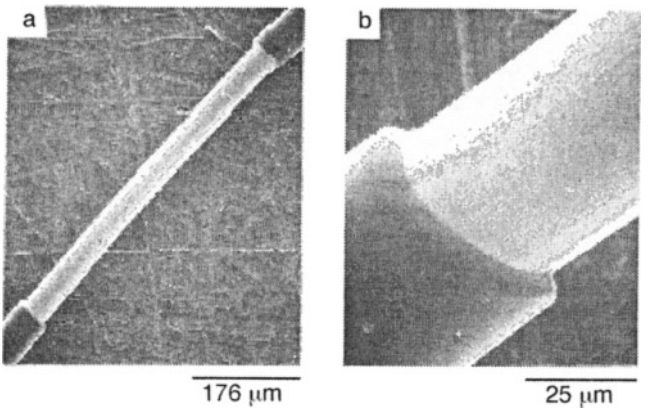


Fig. 5.13. SEM photographs of 248-nm stripping of magnet wire.

One drawback of the excimer laser is the expense required to operate and maintain this device. This is the reason why CO_2 lasers, although not as precise, are often considered "good enough" for the application. Furthermore, once the metal surface is exposed, excessive excimer intensity may cause heating effects that can prove deleterious to the mechanical strength of the wire. These effects are much more of an issue with excimer lasers than with CO_2 lasers, because UV radiation is more strongly absorbed by metals than is IR radiation. For this reason, the number of excimer laser pulses and their intensity must be carefully chosen and maintained.

In conclusion, laser stripping of magnet head wire has proven itself as an industry standard technique for a variety of reasons. First, although trade-offs exist in the merits of excimer laser versus CO_2 laser-stripping techniques, laser stripping provides much higher throughput than non-laser stripping techniques. Second, this higher throughput is coupled with high precision in both the length and placement of the stripped region. Third, the noncontact laser process reliably produces a finished part that ultimately costs less than parts stripped by other techniques.

5.6.5 Laser Texturing of Magnetic Disks

The data storage industry has an increasing need for higher storage capacities.[67] One method for increasing capacity is to increase the areal storage density on the disks within a magnetic hard drive. The physics of the magnetic read/write process dictate that a higher areal density can be achieved by having the magnetic head (which contains the read element) fly closer to the rapidly spinning disk. Currently, the magnetic head flies 30–100 nm above the disk. To have the head fly closer to the disk, and thus achieve higher areal storage densities, requires that the disk surface must be smoother and flatter than in the past.

One problem with reducing the surface roughness of a disk is the increased area contact between the disk and the "slider." (The slider is the smooth ceramic element containing the magnetic read and write head, which is suspended above the disk when it is spinning.) The disk/slider contact results in frictional forces that tend to increase wear on both the slider and the disk. In extreme cases, the slider can become stuck to the disk surface. This phenomenon has earned the moniker "stiction" for the sum total of all attractive forces between the smooth slider and the smooth disk. The tribology problem in the magnetic storage industry has been, and continues to be, finding ways of reducing stiction and wear to acceptable and controllable levels.

Because stiction is proportional to the area of contact between a slider and disk, one way to achieve low stiction is to minimize the contact area. For years, the storage industry has minimized the contact area by performing a full-surface mechanical texturing of the disk. Surface texturing alters, in a controllable fashion, the surface topography such that the contact forces between the disk and the slider are reduced in a known manner. This approach has worked for relatively low areal density disks that do not require very low slider flying heights. However, for the current and future generation of hard drives, this is an unacceptable approach because the full surface roughness does not permit the magnetic head to fly close to the disk. Because of this roughness, the industry has pursued disk texturing in a dedicated "landing zone" near the inside diameter of the disk. In this zone, the requirements of slider-disk tribology can be optimized apart from the requirements of the data zone that constitutes the majority of the disk surface. In the landing zone, the slider can be parked and latched after the drive has been shut down.

For texturing of the landing zone, pulsed laser irradiation has been demonstrated to be effective.[68] At high repetition rates, a short pulsed laser creates discrete topographical features that have domelike protrusions, or "bumps." The bumps are sized and spaced such that hundreds of them are present under the slider, serving as smooth support points. Laser zone texture (LZT) provides an efficient, quick, and high-precision texturing method with excellent tribological performance. This method is so successful that it has permeated the entire magnetic storage industry, becoming a standard manufacturing tool.

The technique of LZT is dependent on the type of disk substrate employed. Current-generation disks consist of an aluminum substrate plated with about 10 μm of amorphous nickel phosphorous (NiP) to improve smoothness and hardness. It is this NiP surface that is zone textured using a solid-state neodymium laser. Subsequently, the magnetic layers are deposited upon the textured NiP layer. A carbon wear layer, followed by a thin lubrication layer, completes the thin film disk. Future requirements of disks demand that they be smoother, flatter, harder, and stiffer than aluminum. Glass substrates nicely fit this demand, and are starting to appear in advanced hard drives. Because of its hardness, the glass surface can be directly textured. The differing optical properties of glass in comparison to metal (highly transparent to 1 μm wavelength Nd laser light) require that a pulsed CO_2 laser perform the texture operation, followed by the deposition of magnetic, wear, and lubrication layers.

For NiP texturing, a high-repetition rate, Q-switched Nd:YLF or Nd:YVO_4 diode-pumped solid-state laser is usually employed.[69] This type of laser is chosen for several reasons: (1) the near-1-μm wavelength radiation is strongly absorbed by the NiP, permitting efficient texturing; (2) the short-duration pulse (40–80 ns) creates high-focus intensities that rapidly melt the NiP and initiate the bump-formation process; (3) because one bump is formed per incident laser pulse, the laser's efficient operation at high repetition rates (20–80 kHz) permits rapid texturing and thus high disk-processing rates; (4) the pulse-to-pulse energy stability is excellent (less than 1% variation), resulting in highly uniform texturing.

Glass substrates strongly absorb near a wavelength of 10 μm. Because this absorption coincides with the output of CO_2 lasers, this laser is employed for LZT of glass.[70] Pulse modulation of CO_2 lasers is obtained by either direct modulation of the RF plasma excitation, or by use of an acousto-optic modulator operating on a cw beam. The latter technique permits rapid pulse repetition rates from 10 to 100 kHz. As with the neodymium lasers, the pulse energy variation is low, consistent with the observed uniform texture process. In contrast, however, the practiced modulation methods limit the pulse duration to no less than a few microseconds. This limitation presents no practical constraint because well-formed texture bumps are nonetheless created.

A schematic of the texture apparatus, for either NiP or glass texturing, is shown in Fig. 5.3. The rapidly pulsed laser is shuttered to control the placement of the texture bumps on the disk. The disk is mounted on a rapidly rotating hub, which in turn is translating. At the same moment, the shutter is opened, and a continuous train of pulses passes through a lens and is focused on the disk surface. Each pulse creates one texture bump. After 1 to 2 s of exposure, the shutter closes and terminates the operation. The spinning and translating motion of the disk creates a spiral of texture bumps in a narrow zone that is 2–5 mm wide. The texture zone placement is precise to within 10 to 50 μm.

One of the distinct advantages of LZT is the microtopography of the generated bumps. Because of this microtopography, the top portions of the bumps are extremely smooth and rounded, which aids in their durability and reduced stiction behavior. For NiP or glass texturing, a single focused pulse creates a single bump with a diameter from 5 to 20 μm, depending on LZT conditions. Further, the height of the texture dome ranges from just a few nanometers to several hundred nanometers. These bumps, though different in geometry for NiP and glass (see below), are extremely shallow, with the slope angle on the side being no more than 1 deg. If the correct pulse energy is established, true nanomanipulation of the texture bump heights can be achieved. In essence, micro- and nanotechnology is positively affecting the data storage market through laser topographic modification.

Figure 5.14 displays the so-called sombrero bump that forms on NiP and the smooth microdome that forms on glass surfaces. Qualitatively, the individual bump geometry is different, reflecting the different formation mechanisms for both surfaces. On NiP, the short and intense pulse

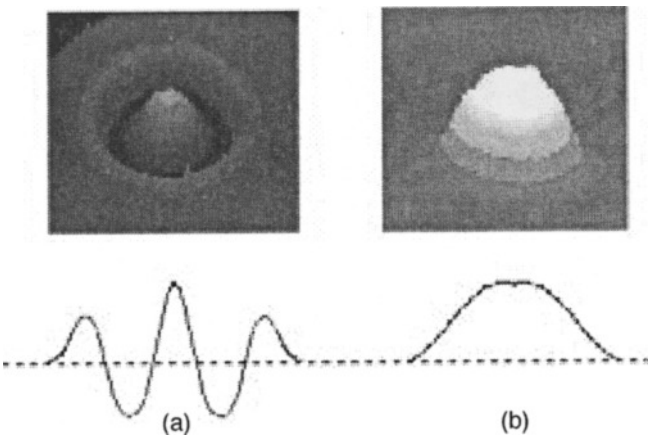


Fig. 5.14. Examples of laser texture bumps. (a) "Sombrero" bump formed by a Nd:YVO_4 laser pulse on a NiP substrate. (b) "Microdome" bump formed by a CO_2 pulse on a glass substrate. Below the photos are typical cross sections of the bumps. (a) sombrero, (b) microdome.

transiently melts the NiP, permitting competing surface tension effects to come into play.[49] The height of the central dome, either above or below the outer rim, can be manipulated by pulse energy adjustment. On glass, the focused CO_2 pulse transiently heats the material to a "softening" point at which material can flow. This flow is induced by normal thermal expansion of the glass and by compressive surface stress that exists because of chemical strengthening of the glass. Because only a small region of material is heated, the main material flow is upward, creating the dome. As with NiP, adjustment of the pulse energy permits variation of the resultant dome height. The NiP melting and the glass softening both create situations where the dome top becomes microscopically smooth. The small laser/substrate contact area and the smoothness of the processed area result in excellent start/stop wear behavior and bump durability.

At a recent 1996 disk industry show, no less than six companies were offering turnkey laser texture tools. In addition, several other companies offer lasers specifically made for disk texturing. Automated laser texture tools (LTTs) for the disk industry offer a high throughput, cassette-based operation. Cassettes of disks are loaded into an input conveyor, and processed cassettes are unloaded from the output conveyor. The procedure is highly controllable, permitting repeated creation of a preselected bump profile. The bump spacing and bump placement are likewise highly controllable. Depending upon the make and model of the automated tool, and the type of processing, process rates extend from 100 to 500 disks per hour.

In conclusion, it is clear that the laser microdome formation process is a success story. The reason for the success of this process, as with all successful laser microengineering techniques, lies in its ability to fill a niche filled by no other process in a cost-effective manner. Although techniques exist for zone texturing by mechanical methods, the zone so produced is of poor quality and has only imprecise placement at best. Similarly, lithographic methods have been used for zone texturing. However, this latter process is relatively expensive and hence not suitable for high-volume manufacturing. The microscopic resolution properties of laser light, coupled with the laser's ability to interact with material surfaces in a unique way, has made the LZT process the solution of choice for the high-end disk drive market.

5.7 A Case Study: Developing a PLD Materials-Processing Tool

5.7.1 Introduction

PLD has become an active area of materials research over the past decade. It is estimated that, to date, over 300 different materials have been deposited using this unique physical vapor deposition (PVD) process. The bulk of the work has focused on various oxide compounds such as high-temperature superconductors (HTSs). However, a significant amount of work has focused on materials that are of potential interest to aerospace engineers, such as wear- or scratch-resistant coatings and tribological coatings.[71] Such materials include diamond-like carbon (DLC), amorphic diamond, cubic boron nitride, and various thin-film lubricants such as WS_4, MoS_2, and TiC.[71] Further research is being conducted on new materials such as C_3N_4 and various other nitrides.

PLD offers many well-known advantages over other PVD techniques, and is therefore an attractive process for materials research. The advantages of PLD can be summarized as follows:

- Ease of stoichiometric transfer of complex target compositions directly to the deposited film
- Ability to deposit films in a wide variety of background gas species
- Ability to conduct ion-, neutral-, or photon-beam-assisted depositions
- Ease of target changes, permitting multilayer film growth

It is difficult to find another PVD process that offers as many features as PLD. When developing a PLD system, certain basic issues must be addressed. The following section focuses on these

issues. The discussion is directed towards aerospace applications but is also relevant to almost all other categories of applications. For the discussion, it is assumed that the reader has some degree of familiarity with the rudimentary aspects of the PLD process. An excellent review of this subject is given by Chrisey and Hubler.[71]

5.7.2 Basics of PLD System

In order to develop a viable PLD tool, the materials scientist must first answer several questions about the nature of the research that is to be conducted. These include the following basic questions.

- What materials or class of materials is to be deposited?
- What substrate materials and shapes will be used?
- What is the substrate size (dictated by the minimum device requirements)?
- What final film thickness will be required for the application?
- What is the maximum temperature the substrates can tolerate? How does this temperature compare to the melting point or crystallization temperature of the material to be deposited?
- Does the material system or device require multilayer film growth?
- What potential background gases will be required for the materials to be deposited?
- Is the process or material of interest sensitive to background gas species such as water vapor?
- Will ion-beam or photon-beam-assisted deposition be necessary to ensure growth of the proper phase?
- What deposition rates and throughput will be necessary for the application?
- What is the available budget for the PLD tool?
- Will the tool be designed and assembled in-house, or purchased as a complete system?

The last two questions should be considered carefully. A typical mistake is to add up the projected costs of all the tool components and then to expect an outside vendor to sell such a system for these costs. Usually ignored in this calculation are thousands of small items (e.g., cables, connectors, water flow switches, power distribution, safety interlocks) as well as the extensive time spent on system design, engineering, and assembly. While design can be conducted "in-house," the cost of design is not typically included in the estimate of a system price. It typically takes about 1 man-yr to properly design, procure, and assemble a working PLD system for a group well versed in deposition and vacuum technology. Therefore, depending on the overall system complexity, a realistic purchase price from a competent vendor is typically two to three times the apparent component "costs." In addition to cost, there are other ancillary issues that should be considered. These issues include the versatility of the tool handling the step that follows PLD in the materials development process, and the question of whether the system should be scaled up to include a larger deposition area, or should branch off into different or more complicated material systems and/or processes. Figure 5.15 shows a schematic of basic large-area PLD system based on a rectangular box design.

When designing a PLD system, the following items should be considered: laser, deposition chamber, substrate heater, target, BDS, pump, deposition rate monitor, and large-area PLD. A discussion of each of these items follows.

5.7.2.1 Laser

In PLD processing, it is strongly recommended that a UV excimer laser be used. Compared to IR or visible lasers, UV lasers generate fewer particulates because of the smaller (~100-nm) absorption depth. In general, higher quality films are grown when UV excimer lasers are used. In addition to the UV output from excimer lasers, the UV output (fourth harmonic) from Nd:YAG (and

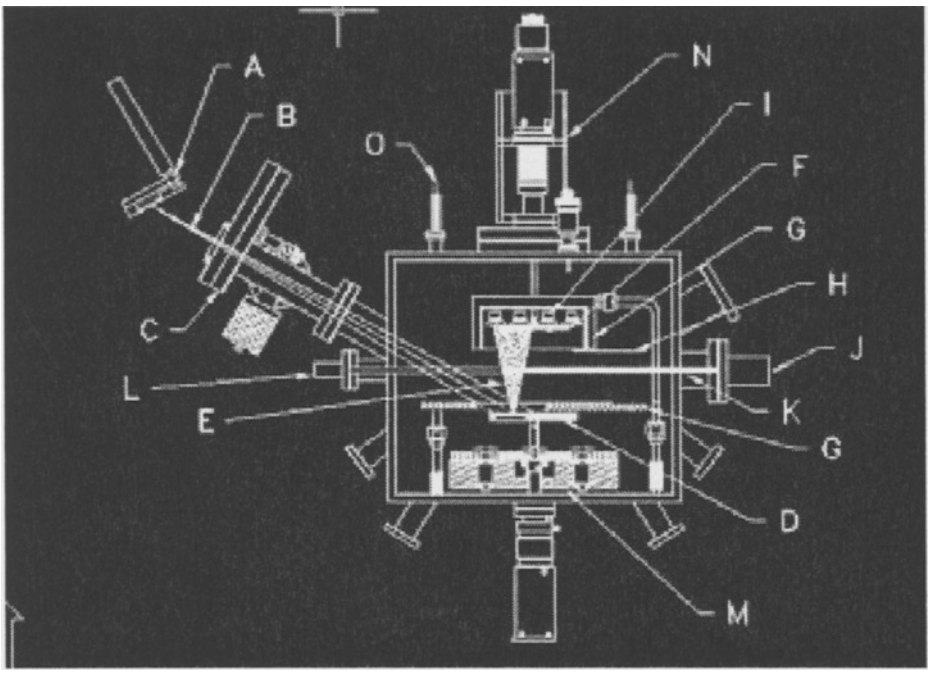


Fig. 5.15. Schematic of a large-area PLD system. A–raster mirror assembly, B–rastered laser radiation, C–Intelligent Window, D–ablation target, E–plume, F–substrate, G–water-cooled shield, H–shutter, I–heat lamp, J–hollow cathode lamp, K–HCL radiation, L–detector, M–target manipulator assembly (only one target shown), N–substrate linear/rotary feedthrough, O–current feedthrough.

other) lasers have also been used in PLD. However, current Nd:YAG lasers produce insufficient UV output powers to obtain reasonable deposition rates over useful substrate sizes. Furthermore, Nd:YAG lasers have a much higher beam quality, which is actually a drawback for PLD. The excellent Gaussian beam profile allows the YAG beam to be focused to a very small spot that yields very high fluences. Energy densities achieved with these lasers can actually melt the target, increasing particle generation tremendously. Expanding the YAG beam to reduce the fluence produces a larger spot with a very nonuniform energy density. The excimer laser, on the other hand, produces a large rectangular output, typically 1 × 2 cm. The energy density over the rectangle is usually uniform (if the laser is working well), except at the edges. Using an appropriate BDS train, the output can be focused down to a spot on the order of a few millimeters square with reasonable uniformity. In summary, excimer lasers offer the best characteristics for the PLD process, because they provide sufficient peak power (~100 MW/cm^2), usable average powers (10–100 W), and multi- or random-mode operation to obtain practical deposition rates and good uniformity. For example, an excimer laser operating at 248 nm (KrF), delivering ~500 mJ per pulse with a repetition rate of about 50 Hz or higher, is more than sufficient for the growth of most materials with reasonable deposition rates (0.5 to ~2 μm/h), over a 2- or 3-in.-diam substrate. Lasers that produce very high pulse energies (~1 J/pulse) are only useful if the ablation threshold for the material is very high. Such lasers rarely operate at the rated output power for an extensive period (irrespective of the vendor claims) and can be very problematic in the long run. Thus, care should be taken in determining the proper size of laser for the application; a 1 J/pulse laser is not necessarily the best choice. Running the laser at lower output is difficult, because the pulse-to-pulse stability is

best when the laser is operated close to the rated output. In high-fluence PLD, most of the incident energy is absorbed in the laser-induced plasma, and the yield of particulate generation tends to increase. High-fluence ablation is typically used for the deposition of DLC films (and usually at 193 nm).[72] High fluence can also be achieved with smaller lasers and the appropriate BDS. However, for several of the materials of interest to aerospace engineers, the deposition rates are typically very low. Materials such as carbon (to form DLC) and carbides used as wear-resistant or protective coatings have very low deposition rates per laser pulse. If these are the materials of interest, a laser with a high repetition rate (~150 Hz or higher) and at least 600 mJ/pulse should be seriously considered in order to obtain reasonable growth rates and throughput.

When purchasing an excimer laser, the user must consider both the physical size of the laser and other additional operational costs. Typical lasers are on the order of 4 to 6 ft long, 3 ft wide, and 2 ft high. Additional costs include the facilities (electrical power, multiple stainless steel gas lines, water cooling, and exhaust vent); supporting table (does not need to have vibration isolation); several high-purity gas regulators; and a cabinet for the halogen gas and high-purity gases. Other ancillary expenses might include an energy detector, a gas purifier, a beam attenuator, safety shields and eyeglasses, and a water chiller (because the higher power lasers and some components in the PLD system may require water cooling). All of this equipment, plus the size of the laser, should be taken into account when procuring the appropriate laboratory space for a PLD system.

In addition to costs just mentioned, there are laser maintenance expenses, which include replacement optics (for the laser and the BDS) and halogen filters. Also, every few years, the internal electrodes and preionization pins must be replaced in the excimer laser. Another issue that impacts cost is the selection of which gas to use in the laser. Argon is significantly less expensive than krypton. However, the reduced amount of energy obtained with argon, and the extra wear and tear on the laser cavity and optics that occurs at the 193-nm (ArF) wavelength, makes krypton (with the laser operating at the 248-nm [KrF] wavelength) a better choice for most PLD applications. As mentioned before, 193 nm is the best wavelength to produce high-quality, optically transparent DLC. If the user plans to use ArF (193 nm) with high repetition rates, then the optical beam path should be purged to remove oxygen. The purging should be done because 193-nm radiation produces ozone in air, which is a serious health hazard. A simple Lucite box purged with nitrogen or argon is sufficient for this purpose and will also keep all of the optics clean.

Excimer lasers also operate at a 308-nm (XeCl) wavelength. The gain at this wavelength is not as high as the gain at 248 nm, but in general the gas lifetime is longer. The 308-nm excimer wavelength is also less punishing to the laser optics but is considered more of an eye safety hazard. Long-term exposure to even low levels of 308-nm scattered light can cause glaucoma. Glaucoma does not occur as readily at 248 nm. However, the 248-nm light can cause an eye condition known as "welder's blindness," which can be treated in 24 h. It is easy to change from ArF (193 nm) to KrF (248 nm) and back again with little downtime. Changing the laser over from a chlorine-based system to an fluorine-based system is not recommended.

Although excimer lasers are valuable for PLD sources, they have only four distinct wavelengths of operation: 351 nm XeF, 308 nm XeCl, 248 nm KrF, and 193 nm ArF. The PLD technique would benefit if the laser wavelength could be altered to match the peak absorbance in the target material. This requires a widely tunable laser with high power. The FEL can be designed to be widely tunable. The Department of Energy (DOE) Jefferson Laboratory FEL is in the commissioning phase to provide tunable kilowatt-pulsed laser power in the IR. The laser is designed to be upgraded to produce tunable laser power in the UV. This facility will provide tremendous insight into the interaction of light with materials. The ability to tune the laser radiation over a

wide spectrum will likely allow the materials scientist to more easily couple the radiation to a specific target material. The strength of the absorption will be strongly dependent on the chemistry (bond nature) of the target. By tuning the laser to individual target chemistry, the user may be able to break specific bonds or to couple energy to the target material in various modes, such as excited electronic or vibrational states. This energy can then be transferred to the growing film, enhancing various film properties such as crystallinity.

Two examples of strong coupling to specific target material as a function of laser wavelength follow. As a first example, the optical quality of DLC films grown from graphite targets is clearly best at 193 nm. The optical quality is best at this wavelength because the more energetic photons obtained with 193-nm radiation break almost all of the carbon-carbon bonds. The result is an energetic plume consisting mostly of individual carbon atoms, with very few dimer molecules.[72] The energetic plume increases the likelihood that sp^3 bonds will form when the carbon condenses at the substrate surface. Films grown with KrF primarily have dimers in the plume, and yield films predominantly made up of the stronger sp^2 bonds. On the other hand, the basic electrical properties of yttrium barium copper oxide (YBCO) (the high-temperature superconducting compound) films are usually very similar, regardless of the laser wavelength used (193 or 248 nm). In this case, the more energetic radiation at 193 nm does not play a significant role (other than a slight reduction in particulates). As a second example, in unpublished work, metallic films made from metal-carbonyl targets were deposited using two different wavelengths. When the target is ablated, it ejects carbonyl compounds that decompose upon impinging on the heated substrate. The volatile carbonyl radicals desorb, leaving a metallic film. Using radiation from 193 nm with fluences of about 1.0 J/cm^2, the film growth rate was on the order of 0.5 μm/h at a 30-Hz pulse rate. Using radiation at 248 nm, just three laser pulses completely covered the entire vacuum chamber walls with several microns of the carbonyl material and covered the hot zone with a thick metallic layer, destroying the substrate heater. While this result was totally unexpected (and the system had to be cleaned and rebuilt, which took several days), it clearly demonstrates that the right combination of target material, coupled with the proper laser wavelength, can increase deposition rates. Also, the right combination can significantly alter the electronic properties of the ablation plume and thus, of the deposited films. While both the excimer and Nd:YAG lasers provide several discrete working wavelengths, they do not offer the capability to continuously vary the wavelength over the wide range that is available with the FEL. Future upgrades of the FEL will include high-power operation in the 190–300-nm range. Another interesting feature of the FEL is its very short pulse length. A short pulse length may significantly reduce the amount of particles that are generated in the PLD process.[71] It is expected that radiation from the FEL will become an active area of research for PLD material scientists in the near future.

5.7.2.2 Deposition Chamber

For the vacuum chamber design, attention should be primarily given to the substrate holder, the substrate heater (if one is required), and the target or targets (if multilayers are required). Several vacuum chamber styles have been used for PLD, including a simple four-way cross, a sphere, and a bell jar, all with extra ports. An alternative design includes a rectangular box chamber with a large hinged access door and several ports. Figure 5.15 shows a schematic of a rectangular box chamber design. Every chamber should be designed around the largest substrate diameter to be used. Once the substrate size is selected (e.g., 50 or 200 mm), the target shape and size and the target-to-substrate spacing, known as throw distance, can be determined. In general, the larger the substrate, the larger the target and the throw distance required. While there are no hard-and-fast rules, a reasonable choice is for the throw distance to be set at least the diameter of the substrate

size if uniform film thickness (±5%) is desired. While some systems allow the throw distance to be varied, such variation is typically not necessary. For most material systems, the same film properties can usually be obtained at different throws by simply changing the background gas pressure, keeping the product of pressure and distance relatively constant. The target diameter should be at least the size of the substrate. Both target and substrate can be mounted in almost any direction; however, there are more advantageous orientations depending on substrate type. Horizontal mounting allows the target (or targets) to be simply held by gravity and the substrate to be suspended at its edges without the need for clamping to a back-plate. Horizontal mounting is important when delicate substrates are used or when the process is sensitive to substrate temperature, or if the quality or integrity of the back side of the substrate is relevant to the application. Regardless of the mounting orientation, if a uniform film thickness is required over substrates larger than 50 mm in diameter, both the substrate and target must be rotated in conjunction with programmable laser beam rastering (discussed below).

The chamber should include ports for substrate manipulation (rotation and translation, if desired); for target rotation (and indexing, if multiple targets are used); for the laser beam entrance; for vacuum pumps and gauging; and for substrate/target transfers (if a hinged door is not used). In addition, the chamber should have several auxiliary ports. These auxiliary ports might include view ports to see the substrate and target during deposition and to see substrate heater components, including current feedthroughs, thermocouples, and water-cooling connections. View ports should use glass that strongly absorbs the laser radiation, or a safety hazard may arise. For excimer lasers, a simple 3-mm-thick Lucite disk placed over a standard low-profile view port is adequate to absorb radiation. In addition to view ports, it is helpful to have a port that controls a substrate shutter and is used for target precleaning prior to deposition. Depending on the ultimate system goals, ports might also be considered for plume diagnostics such as atomic absorption and/or emission; ellipsometry or other spectroscopies; residual gas analysis; and gas processing sources such as atom, ion, or sputter deposition sources, process gas bleed, and vent valve. Port flanges can be knife-edge or O-ring style. It is usually well worth the time to lay out all of the ports on the chamber with each potential component drawn in, prior to fabrication of the chamber. This layout will help ensure that the design can be assembled as intended.

5.7.2.3 Substrate Heater

Careful attention should also be given to substrate heating when designing a PLD tool. Maximum substrate temperatures, along with the type of substrate (as defined by its absorptivity and emissivity) and background gas, are key ingredients in determining the type of heaters to be used. The use of oxygen as a background gas significantly reduces the types of heaters that can be used, especially if the substrate temperature is to reach above ~600°C. Also, if oxygen or other reactive gases are to be used, care must be taken to properly select all the other materials that will become heated in the presence of the gas. Materials such as molybdenum or tungsten should not be used for heating elements, shields, substrate holders, or substrate clips, as they are easily ignited and very dangerous when heated in the presence of oxygen. The design of substrate heaters is considered somewhat of an art and is clearly beyond the scope of this chapter. However, the basic considerations for PLD applications are presented.

One of the first considerations in the design of a substrate heater is whether or not the substrate can be bonded to a heated backing plate with some thermally conductive material such as silver paint or indium. Clamping the substrate to a back plate does not ensure good temperature uniformity and is not recommended for most applications. Although thermal bonding materials (e.g., paint and pastes) provide the easiest way to achieve a given substrate temperature, they have several drawbacks.

- Thermal bonding will be a problem if the back side of the substrate will be used for subsequent film growth (for instance, double-sided YBCO film growth on $LaAlO_3$ substrates), or if the substrates are delicate (such as CdTe or HgCdTe).
- Thermal bonding will be a problem if photolithography will be employed. After deposition, the thermal bonding agent will adhere to the substrate back side and will be difficult or impossible to remove without damaging the deposited film.
- The thermal paste will also outgas a large amount of organic residue during the pumpdown and initial substrate heat cycle.
- If indium is used, it will form an oxide at elevated temperatures. InO has a high vapor pressure and thus will be a possible source of film contamination.
- When bonding large substrates greater than 1 in., it is difficult to ensure uniform bonding after heating up the substrate because the substrate may bow slightly, producing local cold spots.

If bonding is not going to be employed for substrate heating, then more stress is placed on the heating elements to achieve the desired substrate temperature. In this case, however, the substrate back side remains clean, making it easy to either deposit a back-side film or to do postdeposition processing. When not using thermal bonding agents, it is best to hold the substrate only at its edges during the heating process.

Several types of heaters have been used for the PLD process, including projection lamps and resistive heaters based on magnesium oxide (MgO)-sheathed Inconel conductors, nichrome, Kanthal, and platinum. Other combinations of materials have been used for heaters, including Si, SiC, and graphite encapsulated in boron nitride. The latter heating material is very useful for several applications, but if the temperature (not substrate temperature) exceeds ~800°C, the boron nitride is etched by oxygen, if oxygen is used as a process gas. If oxygen is not going to be used, heating elements made from carbon, molybdenum, tantalum, or tungsten wires may be acceptable. When designing a heater, careful consideration must be given to all the materials being used to ensure compatibility with any of the background or process gases and with the ultimate temperature to be reached. Also, care must be taken to make sure that the heating elements or other hot components do not decompose in the desired background gas.

One important figure of merit for any substrate heater is the temperature uniformity attainable across the substrate surface. In order to minimize temperature gradients and the amount of power necessary to heat the substrate to a desired temperature, heating elements should extend out past the edges of the substrate. Several reflecting shields should also be placed above and around the elements and substrate if possible, especially if the substrate temperature is to go beyond ~400°C. These shields will help improve the temperature uniformity and will reduce the amount of power required to heat the substrate to a given temperature. In addition, the shields will reduce the amount of radiation reaching the chamber walls.

Measuring the temperature of a heated substrate can also be difficult, depending on the heater design and substrate materials. For instance, transparent substrates such as sapphire and quartz do not readily absorb IR radiation. Furthermore, when a pyrometer is used to read the temperature of a transparent substrate, the signal from the substrate is obscured by the radiation from whatever is behind the substrate. The result can lead to an incorrect reading. A small thermocouple placed on the substrate itself will not indicate the proper temperature, because the thermocouple will absorb more radiation than the substrate, again indicating the wrong temperature. Thus, measurement of substrate temperature can be very tricky, and care must be taken to obtain the correct value. Sometimes the substrate temperature will not be known, but reproducible growth can be obtained by using a free-floating thermocouple that monitors the radiation environment. The thermocouple should be used as the input to a programmable temperature controller. The temperature

controller can then be used to run predetermined thermal cycles and to hold the temperature constant during deposition.

Another issue when designing a substrate heater is the total amount of energy that will be radiated into the deposition chamber and its effect on other internal components. A properly designed heater for a 3 in. diam substrate needs about 1.5 kW of power to heat a transparent substrate, such as sapphire, to a temperature of 750°C, if no thermal paste is used to bond the substrate to a heated block. The radiation from the heater, if not properly dealt with, can cause several problems. To avoid the problems, properly designed water-cooled plates should be placed behind the heating elements and any of the reflecting shields. This removes excess heat, which otherwise will heat the chamber walls and cause a burn hazard. Furthermore, during processing, hot walls liberate water vapor (usually the dominant source of background gas in any clean and unbaked vacuum system), which can severely influence the properties of the deposited films. A hot substrate (or heater) located a few inches above the target surface can increase the temperature of the target by several hundred degrees. This temperature increase can cause the targets to outgas considerably and may have other, more deleterious effects on film growth. For example, if the target becomes sufficiently hot from the thermal radiation, the absorbed light from the incident laser beam may be sufficient to locally melt the target surface. Local melting greatly increases the ejected particulate density and produces changes in the deposition rate and film properties. Therefore, water-cooled shields should be placed above the target to minimize the thermal radiation, as shown schematically in Figure 5.15. These shields should include slots to expose the necessary area of the target to the laser beam and to allow the plume to impinge on the substrate. Finally, when designing a substrate heater, the following safety issue should be considered: The heater should not be allowed to operate above a few torr. Thus, if the chamber is opened, an operator's hand cannot accidentally touch one of the electrical feedthroughs, causing electric shock.

5.7.2.4 Target

For simple single-layer film growth, only a simple target holder and rotary feedthrough are required. Target rotation is necessary as a minimum; otherwise, the target morphology changes very quickly under laser irradiation, greatly altering the plume shape and direction. Several types of rotary feedthroughs are available for target rotation. Rotation speeds between 6 and 30 rpm are more than adequate for most applications. If the application requires multilayer film growth of different materials, then a multitarget manipulator is required. A manipulator allows the user to readily change, either via computer or manual control, the active (ablation) target without breaking vacuum. Typical manipulator configurations can include from three to six targets of a given size. Figure 5.16 shows a photograph of a multitarget manipulator that holds four 2 in. diam targets. This particular manipulator is mounted on a large 12 in. diam mother flange and has a linear translation stage that provides up to 4 in. of z-axis motion. A programmable stepper motor is supplied, which allows the targets to be quickly indexed into the ablation position in any order desired. This manipulator is based on a dual-axis, magnetically coupled rotary feedthrough. One axis provides continuous target rotation up to 35 rpm; the other axis provides target indexing. The magnetically coupled feedthrough is used because it provides longer life than a welded-bellows feedthrough. The manipulator in Figure 5.16 also has a water-cooled shield that sits directly above the targets. The shield has an open slot to allow rastering of the laser beam over the active target. Note that the shield design protects the unused targets from backscattered vapor, which minimizes cross contamination, and the water-cooling keeps the targets, gears, and bearings isolated from thermal radiation.

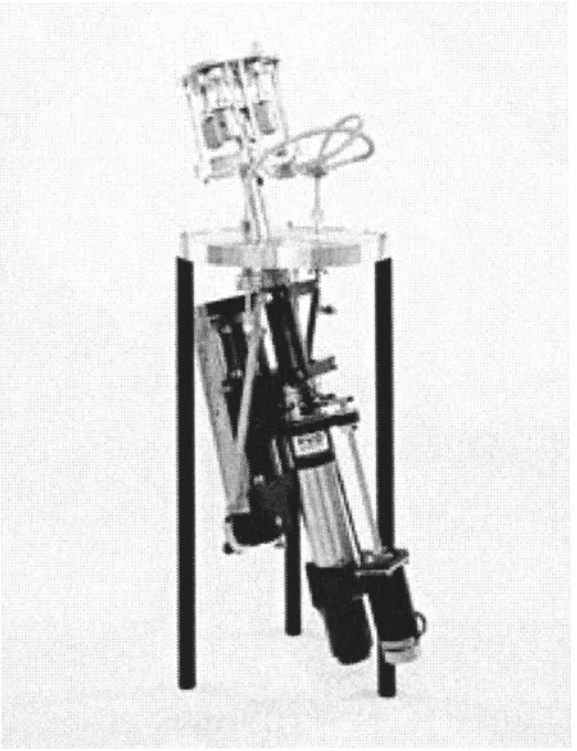


Fig. 5.16. A four-position PLD target manipulator. Photo Courtesy of Epion Corporation.

Almost any material can be deposited by PLD. It is often said that if you can make a solid target of any size and shape, then you can deposit a film by PLD. One interesting example comes to mind. At a Materials Research Society meeting in Boston, a researcher showed an x-ray diffraction θ-2θ scan obtained from a PLD film grown from a small rock taken from the Berlin wall! This example also highlights the fact that the purity of the target is very important. If the element (desired or not) is in the target, it is most likely going to end up in the film. Thus, if the material being deposited is strongly affected by impurities, care should be taken to purchase a target with very high purity. In addition, the targets should be handled with clean gloves.

The target properties (size and density) can play a role in the quality of the films that are produced. Typically, round targets, from one to several inches in diameter, are used. Target thickness can vary from below 0.1 to 0.25 in. or greater. In general, larger diameter targets are more effective than smaller diameter targets, especially if large substrates are to be coated or if very thick films are desired. Larger targets are effective because a large amount of material will be removed from the target. If a large target is used in conjunction with laser beam rastering, the target surface will remain flat, which is preferable. Conversely, if small-diameter targets are used without rastering, the target surface will become trenched as the target is rotated. As a result, the ablation plume angle will be grossly altered, changing the deposition rate at the substrate.[73] Also, the trenching of the target surface will alter the laser beam spot size and thus the laser fluence. Laser rastering using a programmable mirror, discussed in the section below, eliminates trenching of the target surface and provides for more uniform and reproducible deposition.

When obtaining targets for the PLD process, density is also a consideration. It is widely believed that dense targets produce films with the least amount of particulates. This belief is often,

though not always, borne out in reality. Though a dense target is desirable, it is also important that the target be homogeneous with very small grains. However, in order to obtain high density, the target is sintered for a long time, but this also yields larger grain sizes. Therefore, a compromise must be struck between target density and the grain size within the target. A low target density means it is porous but has a small grain size. This porosity will result in outgassing, and for large targets, a considerable amount of pump-out time might be needed before the chamber reaches a base pressure. For some materials, target densities in the 80% range have yielded very good film properties. In other materials, like the carbide targets, vendors will supply either hot-pressed or CVD-prepared targets. The hot-pressed targets tend to generate films with more particulates than CVD-prepared targets. However, CVD carbide targets are very difficult and expensive to obtain. In some cases, powder of the proper material can be ablated. Also, liquid Ga has been used as a target to form GaN films in various background gases. For most materials, the target composition should be exactly the same as the composition desired in the film. However, for certain materials, such as those containing high vapor pressure elements like Pb or Li, nonstoichiometric targets should be considered. The reason is that some of the elements will re-evaporate from the heated substrate surface before being oxidized. In some cases, such as the oxide films, the target can be nonstoichiometric. For instance, MgO and SiO_2 may be formed in an oxygen background using either a simple Mg or SiO target, respectively. Stoichiometric targets would be difficult to obtain directly because of the poor absorption in the UV of these materials.

The morphology of the target surface changes as the ablation process is carried out.[73] For systems that use a fixed-position (nonrastered) laser beam, the target needs to be resurfaced after each run, or after every few runs. Target resurfacing wastes material and raises the issue of contamination of subsequent films.

5.7.2.5 BDS

While the laser has already been discussed, consideration must also be given to the entire BDS that delivers the laser radiation to the ablation target. Two basic approaches to the BDS can be used: focusing and imaging. Usually, fluences between 1 and 3 J/cm^2 are sufficient for most materials, with the actual fluence hitting the target being between 50 and ~500 mJ. Again, because several of the materials of interest to aerospace engineers have high ablation thresholds and low deposition rates, fluences approaching 5 J/cm^2 may be necessary. In general, low fluence and high laser repetition rates are better for producing films than high fluence and low repetition rates. In a simple BDS, the basic elements are an aperture used to define and remove the nonuniform portions of the excimer beam, a focus lens (spherical or cylindrical), and a chamber entrance window. The excimer laser beam is usually 1 × 2 cm in size. Thus, optics of at least 50 mm in diameter should be used. Anti-reflective (AR) coatings are useful for minimizing the laser energy loss at the lens and/or window surfaces. The AR coatings are wavelength specific and will be damaged by radiation at other wavelengths. A more complex BDS can be employed, including multiple focusing lenses, raster mirrors, and beam homogenizers. Laser beam homogenizers are usually not necessary for the PLD process, especially if the laser is operating properly. Raster mirrors are very useful for scanning the laser beam over a large diameter target.[73] Rastering the beam greatly improves the uniformity of film properties such as film thickness and composition. Furthermore, by rastering the laser beam over a large target, the target surface morphology does not significantly change during time. The result is reproducible film growth without the need for resurfacing of the target after each deposition.

Care should be taken to properly deal with reflections that occur at all of the BDS surfaces. These reflections can cause damage to optical components and can become a safety issue. Thus,

it is wise to enclose the BDS in an appropriate box that will absorb the stray radiation. For excimer lasers, a Lucite box is strongly absorbing and quite adequate for the job.

During deposition, the ablated products typically coat the entrance window of the chamber.[74] The amount of the coating depends on several factors, including typical operation pressure, the distance from the target to window, and the angle that the laser makes with the target surface. This coating reduces the laser fluence that is incident on the ablation target. The consequence of this reduction is a change in the deposition rate and the energetics of the ablation process. This change potentially results in nonuniform deposition and in process variability during the film growth. Furthermore, it is well known that changes in fluence affect the as-deposited stress in PLD films. This deleterious effect can be somewhat minimized by the injection of a "curtain" of process gas over the entrance window and by the intelligent choice of the incident angle (e.g., at least 45 deg, or smaller than 30 deg relative to the surface normal) that the laser beam makes with respect to the target surface. Figure 5.17 shows a PLD product, called the Intelligent Window, that not only helps to minimize coating on the window but also provides a direct measure of the energy that actually enters the deposition chamber. For this window, a large transparent disk is housed inside a pair of vacuum flanges. The laser radiation enters the Intelligent Window through a high-quality AR-coated window. The backscattered material is deposited onto the transparent disk over a small area defined by an internal aperture. When this area has become coated with vapor, the disk can be easily rotated, exposing a clean surface via an external feedthrough. A small port is included to bleed the process gas into the area around the disk and aperture, thereby raising the local gas pressure, which also helps keep the disk clean. When the disk becomes fully coated, it can be easily removed and replaced with another disk. The disks can be polished and reused multiple times. Another important feature of the Intelligent Window is that it allows the user to monitor the energy that enters the chamber just prior to deposition of the film. This capability enables the user to achieve reproducible deposition rates and overall film quality. Monitoring the energy is preferable to relying on the laser's energy meter. This meter does not provide a good measure of the energy hitting the target, for a variety of reasons.

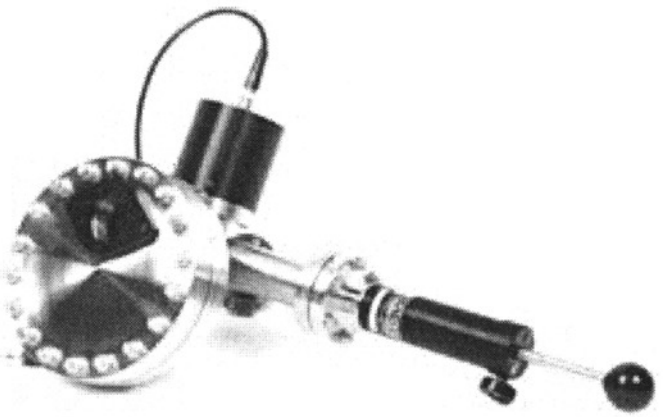


Fig. 5.17. Intelligent Window, which keeps the vacuum chamber window clean for extended periods of time and allows the user to monitor the energy that enters the deposition chamber. Photo courtesy of Epion Corporation.

- The laser beam is multimoded, and the beam divergence depends on factors such as gas fill quality and laser optics. As the beam divergence changes, the energy hitting the target can change considerably (by as much as 25%), depending on other aspects of the overall optical train.
- Quartz optics are well known to produce so-called color centers under intense radiation in the UV. The optics produce a red fluorescence under exposure when a sufficient number of color centers have been produced. When the color centers are activated, the optics absorb a significant amount of the incident energy, significantly reducing what hits the ablation target.
- Losses also occur as the coatings and reflecting surfaces of mirrors and lenses slowly degrade over time because of UV laser exposure.

Thus, the ability to monitor the energy that actually enters the chamber is the only way to know what is really incident on the target. Because many of the film properties depend strongly on laser fluence, the ability to monitor this energy is key to reproducible film growth results.

5.7.2.6 Pump

Several types of vacuum pumps–such as oil diffusion, cryo-, and turbomolecular pumps supported by a rough-pump–are suitable for the PLD process. Ion pumps are not considered suitable unless one is using a load-locked system and is not planning on using any background process gas during the deposition process. Oil diffusion pumps are not the top recommendation because they can be a source of oil contamination in the deposited films. Turbomolecular pumps offer the best option for most PLD applications. Note that these pumps should not be placed on the bottom of the chamber, where they may be seriously damaged by anything that falls. Turbomolecular pumps come in several sizes and varieties, including a molecular drag style. The drag pumps provide very high gas throughput but reduce the overall compression ratio of the pump. A high compression ratio, however, is not needed for the PLD process. The selection of pump size, measured in liters per second, should be based on the chamber size and the required speed for pumpdown to the necessary base pressure. A properly selected pump should achieve an initial pressure below 1×10^{-6} torr within 2 h of pumping and a base pressure in the low to mid 10^{-7} torr range. Because some targets may outgas considerably, there may be a limit to the base pressure that can be obtained. Appropriate pump sizes are between 100 and 1000 l/s, depending on the overall chamber size. At the base pressure, it is likely that the dominant background gas species will be water vapor. If the materials to be deposited are sensitive to water vapor, then alternative pumping approaches need to be considered, or the chamber walls must be baked out to remove water. The latter approach is not particularly convenient, because it is usually takes time, thus limiting the throughput of the PLD tool. An alternative approach is to incorporate, with added cost, a load-lock facility that minimizes the exposure of the main deposition chamber to atmosphere. Adequate consideration should also be given to the selection of a rough-pump. While dry pumps offer oil-free rough pumping, they are expensive and not needed if the turbomolecular pump is handled properly. A standard rotary-vane mechanical rough-pump is more than adequate for the job. However, if oxygen is to be pumped, then consideration must also be given to the pump oil used (e.g., Fomblin oil or the equivalent), because standard hydrocarbon-based oils become highly explosive with sufficient oxygen entrapment. Similar considerations as those just mentioned should be taken into account if ozone or another toxic gas is being pumped using a cryopump.

In the PLD process, vacuum gauging is necessary, not only to measure the base pressure, but also to monitor the pressure of added background gases. Vacuum gauging should include an ion gauge to determine the chamber base pressure, and either a thermocouple, Convectron, or Pirani gauge to monitor the pressure during the initial stages of the pump-down cycle. A capacitance

manometer capable of measuring from about 1×10^{-4} torr to 200 mtorr is recommended for monitoring the pressure during deposition. The manometer should be located so that it actually samples the pressure in close proximity to the substrate. Monitoring the pressure at the chamber wall is usually inadequate, especially if heated substrates are used, because the pressure near the substrate is strongly dependent on the substrate temperature.

During deposition, a background gas is typically used to either thermalize the plume or to improve the stoichiometry of the gas species in the film. For thermalization, an inert gas such as Ar can be used. It is important to control the background gas pressure for several reasons. First, as the background gas pressure is increased, the amount of gas-phase scattering that occurs in the plume also increases. The backscattering of the atomic and molecular species in the plume is considerably higher than the backscattering of the small particulates generated in the plume. For high gas pressures, the atomic species are scattered away from the substrate. This scattering results in films with poor surface morphology. Thus, the deposition process should be run at the lowest pressure that is compatible with obtaining the other properties desired from the material, if film morphology is also important for the application. Second, it is well known that the background gas pressure plays a significant role in the stress in laser-deposited films. At very low pressures, large compressive stresses are usually generated. As the pressure is increased, the magnitude of compressive stress can be reduced. In some cases, tensile stress can be generated, depending on the film and substrate materials.[75] Stress is usually an issue in very thick films, or in films that will be used for protective or tribological coatings.

In order to control the background gas pressure, several alternate approaches can be used instead of the usual method. In the usual method, gas is bled into the vacuum system using either a needle valve or a mass flow control valve. The pressure is then adjusted by either throttling the gate valve or by varying the speed of the turbomolecular pump. In one alternate approach, a secondary valve with a much smaller conductance is used to allow the gas to be bled from the chamber. In another approach for more advanced PLD tools, the capacitance manometer is used in conjunction with a small stepper-motor-controlled bleed valve. The manometer and bleed valve are employed in a closed-loop feedback system to accurately control the pressure during deposition. Usually, it is best to bleed the process gas into the chamber far from the deposition region in order to provide a more static gas environment around the substrate. As mentioned above, the best place to bleed gas into the system is by the laser entrance window, which also helps the window stay clean. Pointing the gas nozzle directly at the substrate is not recommended, because it will produce a dynamic flow environment around the substrate, resulting in nonuniform film properties.

5.7.2.7 Deposition Rate Monitor

For most applications, the film thickness needs to be well defined and reproducible. Thus, the rate of film growth, and the final film thickness, need to be monitored. There are several types of deposition rate monitors that work well for most PVD processes. The most well-known monitor is the quartz crystal microbalance (QCM). At first glance, it may appear that the QCM is ideal for the PLD process. However, this is not the case, for many different reasons. First, if heated substrates are used, the thermal radiation is usually sufficient to cause the QCM to become unstable. Second, the PLD process produces a highly forward-directed plume. In order to achieve any uniformity over reasonable substrate sizes, laser beam rastering is employed, as previously discussed. This rastering produces a dynamic tooling factor problem for the QCM and makes any real thickness measurement difficult to interpret. Third, the PLD process tends to produce films with a high amount of intrinsic stress. Such stress is sufficient to cause the QCM to stop oscillations or to become highly nonlinear. Therefore, instead of the QCM, ellipsometry or optical

transmission techniques may be used to monitor the deposition rate. These techniques are usually difficult to implement as an *in situ* process monitor.

An alternative approach to deposition rate monitoring for the PLD process is that of atomic absorption (AA).[74] An AA monitor is depicted in Fig. 5.15 as items J (a hollow cathode lamp [HCL]), K (HCL collimated radiation), and L (a detector). In this figure, the monochromatic light beam (K) produced by a hollow cathode lamp (L) (J), selected specifically for one of the materials in the target, passes through the chamber, intercepts the ablation plume, and hits the detector (L). The amount of HCL light absorbed by the specific species within the plume is then a measure of the net flux to the substrate surface. Even when the laser is rastered over the target, the HCL beam intercepts the same volume of the ablation plume. With proper integration techniques and careful calibration, this system can be turned into a useful rate monitor. Such systems are currently under development and are expected to be on the market in the near future.

A wide variety of HCLs are available for most materials of interest to the aerospace engineer, except for carbon. However, a variety of tunable lasers are now coming on the market that may replace the HCL for carbon and for other applications. Thus, it would be wise to include ports for AA in any system that is contemplated.

5.7.2.8 Large-Area PLD

Initially, there was a lot of skepticism that PLD could be scaled to substrates much larger than approximately 1 in. This skepticism was held because of the nonuniform and highly directional nature of the PLD plume.[71] Most applications require the various physical, electrical, optical, and tribological properties of the film to be uniform over much larger areas. Indeed, with the proper techniques previously discussed, the PLD *is* readily scalable to large substrates.[73,76] Using large-diameter rotating targets, in conjunction with programmable laser beam rastering with rotating substrates, excellent uniformity can be achieved over substrates up to 8 in. in diameter. Film thickness uniformity of better than ±4% has been achieved for Y_2O_3 films deposited over 200-mm (8 in.) diam substrates.[76] Furthermore, the compositional uniformity obtained over a 150-mm-diam substrate from a YBCO target was ±1.48%, ±0.17%, and ± 0.36% for the Y, Ba, and Cu species, respectively.[76] Uniform electrical properties such as the T_c (critical temperature) and J_c (critical current density) for YBCO over 75-mm-diam $LaAlO_3$ substrates has also been demonstrated.[71] More recent (unpublished) results for the critical temperature of HTS films indicate that *in-situ* YBCO can be deposited with very high quality over 125 mm (5 in.) diam substrate areas with T_c's as high as 89.6 K and with variations of ±0.5 K. It is expected that with proper engineering, PLD can be scaled to much larger sizes, if needed.

PLD is still an emerging technology. At present, applications for PLD-deposited films have not demanded production-style machines. However, several applications for small-scale production PLD are emerging, including the HTS market. Also, complex films deposited over a 1 m length are being seriously considered for a roll-to-roll application. Thus, it is expected that PLD will soon be a standard production deposition technique for a variety of otherwise hard-to-deposit materials.

5.7.3 The PLD System

A properly designed PLD system will offer an engineer or scientist many years of research and development capabilities without the need for major modification or upgrades. Several vendors that offer commercially available systems are listed in Table 5.6. These vendors can also provide components for "in-house" systems. Therefore, the user must decide whether to purchase a complete system or components or whether to assemble the unit in-house. Care should be taken when purchasing a PLD system from a vendor. Vendors who use their own systems to deposit material

Table 5.6. Vendors of PLD Systems

Vendor	Location
DCA Instruments, Inc.	Woburn, Massachusetts
Epion Corporation	Bedford, Massachusetts
Kurt J. Lesker Co.	Clairton, Pennsylvania
Neocera, Inc.	College Park, Maryland
Surface Equipment, Ltd.	Huckelhoven, Germany
Thermionics, Inc.	Hayward, California

on a routine basis are much more likely to deliver a working end-product. While complete systems assembled by vendors may be more expensive, they should provide the user with an operational system in a comparatively short time. Thus, the user will be able to focus on materials development rather than deposition system design.

Figures 5.18 and 5.19 display photographs of a complete PLD system based on a rectangular box design. This load-lock compatible PLD system can handle up to 3 in. diam substrates and can provide very uniform thin films. The system includes a three-position large-diameter target manipulator, an Intelligent Window, a turbomolecular pump, and a rastered optical train. The substrate heater uses IR heat lamps and can heat transparent substrates to temperatures in excess of 800°C in oxygen. In Fig. 5.18, the excimer beam is seen traveling through the optical train on the

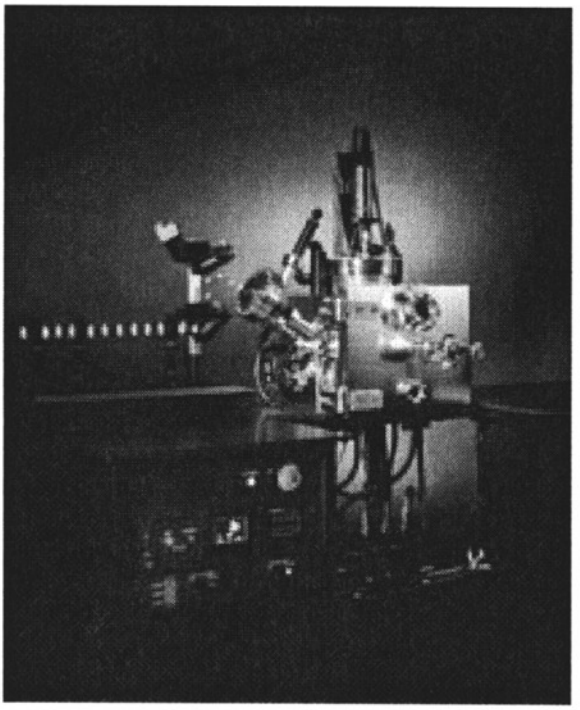


Fig. 5.18. A PLD system for coating 3-in.-diam substrates. Photo courtesy of Epion Corporation.

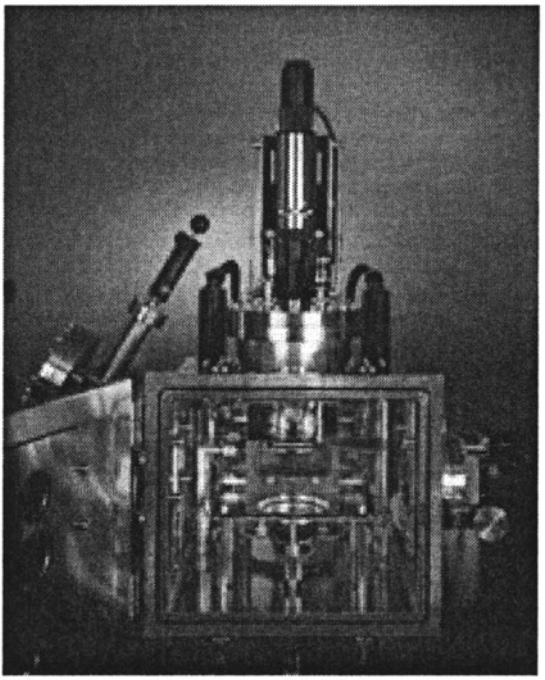


Fig. 5.19. Interior of the PLD system shown in Fig. 5.18 with door ajar. Photo courtesy of Epion Corporation.

left, where it enters the chamber through the Intelligent Window. Also shown in this figure is the linear/rotary feedthrough, on the top of the chamber, for substrate manipulation; the temperature control unit; an emergency off (EMO) button; and the chamber frame and support table. A water manifold is seen below the chamber, which provides several water lines that enter the bottom of the vacuum system. The large door on the chamber can be opened upon venting, as seen in Fig. 5.19, which allows the user to quickly change targets and substrates without removing any of the flange assemblies. Also shown in this figure is the target assembly (below the flat water-cooled plate), the water-cooled heater boxes, the water lines, the shutter, and the substrate rotation stage.

5.8 Conclusions and Brief Overview of Trends

It is clear that laser material processing tools will play a significant role in the future development of new materials and devices. This prediction is made because, by its very nature, laser processing is a "dry," environmentally friendly technique, which under controlled conditions can process materials with minimum waste. We believe that in the near term, laser processing will more likely be applied "at the back end" of a manufacturing process–used for customizing items, enabling precision modifications, and performing repairs. However, the trend for "front-end" laser-processing operations will grow as the cost and reliability of lasers are improved. There are already applications where laser processing in the front end of a manufacturing line is close to being cost effective (e.g., in surface texturing). The Department of Energy's (DOE's) JLAB FEL group is attempting to tip the scales by demonstrating that laser photons can be made for less than 10 cents per kJ of light. Several major U. S. corporations (e.g., Dupont, 3M, IBM, and Northrop Grumman) have formed an industrial consortium (Laser Processing Consortium) in support of this FEL demonstration.

We see laser-based processing techniques especially growing in microengineering applications. We forecast the development of laser tools (e.g., for soldering, phase hardening, sintering) that will enable the manufacturing of microfabricated devices on a desktop[77] to larger laser systems that are designed for volume manufacturing applications (e.g., laser marking, fabrication of masters, selective removal etching, surface texturing).[78] Furthermore, with the desire in the aerospace community to fabricate better or more robust materials (e.g., diamond and SiC for hypersonic vehicles), the use of laser-coating tools like PLD will be more in demand.

The laser vendors and the laser-processing community are also taking steps to provide a better product and more reliable processes. Lasers will continue to get more reliable as they are implemented in manufacturing. A strong technological driver that will increase laser reliability is the decision of the microelectronics industry to use lasers in the lithography of submicron circuits. As a consequence, developers of laser-processing tools are now implementing user-friendly software and additional controls to increase reliability and reduce process variability. In the horizon are other laser sources, currently under development, that will accelerate laser processing. For example, single-mode fiber lasers have achieved power levels approaching 1 W, and one can conceive of bundling many fibers together for higher power. Furthermore, there continues to be an increase in the power of diode-pumped solid-state lasers. These systems are approaching kilowatt power levels, and their total footprint in size is diminishing. Also expected in the market are femtosecond lasers and vacuum UV (VUV) (157-nm) excimer-laser-based material processing tools.

Finally, in the aerospace community, and especially for space applications, there is a growing trend to design systems, and for that matter, satellites, as complete integrated packaged units.[79] To implement this concept, the aerospace community will need to borrow heavily from the microelectronics industry. This means more automation in the manufacturing segment, use of clean processes, more use of CADCAM software, reliability assessments based on statistical measurements, and the general miniaturization and integration of common systems. Laser-based tools are poised to make significant impact to this development.

5.9 References

1. R. S. Muller, R. T. Howe, S. D. Senturia, R. L. Smith, and R. M. White, *Microsensors* (IEEE Press, NY, 1991); J. Bryzek, "MEMS: A Closer Look," *Sensors*, 5 (July 1996); *Microengineering Technology for Space Systems*, Monograph 97-02, edited by H. Helvajian (The Aerospace Press, Los Angeles, 1997).
2. H. Helvajian, "Laser Material Processing: A Multifunctional in-situ Processing Tool for Microinstrument Development," in *Microengineering Technology for Space Systems*, Monograph 97-02, edited by H. Helvajian (The Aerospace Press, Los Angeles, 1997), p. 67 and references therein.
3. R. E. Russo, D. B. Geohegan, R. F. Haglund, Jr., and K Murakami, editors, *Laser Ablation* (Elsevier, New York, 1997); J. F. Ready, *Effects of High-Power Laser Radiation* (Academic Press, NY, 1971); J. Y. Tsao and D. J. Ehrlich, editors, *Laser Microfabrication –Thin Film Processes and Lithography* (Academic Press, NY, 1989); R. Haglund, Jr., and R. Kelly, "Electronic Processes in Sputtering by Laser Beams," in *Fundamental Processes in Sputtering of Atoms and Molecules*, edited by P. Sigmund (Munksgaard, Copenhagen, 1993); L. D. Laude, D. Bäuerle, and M. Wautelet, editors, *Interfaces Under Laser Irradiation*, NATO ASI series E-134 (Martinus, Nijhoff Publishers, Boston, 1987); Y-K Swee, H-Y Zheng, and R. T. Chen, *Microelectronic Packaging and Laser Processing* (SPIE Publications, Washington, 1997); W. M. Steen, *Laser Material Processing* (Springer, NY, 1998); H. E. Pontath and G. I Stegeman, editors, *Nonlinear Surface Electromagnetic Phenomena, Modern Problems in Condensed Matter Physics* (North Holland Press, NY, 1991), Vol. 29; R. M. Osgood, Jr., editor, *Laser-Assisted Microtechnology*, Springer Series in Materials Science, (Springer, NY, 1998), Vol. 19.
4. W. W. Duley, *Laser Processing and Analysis of Materials* (Plenum Press, NY, 1983), p. 37.

5. An-S. Chu, S. H. Zaidi, and S. R. J. Brueck, "Fabrication and Raman Scattering Studies of One-Dimensional Nanometer Structures in (110) Silicon," *Appl Phys. Lett.* 63 (7), 905 (1993).
6. Y. Lu *et al.*, "Wet-Chemical Etching of Mn-Zn Ferrite by Focused Ar^{+}-Laser Irradiation in H_3PO_4," *Appl. Phys.* A47, 319 (1988).
7. M. Rothschild and D. Ehrlich, "A Review of Excimer Laser Projection Lithography," *J. Vac. Sci. Technol.* B6, 1 (1988).
8. H. Kumagai *et al.*, "Ablation of Polymer Films by a Femtosecond High Peak Power Ti:Sapphire Laser at 798 nm," *Appl. Phys. Lett.* 65, 1850 (1994).
9. J. Ouellette, "Free-Electron Lasers: A Radical Alternative," *Indust. Phys.* 3, 18 (1997).
10. S. Silverman, R. Aucoin, J. Mallatt, and D. Ehrlich, "Laser Microchemical Technology: New Tools for Microsystems Engineering, Debug and Failure Analysis," *SPIE* 2991, 129–137 (1997).
11. J. H. Brannon, J. R. Lankard, A. I. Baise, F. Burns, and J. Kaufman, "Excimer Laser Etching of Polyimide," *J. Appl. Phys.* 58, 2036–2043 (1985).
12. W. W. Hansen, S. W. Janson, and H. Helvajian, "Direct-Write UV Laser Microfabrication of 3D Structures in Lithium-alumosilicate Glass," *SPIE* 2991, 104–112 (1997).
13. N. Nassuphis, R. H. Mathews, S. T. Palmacci, and D. J. Ehrlich, "Three Dimensional Laser Direct Writing: Applications to Multichip Modules," *J. Vac. Sci. Technol. B.* 12(6), 3294–3295 (1994).
14. G. M. Daly, D. B. Chrisey, J. M. Pond, M. Osofsky, M. Miller, P. Lubitz, J. S. Horwitz, R. C.Y. Auyeung, and R. J. Soulen, Jr., "Pulsed Laser Deposition of High Temperature Superconducting and Metallic Thin Films for Novel Three Terminal Device Applications," *SPIE* 2991, 226–237 (1997).
15. V. I. Konov, F. Dausinger, S. V. Garnov, S. M. Klimentov, T. V. Kononenko, and O. G. Tzarkova, "Ablation of Ceramics by UV, Visible and IR Pulsed Laser Radiation," *SPIE* 2991, 151–160 (1997).
16. F. Gonella, G. Mattei, P. Mazzoldi, E. Cattaruzza, G. W. Arnold, G. Barraglin, P. Calvelli, R. Polloni, R. Bertoncello, and R. F. Haglund, Jr., "Interaction of High-Power Light with Silver Nanocluster Composite Glasses," *Appl. Phys. Lett.* 69, 3101–3103 (1996).
17. M. Rothschild, C. Arnone, and D. J. Ehrlich, "Excimer-Laser Etching of Diamond and Hard Carbon Films by Direct-Writing and Optical Projection," *J. Vac. Sci. Technol. B.* 4, 310–314 (1986).
18. I. W. Boyd, "Doping and Oxidation," in *Laser Microfabrication–Thin Film Processes and Lithography*, edited by J. Y. Tsao and D. J. Ehrlich (Academic Press, NY, 1989), p. 542.
19. S. H. Zaidi and S. R. J. Brueck, "Multiple-Exposure Interferometric Lithography," *J. Vac. Sci. Technol. B.* 11(3), 658 (1993).
20. A. G. Cullis, H. C. Webber, and P. Bailey, "A Device for Laser Beam Diffusion and Homogenization," *J. Phys. E. Sci. Instrum.* 12, 688 (1979).
21. H. Kogelnik, and T. Li, "Laser Beams and Resonators," *Proc. IEEE 54* (1966), p. 1312.
22. A. E. Siegman, *Lasers* (University Science Books, Mill Valley CA, 1986), p. 676.
23. M. Born and E. Wolf, *Principals of Optics*, 5th ed. (Pergamon Press, NY, 1975) p. 419.
24. D. J. Ehrlich, J. Y. Tsao, C. O. Bozler, "Submicron Patterning by Projected Excimer Laser-Based Beam Induced Chemistry," *J. Vac. Sci. Technol. B3*, 1 (1985).
25. Y. S. Liu, "Sources, Optics and Laser Microfabrication Systems for Direct Write and Projection Lithography," in *Laser Microfabrication--Thin Film Processes and Lithography*, edited by J. Y. Tsao and D. J. Ehrlich (Academic Press, 1989), p. 3.
26. K. M. A. El-Kader, J. Oswald, J. Kocka, and V. Chab, "Formation of Luminescent Silicon by Laser Annealing of a-Si:H," *Appl. Phys. Lett.* 64, 2555 (1994).
27. G. V. Treyz, R. Beach, and R. N. Osgood, Jr., "Rapid Direct Writing of High Aspect-Ratio Trenches in Silicon," *Appl. Phys. Lett.* 50, 475 (1987).
28. G. B. Shinn, F. Steigerwald, H. Stiegler, R. Sauerbrey, F. K. Tittle, and W. L. Wilson Jr., "Excimer Laser Photoablation of Silicon," *J. Vac. Sci. Technol. B4*, 1273 (1986).
29. M. Ishii, T. Meguro, T. Sugano, K. Gamo, and Y. Aoyagi, "Digital Etching by Using a Laser Beam: On the Control of Digital Etching Products," *Appl. Surf. Sci.* 79/80, 104 (1994).
30. J. Y. Tsao and D. J. Ehrlich, editors, *Laser Microfabrication –Thin Film Processes and Lithography*, edited by (Academic Press, NY, 1989); R. Haglund, Jr., and R. Kelly, "Electronic Processes in Sput-

tering by Laser Beams," in *Fundamental Processes in Sputtering of Atoms and Molecules*, edited by P. Sigmund (Munksgaard, Copenhagen, 1993).
31. M. Eyett and D. Bauerle, "Influence of the Beam Spot Size on Ablation Rates in Pulsed-Laser Processing," *Appl. Phys. Lett.* 51, 2054 (1987).
32. C. I. H. Ashby, "Laser Driven Etching" in *Thin Film Processes II* (Academic Press, NY, 1991), p. 783.
33. D. V. Podlesnik, H. H. Gilgen, and R. M. Osgood, Jr., "Waveguiding Effects in Laser-Induced Aqueous Etching of Semiconductors," *Appl. Phys. Lett.* 48, 496 (1986).
34. R. J. Wallace, M. Bass, S. M. Copley, "Curvature of Laser-Machined Grooves in Si3N4," *J. Appl. Phys.* 59, 3555 (1986).
35. C. Kittel, *Introduction to Solid State Physics*, 6th ed. (John Wiley & Sons, NY, 1986).
36. D. Ashkenasi, A. Rosenfeld, H. Varel, M. Wahmer, and E. E. B. Campbell, "Laser Processing of Sapphire with Picosecond and Sub-picosecond Pulses," *Appl. Surf. Sci.* 120, 65 (1997).
37. L. Wiedeman and H. Helvajian, "Laser Photodecomposition of Sintered YBCO: Ejected Species Population Distributions and Initial Kinetic Energies for the Laser Ablation Wavelengths 351, 248, and 193," *J. Appl. Phys.* 70, 4513 (1991); H. Helvajian and R. Welle, "Threshold Level Laser Photoablation of Crystalline Silver: Ejected Ion Translational Energy Distributions, *J. Chem. Phys.* 91, 2616 (1989); H. Helvajian, "Surface Excitation Mediated Physics in Low-Fluence Laser Material Processing," *SPIE* 2403 1 (1995); R. H. Ritchie, J. R. Manson, and P. M. Echenique, "Surface Plasmon-Ion Interaction in Laser Ablation of Ions from a Surface," *Phys. Rev. B.* 49, 2963 (1994); D. P. Taylor, W. C. Simpson, K. Knutsen, M. A. Henderson, and T. M Orlando, "Photon Stimulated Desorption of Cations from Yttria-Stabilized Cubic ZrO_2(100)," in *Laser Ablation*, edited by R. E. Russo, D. B. Geohegan, R. F. Haglund, Jr., and K Murakami (Elsevier, NY, 1997), p. 101.
38. J. F. Ready, *Effects of High-Power Laser Irradiation* (Academic Press, NY, 1971).
39. C. I. H. Ashby, J. Y. Tsao, "Photophysics and Thermophysics of Absorption and Energy Transport in Solids" in *Laser Microfabrication--Thin Film Processes and Lithography*, edited by J. Y. Tsao and D. J. Ehrlich (Academic Press, NY, 1989), p. 272.
40. J. T. Dickinson, S. C. Langford, J. J. Shin, and D. L. Doering, "Positive Ion Emission from Excimer Laser Excited MgO Surfaces," *Phys. Rev. Lett.* 73, 2630 (1994).
41. F. Wood, and D. H. Lowndes, "Laser Processing of Wide Band Gap Semiconductors and Insulators," *Cryst. Latt. Def. and Amorph. Mat.* 12, 475 (1986); A. H. Guenther, and J. K. Mciver, "The Role of Thermal Conductivity in the Pulsed Laser Damage Sensitivity of Optical Thin Films," *Thin Solid Films* 163, 203 (1988).
42. L. P. Welsh, J. A. Tuchman, and I. P. Herman, "The Importance of Thermal Stresses and Strains Induced in Laser Processing with Focused Gaussian Beams," *J. Appl. Phys.* 64, 6274 (1988).
43. A. L. Dawar, S. Roy, T. Nath, S. Tyagi, and P. C. Mathur, "Effect of Laser Annealing on Electrical and Optical Properties of n-Mercury Cadmium Telluride," *J. Appl. Phys.* 69, 3849 (1991).
44. M. Raff, M. Schutze, C. Trappe, R. Hannot, and H. Kurz, "Laser-Stimulated Nonthermal Particle Emission from InP and GaAs Surfaces," *Phys. Rev. B* 50, 11031 (1994);
45. H. Helvajian and R. Welle, "Threshold Level Laser Photoablation of Crystalline Silver: Ejected Ion Translational Energy Distributions," *J. Chem. Phys.* 91, 2616 (1989).
46. L. P. Smith, "The Emission of Positive Ions from Tungsten and Molybdenum," *Phys. Rev.* 35, 381 (1930).
47. A. N. Pirri, R. G. Root, P. K. S. Wu, "Plasma Energy Transfer to Metal Surfaces Irradiated by Pulsed. . .," AIAA J. 16, 1296 (1978).
48. L. Spitzer, *Physics of Fully Ionized Gases* (Wiley-Interscience, NY, 1956).
49. A. C. Tam, *et al.*, "Experimental and Theoretical Studies of Bump Formation During Laser Texturing of NiP Disk Substrates," *IEEE Trans. Mag.* 32, 3771 (1996).
50. S. Lugomer, *Laser Technology: Laser Driven Processes* (Prentice Hall, Englewood Cliffs, 1990), Chap. 5.

51. L. F. Thompson, "Microlithography: The Physics," in *Introduction to Microlithography*, edited by L. F. Thompson, C. G. Willson, and M. J. Bowden (American Chemical Society, Washington, D.C., 1983), Chap. 1.
52. J. Brannon, *Excimer Laser Ablation and Etching*, AVS Monograph M-10 (American Vacuum Society, NY, 1993), Chap. 6.
53. E. Hecht and A Zajac, *Optics* (Addison-Wesley, Reading, 1974), Chap. 10.
54. A. Marchant, *Optical Recording* (Addison-Wesley, Reading, 1990), Chap. 7.
55. J. Brannon, "Micropatterning of Surfaces by Excimer Laser Projection," *J. Vac. Sci. Technol.* B7, 1064 (1989).
56. Y.S. Liu "Sources, Optics, and Laser Microfabrication Systems for Direct Writing and Projection Lithography," in *Laser Microfabrication: Thin Film Processes and Lithography*, edited by D. J. Ehrlich and J. Y. Tsao (Academic Press, NY, 1989), p. 3; J. J. Ritsko "Laser Etching" in *Laser Microfabrication: Thin Film Processes and Lithography*, edited by D. J. Ehrlich and J. Y. Tsao (Academic Press, NY, 1989), p. 33.
57. A. Berezhnoi, *Glass Ceramics and Photo-Sitalls* (Plenum Press, NY, 1970).
58. *IBM J. Res. Dev.* 36(5) (1992). The September issue is fully devoted to the ES9000 semiconductor and packaging technologies.
59. R. Srinivasan and V. Mayne-Banton, "Self-Developing Photoetching of Poly(ethylene terephthalate) Films by Far-Ultraviolet Excimer Laser Radiation," *Appl. Phys. Lett.* 41, 576 (1982).
60. J. Andrew *et al.*, "Direct Etching of Polymeric Materials Using a XeCl Laser," *Appl. Phys. Lett.* 43, 717 (1983).
61. J. Lankard and G. Wolbold, "Excimer Laser Ablation of Polyimide in a Manufacturing Facility," *Appl. Phys.* A54, 355 (1992).
62. K. Prasad and E. Perfecto, "Multilevel Thin Film Packaging: Applications and Processes for High Performance Systems," *IEEE Trans. Comp. Hybrid Mfg. Tech.* 17, 38 (1994).
63. G. Wolbold, C. Tessler, and D Tudryn, "Polymer Ablation with a High-Power Excimer Laser Tool," *Microelec. Eng.* 20, 3 (1993); J. Lankard and G. Wolbold, "Excimer Laser Ablation of Polyimide in a Manufacturing Facility," *Appl. Phys.* A54, 355 (1992); R. Patel *et al.*, "Laser Via Ablation Technology for MCM-D Fabrication at IBM Microelectronics," *Int. J. Microcircuit Elec. Pack.* 18, 266 (1995).
64. R. Patel *et al.*, "Laser Via Ablation Technology for MCM-D Fabrication at IBM Microelectronics," *Int. J. Microcircuit Elec. Pack.* 18, 266 (1995).
65. W. F. Iceland, "Design and Development of Equipment for Laser Wire Stripping," *SPIE* 86, 68 (1976); J. P. Wheeler, "Industrial Applications of Low-Power CO_2 Lasers," *SPIE* 668, 236 (1986); R. T. Miller, "Laser Wire Stripping," *SPIE* 744, 94 (1987).
66. J. Brannon, A. C. Tam, and R. Kurth, "Pulsed Laser Stripping of Polyurethane-Coated Wires: A Comparison of KrF and CO_2 Lasers," *J. Appl. Phys.* 70, 3881 (1991).
67. K. Ashar, *Magnetic Disk Drive Technology* (IEEE Press, NY 1997), Chap. 1.
68. P. Baumgart, D. Krajnovich, T. Nguyen, and A. C. Tam, "A New Laser Texturing Technique for High Performance Magnetic Disk Drives," *IEEE Trans. Mag.* 31, 2946 (1995).
69. P. Baumgart, D. Krajnovich, T. Nguyen, and A. C. Tam, "Safe Landings: Laser Texturing of High-Density Magnetic Disks," *Data Storage* (March 1996).
70. A. C. Tam, J. Brannon, P. Baumgart, and I. Pour, "Laser Texturing of Glass Disk Substrates," *IEEE Trans. Mag.* 33 (1997).
71. D. B. Chrisey and G. K. Hubler, editors, *Pulsed Laser Deposition of Thin Films* (John Wiley & Sons, Inc., NY, 1994).
72. A.A. Puretzky, C. B. Geohegan, G.E. Jellison, Jr., and M. M. McGibbon, "Amorphous Diamond-Like Carbon Films Growth By KrF- and ArF- Excimer Laser PLD:Correlation with Plume Properties," *MRS* (1995); D. L. Pappas, K.L. Saenger, J. Bruley, W. Krakov, J.J. Cuomo, and R. W. Collins, *J. Appl. Phys.* 71, 5675 (1992).

73. J. A. Greer and M. D. Tabat, "Large-Area Pulsed Laser Deposition: Techniques and Applications," *J. Vac. Sci. Technol. A* 13(3), 1175–1181 (1995).
74. J. A. Greer, M. D. Tabat, and C. Lu, "Future Trends for Large-Area Pulsed Laser Deposition," *Nuclear Instruments and Methods in Physics Research*, B(121), 357–362 (1997).
75. J. A. Greer and M. D. Tabat, "Properties of Laser-Deposited Yttria Films on CdTe and Silicon Substrates," *Mat. Res. Soc. Symp. Proc.* 341, 87–94 (1994).
76. J. A. Greer and M. D. Tabat, "On- and Off-Axis Large-Area Pulsed Laser Deposition," *Mat. Res. Soc. Symp. Proc.* 388, 151–161 (1995).
77. C. Roychoudhuri, "Desk-Top Manufacturing Using Diode Lasers," *SPIE* 3274, 162–170 (1998).
78. G. Ogura and B. Gu, "Review of Laser Micromachining in Contract Manufacturing," *SPIE* 3274, 171–182 (1998).
79. S. Janson, "Spacecraft as an Assembly of ASIMS," in *Microengineering Technology for Space Systems*, Monograph 97-02, edited by H. Helvajian (The Aerospace Press, El Segundo, CA, 1997), p. 143. First published as The Aerospace Corp. Report no. ATR-95(8168)-2 (1995).

6

Rechargeable Li-ion Batteries for Satellite Applications: Pros and Cons*

J-M. Tarascon† and G. G. Amatucci‡

6.1 Introduction

For many years rechargeable lithium batteries have been considered a superior alternative power source for a wide variety of applications. However, it is only since the 1990s that rechargeable lithium-ion (Li-ion) cells have become key components of the portable, entertainment, computing, and telecommunications equipment required by an information-rich, mobile society. The electric vehicle (EV) industry, always searching for long-lasting batteries with greater autonomy, is seriously considering the rechargeable Li technology as a viable solution. The aerospace industry is also in need of improved batteries. Satellite reliability, cost-effectiveness, and performance depend on many factors, a critical one being the selection of an appropriate battery technology.

This chapter will discuss the advantages and disadvantages of the Li-ion technology as a powering source for satellite applications. A rechargeable plastic Li-ion battery (PLiON™) technology, a modified version of the liquid Li-ion technology first commercialized in the 1990s by Sony, has recently been developed. Advantages of this technology with respect to satellite and, more specifically, nanosatellite applications will be discussed. Performance and design flexibility of this technology will be compared with those of Ni/Cd or Ni/H_2 batteries, currently used for satellite applications.

A successful satellite mission depends on the proper function of the power system of the spacecraft in orbit over extended periods of time; therefore, continual efforts are being made to realize more reliable power systems. The electrical power onboard a spacecraft generally involves four basic elements:

- A primary source of energy such as a solar cell
- A device for converting the primary energy into electrical energy
- Chemical batteries for storing the electrical energy to meet the peak and/or eclipse demands
- A system for conditioning charging/discharging[1]

The main source of primary power for satellites is solar radiation. However, such a primary source of power, if not coupled with a supplementary source that can store electrical energy, is of little use in most applications. Chemical sources such as rechargeable batteries serve such a purpose. Specifications of a battery for satellite applications depend on the spacecraft power requirements, which are contingent to a large extent upon the nature of the mission.

In any spacecraft power system based on the use of solar energy, the storage battery is the main source of continuous power since it is called upon to respond to peak and eclipse power demands. Such demands depend upon the satellite orbit. For low Earth-orbit (LEO) spacecraft, the number of eclipses increases as the altitude decreases. Typically, for a 550-km orbit, there will be about 15 eclipses per day, 5500 per year, or about 40,000 per usual satellite lifetime (e.g., 7 years). This

---

†*Universite de Picardie Jules Verne*, Amiens, France

‡Bellcore, Energy Storage Research Group, Red Bank, New Jersey

translates into 40,000 charge/discharge cycles for the selected battery.[1] These cycles will not be deep cycles (the percentages will be defined later) since for an orbital radius of 550 km, the duration of each eclipse is estimated to be 36 min. Although partial charge/discharge cycles are required, with such a large number it is mandatory that the power storage system proceeds reversibly without loss (e.g., without any capacity loss between subsequent charge/discharge cycles). An additional factor of importance for satellite applications is the weight, since each additional pound shortens the lifetime of a satellite in orbit by one month and adds to the launch costs associated with the satellite. Therefore, the rechargeable batteries must be lightweight (e.g., batteries with large gravimetric energy densities).

In short, "an ideal storage cell" for space applications has the following primary requirements.

- Ability to accept and deliver power at high rates
- Long charge/discharge cycle life under a wide range of conditions
- High recharge efficiency
- Low impedance
- Good hermetic seals throughout thousands of electrical cycles involving concurrent pressure and thermal changes
- Operation in all physical orientations
- Ability to withstand launch and space environments
- Stable long-term overcharge characteristics
- Maximum usable energy per unit weight and volume at low cost with high and proven reliability

At present no single battery technology can meet all these requirements. Choosing a technology, therefore, depends mainly on a knowledge of the available and emerging rechargeable batteries, which requires a detailed survey of such technologies. Rather than conduct a complete survey of various technologies, we present how an emerging technology, that is, the Li-ion technology both in its liquid and plastic form, can provide a viable solution to the powering issues faced by satellite applications.

The chapter is in three sections. The first introduces battery nomenclature and briefly retraces the historical development of lithium batteries. The second deals with the chemistry, materials issues, and performances of Li-ion technology in conjunction with a comparative study of the performance of the Ni/Cd and Ni/H_2 battery technologies already accepted for use in today's satellites. The final section introduces the plastic Li-ion technology and its potential for space applications through exploitation of its intrinsic design and flexibility. A tentative time is forecast for the implementation of the Li-ion technology either in its liquid or plastic forms with respect to space applications, keeping in mind that the choice of a battery depends on suppliers and availability as well as technical characteristics.

6.2 Historical Development of Lithium Batteries

6.2.1 A General Introduction to Batteries

Just as a molecule is composed of several atoms, a battery is composed of several electrochemical cells. These cells are connected in series and/or parallel to provide the required voltage and capacity, respectively. Each cell consists of a positive electrode and a negative electrode (both sources of chemical reactions) immersed in an electrolyte medium, a solution containing dissociated salts, which allows ion transfer between the two electrodes. Once these electrodes are externally connected through a resistor, chemical reactions proceed in tandem at both electrodes, thereby liberating electrons and allowing a current to flow through the resistor (to perform work). Thus a cell can simply be viewed as an electrochemical device that stores energy in the chemical

form and converts this chemical energy into electrical form during discharge. Depending upon the nature of the chemical reactions taking place at the electrodes, primary and secondary rechargeable cells will be distinguished.[2]

In a primary cell, the chemical reactions are not reversible, so that once discharged, the cell cannot convert electrical energy back into chemical energy and must be disposed of. In contrast, with secondary cells, the chemical reactions are perfectly reversible so that chemical energy can be transformed several times into electrical energy and vice versa.

The amount of electrical energy that a cell is able to deliver is a function of the cell potential and capacity, both linked directly to the chemistry of the system. The cell potential (cell voltage, V_{cell}) is the difference between the potential of the redox reactions occurring at the positive (V^+) and negative (V^-) electrodes simultaneously.

$$V_{cell} = V^+ - V^- \tag{6.1}$$

The cell capacity refers to the total quantity of electricity (number of electrons transferred) involved in the electrochemical reaction and is defined in ampere-hours (*Ah*). The gravimetric capacity of an electrode in *Ah/g* units is defined as the number of electrons involved in the redox reaction multiplied by 26.8 *Ah* (26.8 *Ah* being the capacity delivered by 1 gram-equivalent weight of material) and divided by the molecular weight of the active material (*Ah/g*).[3]

$$\text{Capacity}\left(\frac{Ah}{g}\right) = \frac{26.8ne}{MW}\left(\frac{Ah}{g}\right) \tag{6.2}$$

The product of the *Ah* of an electrode material (or complete cell) and its potential will give its energy density (*Wh*),

$$\text{Energy density } (Wh) = Ah \cdot V, \tag{6.3}$$

which can be expressed either per gram (gravimetric *Wh/kg*) or per liter (volumetric *Wh/l*). Both energy density characteristics are critical in evaluating the performance of battery technology. The power energy of a cell refers to the rate at which the cell can release its chemical energy. We will use the C-rate scale, where 1 C is the current at which the battery energy will be fully utilized in one hour. The capacity delivered at 1 C is compared with theoretical capacity and is given on a percentage scale. For example, percentage capacity can be given for different rates, C/2 for 2 hours and 2 C for half an hour.

6.2.2 Why Li Metal?

The motivation for using a battery technology based on Li metal lies in the fact that it is the lightest and also the most electropositive metal in the electromotive series. The low atomic mass of lithium compared with that of lead, for instance, results in a specific capacity of 3800 *Ah/Kg*, 14 to 15 times higher than that of lead or 4 to 5 times higher than that of nickel (260 *Ah/Kg* and 900 *Ah/Kg*, respectively). The advantage of using Li metal was first demonstrated in primary Li cells.[4] Such cells used Li as anode and an inorganic compound as cathode, which when electrochemically reacted with Li^+, led to an irreversible displacement/decomposition reaction of the general type:

$$2M^{4+}O_2 + 2Li^+ \rightarrow M_2^{3+}O_3 + Li_2O \tag{6.4}$$

Where *M* is a transition metal, this type of reaction is irreversible.[2,4] However, there now exists an entirely different family of inorganic compounds, called intercalation compounds, that are able

to reversibly intercalate cations ($A^+ + e^- + MX_n \leftrightarrow AMX_n$) while maintaining their framework structure.[5,6] Within an intercalation reaction a guest specie (A^+) can reversibly enter and be removed from the vacant structural sites of the host structure (MX_n). Intercalation can only occur if the host material has both a crystallographic and electronic structure that is able to accept ions and electrons, respectively. Among the basic requirements for the solid-state intercalation electrode materials are

- Ability to reversibly insert maximum amount of Li^+ (resulting in a large electrochemical capacity)
- High diffusion of the guest species in the host (resulting in high-power densities)
- Minimal structural change (e.g., resulting in highly reversible reaction)
- Good electronic conductivity to eliminate the need for conducting additives
- Low solubility in the electrolyte to prevent high self-discharge
- Large free energy of reaction (high voltage)[2]

The chemical potential of Li into the intercalation material is governed by the redox potential of the transition metal $M^{+n} / M^{+(n-1)}$ redox reaction, which is a strong function of the ionic-covalent nature of the $M-X$ (X is an anion) bond.[7] The highest potential is established by bonds with the greatest ionicity, which is why oxides exhibit greater potential than the more covalently bonded chalcogenides. In addition, the chemical potential is also governed by the local environment of the intercalating Li ion. For example, a lithium ion tetrahedrally coordinated by the oxygen ligands may have a higher voltage than a lithium ion situated in an octahedral coordination.

In the 1970s, researchers recognized the possibility of utilizing intercalation reactions at the cathode of a galvanic cell. A typical rechargeable lithium intercalation battery [Fig. 6.1(a)], proposed in the 1970s on this principle, uses as the positive electrode, an intercalation material;[5,6,8] as the negative electrode, lithium metal; and as the electrolyte, a solution of some Li-bearing salt in an organic electrolyte (for instance $LiPF_6$ in a mixture of ethylene carbonate and dimethyl carbonate).

During the charge and discharge cycles of a rechargeable Li cell, as indicated by Eqs. (6.5)–(6.7), the positive electrode material undergoes a bulk reversible electrochemical reaction (lithium deintercalation↔intercalation within the open structure of the material).

$$xLi \leftrightarrow xLi^+ + xe^- \text{ (negative electrode)} \tag{6.5}$$

$$xLi^+ + xe^- + MO_n \leftrightarrow Li_xMO_n \text{ (positive electrode)} \tag{6.6}$$

$$MO_n + xLi \rightarrow Li_xMO_n \tag{6.7}$$

Surface reactions, namely lithium plating and stripping on the electrode surface, occur at the lithium metal electrode.

Despite numerous intercalation electrode materials, only primary lithium metal cells such as Li/CF_x and others have been commercialized to date.The difficulties in commercializing room-temperature rechargeable Li metal batteries, despite their promises in laboratory prototypes, can largely be traced to safety problems associated with the use of lithium metal as the anode. During subsequent charge/discharge cycles, Li is removed and replated on the metallic Li negative electrode. Dendrites or low-density, highly reactive, "mossy" lithium may be formed instead of smooth replating on the Li metal surface. The dendritic growth of lithium during the recharge cycle can lead to electrical shorts through the separator and catastrophic failure of the battery.[9]

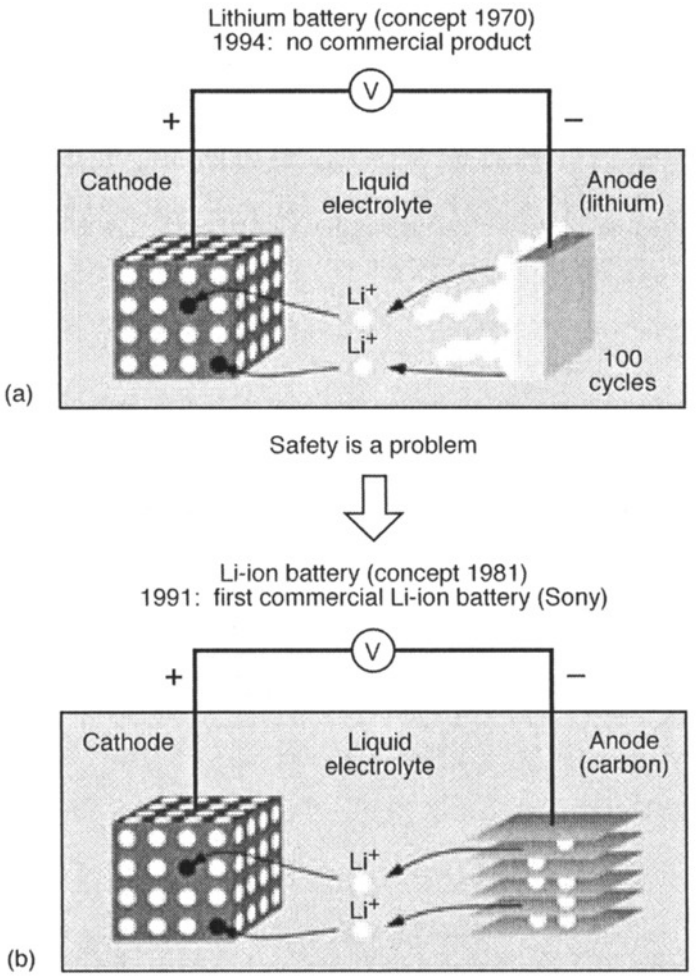


Fig. 6.1. (a) Graphical representation of the $LiMn_2O_4$/liquid electrolyte/Li-metal battery showing formation of deleterious Li metal dendritic growth on the anode. (b) Representation of the Li-ion battery technology in which the Li metal anode is substituted for a carbon-based intercalation compound.

6.3 The Li-ion Battery

In the early 1980s, two approaches were proposed to circumvent the problems associated with the growth of lithium dendrites at the negative electrode. One approach replaces the liquid electrolyte with a solid polymer electrolyte, which impedes the growth of dendrites between the two electrodes.[10] This research led to the so-called lithium solid polymer batteries, extensively studied, by Hydroquebec of Canada for example, during the last 20 years.[11] While this approach is compatible with large-scale manufacturing techniques, it has not completely solved the dendrite growth problem. The other approach replaces the metallic lithium at the negative electrode by another intercalation compound (e.g., another lithium sponge) so that the lithium activity is reduced, since no lithium metal exists in this battery, only lithium ions.[12–14] A battery fabricated this way, in which the lithium ions can be shuttled (or rocked) from one sponge to another as the cell is cycled, is commonly known as a Li-ion or "rocking chair" battery [Fig. 6.1(b)]. When optimized,

these batteries have lower energy densities than the Li-metal batteries; however, they exceed existing battery technologies in both gravimetric and volumetric energy densities (Fig. 6.2).

The Li-ion battery is typically fabricated in its discharged state, with the lithium contained as ions within structural sites of the cathode material. In Li-ion technology, no lithium metal exists. The schematic in Fig. 6.3 illustrates the reactions involved during the charging and discharging of the cell.[16] During the charge cycle of the Li-ion battery, Li-ions are extracted (deintercalated) from the cathode structure and passed into the ionically conductive electrolyte. Simultaneously, an electron is extracted from the transition metal in the cathode and is passed through the external circuit. The result is oxidation of the transition metal.The lithium ion is then intercalated (instead of plating on Li metal) into the anode intercalation material from the electrolyte simultaneously with the electron from the external circuit.

During the discharge cycle, the reverse reaction takes place. The electron is removed from the anode and is passed through the external circuit where the subsequent electric current is used to do work. It then reinserts into the cathode by charge transfer and reduces the transition metal. Simultaneously, the lithium ion is extracted from the anode, passes through the electrolyte, and reintercalates into the cathode. The output voltage of the cell is simply the difference in chemical potentials versus Li/Li^+ of the cathode and anode.

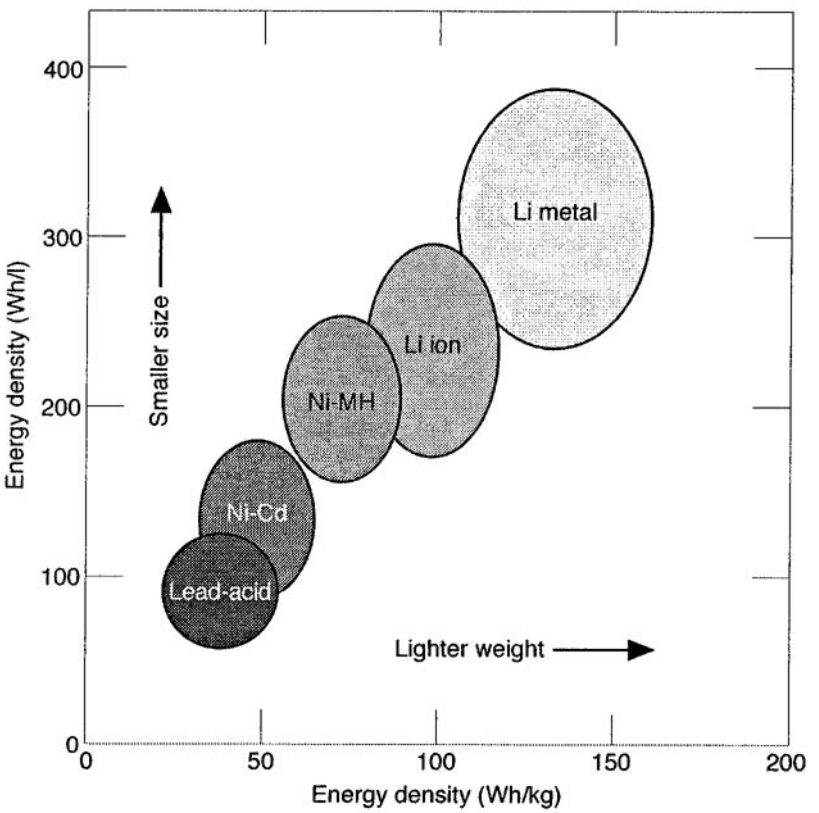


Fig. 6.2. Comparison of volumetric and gravimetric energy densities for a variety of rechargeable battery systems.[15]

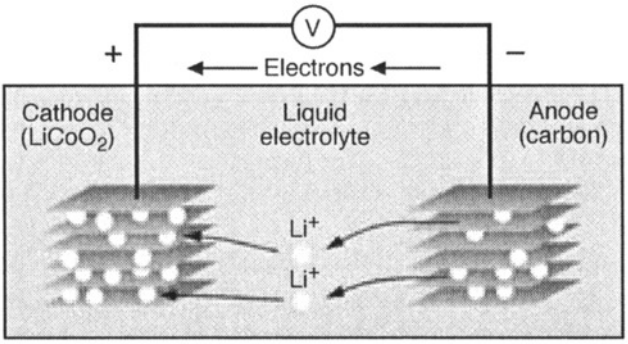


Fig. 6.3. Illustration of the electrochemical reactions that occur in the Li-ion battery upon charge and discharge.

This type of battery is inherently safer because in contrast to the metallic state where Li is reduced in Li-metal batteries, in the Li-ion battery, lithium is always confined to the ionic state. Sony was the first company to commercialize a Li-ion cell in June 1991,[17] and several other battery companies (Sanyo, Matsushita, Fuji, Toshiba, Hitachi, Yuasa, Moli[18]) are currently developing the Li-ion technology using a liquid organic electrolyte. The feasibility of large-size Li-ion batteries for EV applications was demonstrated in 1996 by Sony. This demonstration of the Li-ion technology scaleability opens new markets for the Li-ion battery, namely, the space application market where Li-ion technology could favorably compete with the Ni-Cd and Ni-H_2 technologies that dominate this application domain.

6.3.1 The Chemistry

To alleviate the safety problems associated with a lithium-metal based battery, a Li-ion cell utilizes intercalation reactions at the cathode and at the anode, as indicated in Eqs. (6.8)–(6.10). However, a price is paid in terms of average output voltage and energy density as compared to a lithium metal cell. In addition, replacing Li metal by another intercalation compound results in an overall cell capacity penalty, which can be minimized by selecting intercalation anodes having large electrochemical capacities. The following reaction equations show intercalation reactions for positive and negative electrodes using transition metal (M, N) oxides as intercalation hosts for both the positive and negative electrodes.

$$Li_1NO_n \leftrightarrow Li_{1-x}NO_n + Li^+ + e^- \text{(positive electrode)} \tag{6.8}$$

$$xLi^+ + xe^- + MO_n \leftrightarrow Li_xMO_n \text{(negative electrode)} \tag{6.9}$$

$$Li_1NO_n + MO_n \rightarrow Li_{1-x}NO_n + Li_xMO_n \tag{6.10}$$

A Li-ion cell's voltage is defined by the difference in the chemical potential of lithium within each intercalation electrode material; therefore, to ensure a high cell voltage and high energy density, strongly oxidizing (e.g., high V^+) and strongly reducing (e.g., low V^-) intercalation compounds must be used for the positive and negative electrodes, respectively. A V^- that is too close to 0 V is not totally satisfactory either, since it will enhance the risk of the Li plating at the negative electrode during the recharge of the Li-ion cell, thereby defeating the purpose of the Li-ion concept.[19] Because of this danger, compounds with intercalation voltages slightly above that of 0 V versus Li/Li^+ are used. Present day compounds are now approximately 0.01–0.05 V versus

Li/Li^+. Furthermore, while very attractive, the use of both highly oxidizing and reducing intercalation electrodes also necessitates electrochemically stable electrolytes that can operate over a wide range of potential (0–5 V).

Within a Li-ion cell, the positive electrode stands as the only source of Li, so for practical manufacturing, this electrode must be stable in air. In light of this, only lithiated compounds for which the chemical potential of Li is greater than 3.4 V (greater than the oxidation potential of water vs Li) present a potential interest for Li-ion batteries. If Li potential is lower than 3.2 V, then this compound will be oxidized by water ($Li_xMO_n \rightarrow MO_n + LiOH + H_2$).

In short, implementing the Li-ion concept requires high-performance electrode and electrolyte materials. This is the reason why Li-ion technology took many years to reach the marketplace. The first functioning rocking-chair batteries were based on lithiated negative electrode materials. Systems using $LiWO_2$,[20,21] $Li_6Fe_2O_3$,[22] or $Li_9Mo_6Se_6$[23] as the negative electrode combined with TiS_2, NbS_2, or Mo_6Se_6 as the positive electrode were built and tested.[24] Because of the low V^+ and high V^- output voltages of these materials, the specific energy density for such Li-ion cells was at least 3 times lower than that for their Li-metal counterparts and even lower than that of Ni-Cd cells. In addition, the negative materials were not air stable. To improve some of these drawbacks, rocking-chair cells based on new lithiated positive electrode materials were proposed in which the air-stable layered intercalation compound $LiCoO_2$ was coupled with either MoO_2,[25] WO_2,[25] or TiS_2[26] negative electrodes. The highest Li intercalation voltage of 4 V for the $LiCoO_2$ positive electrode material allowed the use of TiS_2 (previously given in example as a positive electrode) as a negative electrode in the cobalt system. The TiS_2-based cell was found to have the largest capacity, but because of the low output cell voltage (due to the remaining large V^-), its energy was well below that of the Li-metal counterpart cell or just similar to that of the Ni-Cd cells. The poor performance of the Li-ion systems alluded to above combined with the costly synthesis of the initial negative electrode materials from the non-lithiated material did not justify their commercialization.

These studies confirmed the viability of the Li-ion technology, but also showed that the practical benefits of this technology were limited by the lack of high-capacity negative electrode materials. The discovery by Japanese researchers[27,28] that some forms of carbon (C) could be used as reversible low-voltage lithium intercalation materials drastically changed the scenario. Following this discovery, a carbon (V^-) / $LiCoO_2$ (V^+) rocking-chair cell with a performance exceeding that of Ni-Cd cells was demonstrated. Recent developments in the field of rechargeable cells utilizing carbon anodes as the negative intercalation compound have resulted in batteries with specific energies almost twice that of Ni-Cd batteries.[22] Moreover, the carbon materials used as negative electrodes have the important advantage of being abundant, inexpensive, and nontoxic. They exhibit a highly reversible electrochemical behavior, thus fulfilling the three important requirements for a technology to succeed: safety, performance, and affordable cost.

6.3.2 Li-ion Battery Technology: Materials Issues

6.3.2.1 Positive Electrode

Over the years, numerous materials have been evaluated for their potential use as intercalation electrodes. Figure 6.4 shows the voltage range for lithium intercalation in various known lithium intercalating materials. The data clearly show which materials are appropriate to build a high-

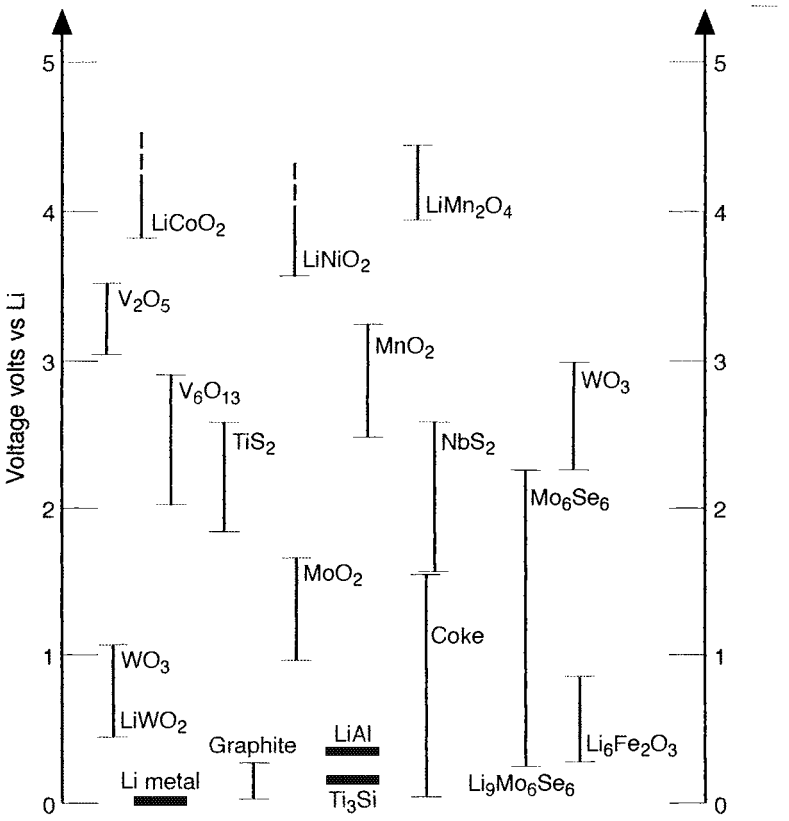


Fig. 6.4. Schematic showing the voltages vs Li for a variety of potential positive and negative electrode materials.

voltage, rocking-chair lithium battery. Materials for the positive or negative electrode should be chosen in the high or low potential range, respectively. Besides $LiCoO_2$, other oxides such as $LiNiO_2$[18] and $LiMn_2O_4$[29–35] were proposed for use in carbon-based Li-ion systems as they all retain voltages versus Li/Li^+ in the 4 V range. $LiMO_2$ (M = Ni, Co) compounds, which can cycle about 0.5 Li^+ ion per transition metal atom, and $LiMn_2O_4$ materials result in carbon-based rechargeable cells performing near the theoretical limit. The $LiM^{3+}O_2$ (M = Co and/or Ni) are compounds composed of MO_6 metal oxide octahedra connected at the edges to form single-layer MO_2 sheets (Fig. 6.5). Layers of these sheets are separated by interlayer sheets containing Li-ions octahedrally coordinated by oxygen anions contained in the MO_2 sheets. This results in two-dimensional (2D) pathways, which allow fast lithium-ion diffusion into and out of the structure. In addition, these materials have good electronic conduction, allowing electron charge transfer to the transition metal cations (Co or Ni). Upon Li intercalation and electron charge transfer from the external circuit, the tetravalent transition metal cations, $M^{4+}O_2$, are reduced to trivalent cations, $LiM^{3+}O_2$.

The $LiMn_2O_4$ structure differs from the layered $LiMO_2$ structure in that it contains MnO_6 octahedra connected to form a three-dimensional (3D) network (Fig. 6.6). This network creates 3D pathways from which Li-ions can be reversibly removed, but the gravimetric and volumetric energy densities are lower for $LiMn_2O_4$. However, Li-ion cells consisting of $LiMn_2O_4$ electrodes

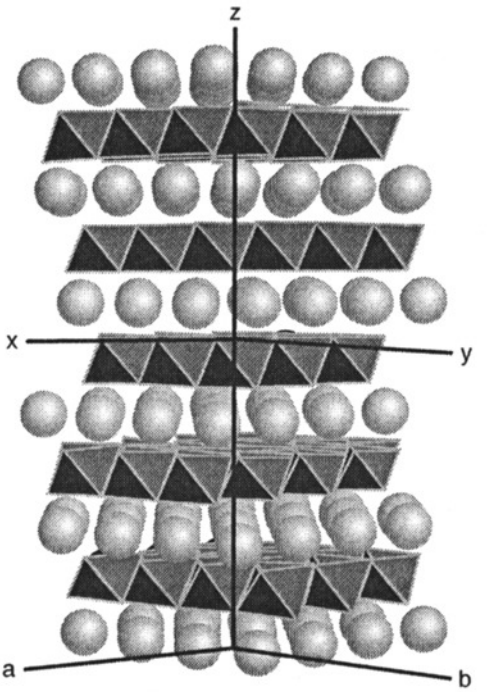


Fig. 6.5. The layered structure of $LiMO_2$ (M = Ni and/or Co) showing the 2D pathways for Li diffusion. Balls represent Li^+.

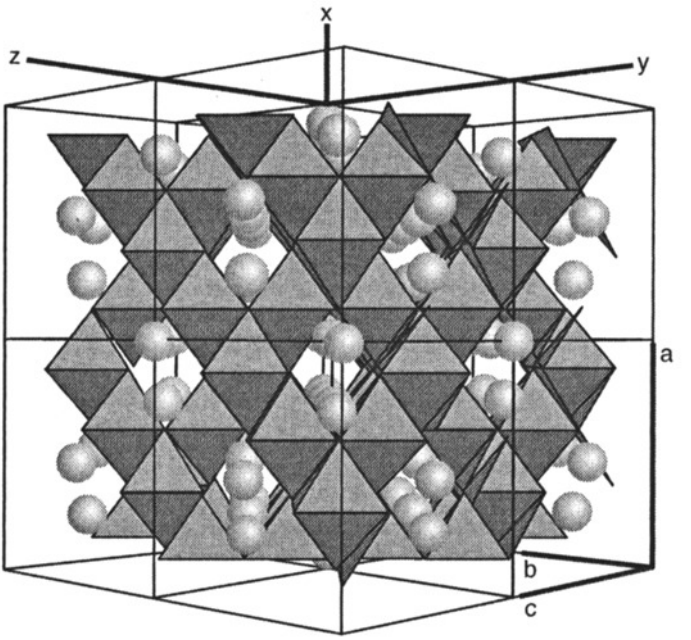


Fig. 6.6. The 3D spinel structure of the $LiMn_2O_4$ positive electrode material showing the 3D pathways for Li diffusion. Balls represent Li^+.

offer several advantages over the Li-ion cells based on Ni or Co. These include a lower electrode cost resulting from natural abundance and a lower cost for Mn in comparison with Co or Ni; considerably less-toxic manganese-based oxide materials, which have well-established recycling methods; and improved safety in the charged state offered by the 3D framework structure. The layered structures have a greater tendency to release oxygen under extreme-abuse conditions where internal temperatures approach 200°C.[36] Normally, oxygen is bonded tightly by the transition metals (M) in edge-shared MO_6 octrahedra, which form MO_2 layers. These MO_2 layers are stabilized by interlayer occupation of Li-ions. When the battery is charged and the Li-ions are removed, these layers can become unstable at extreme-abuse conditions (>200°C). When this happens, transition metals move into the interlayer space, resulting in a structural decomposition concurrent with the release of oxygen. The evolved oxygen in combination with the organic electrolyte creates a potential combustion hazard. The 3D framework structure of the $LiMn_2O_4$ spinel offers greater resistance to this evolution. Thus, based on its economy, safety, and environmental acceptability, this system is desirable for manufacturing. Although this material appeals to battery manufacturers, they are still hesitant in implementing it in commercial cells because of solubility issues with today's electrolytes.

6.3.2.2 Negative Electrode

The most popular negative electrode materials are based on carbon materials. Two major forms of carbon, coke and graphite, have been used in Li-ion batteries. Coke results from annealing carbon in inert atmospheres to temperatures of 1000°C. Graphite is formed after annealing carbons to temperatures approximately 2800°C. Graphite is a hexagonal, layered structure that allows 2D intercalation pathways of Li-ions much like that of the $LiMO_2$ positive electrodes. The output voltage (V^-) of graphite negative electrodes are satisfactory; however, improvement in capacity has been the main focus of latest research efforts. Recently, attempts to increase the capacity of the negative electrode have focused in two areas: (1) enhancing the electrochemical characteristics of the carbonaceous negative electrode and (2) finding alternative materials as a substitute for the carbonaceous negative electrode. Chemical or physical means have been used to improve the reversible capacity of the carbonaceous materials. For instance, several enhanced capacity electrode materials have been obtained either by means of a pyrolytic processing of organic materials[37] or by mechanical processing[38] (e.g., mechanical grinding) of the negative electrode material. These approaches, however, have a tendency to produce carbonaceous materials with large irreversible losses and low packing density, so that no significant improvements result when implemented in Li-ion cells. Yoshio[39] has recently reported that some lithiated vanadium oxide-based electrodes, when discharged to voltages lower than about 0.1 V, could reversibly intercalate Li-ions in amounts up to 7 lithium per transition metal so that capacities 2 to 3 times greater than those of graphites could be obtained. However the capacity retention is poor. This was later confirmed by Sigala *et al.*[40] In substituting oxides for graphites as the negative electrode in rechargeable Li-ion batteries, however, a price is paid in terms of cell output voltage since the average voltage at which these V-based oxides intercalate lithium is of 1.4 V compared with 0.3 V for graphite. The energy density of the Li-ion cell based on an oxide rather than graphite is the same within 5%.

6.3.2.3 The Electrolyte

An ideal nonaqueous electrolyte for a Li-ion cell should have the same requirements as for rechargeable Li metal cells:

- Low cost
- High ionic conductivity both at room temperature and at –20°C

- Chemically stable over a wide temperature range (–30 to 100°C)
- Demonstrated electrochemical stability over a wide range potential (0–5 V)

Recent electrolytes,[41] consisting of Li-PF_6 salts and a combination of solvents—e.g., ethylene carbonate (EC) with dimethyl carbonate (DMC)—satisfy most of the required conditions except low-temperature performance. However, adding a third solvent, such as propylene carbonate (PC) for example, allows these cells to function properly down to –20°C. The main drawback of today's electrolytes is their cost, which to a certain extent dictates the price of the Li-ion technology. The high price is a result of the stringent processing conditions to ensure that the salt and solvent are as dry as possible, since moisture contaminants will result in the formation of hydrogen fluoride (HF). HF has been shown to be detrimental to the performance of Li-ion batteries when present in the system in excess.

6.3.3 The Li-ion Battery: Assembly and Function

Like Ni-Cd batteries, Li-ion batteries are assembled in their discharged state and must be charged before use. Discharged Li-ion cells have a potential near zero, which greatly simplifies manufacturing and eliminates the possibility of accidental short-circuiting during assembly. For a Li-ion battery to function properly and efficiently, the mass of the positive and negative electrodes must be adjusted so that they have the same capacity. Cell imbalance may result in either Li plating or electrolyte decomposition. Balancing a Li-ion cell is complicated by the fact that each intercalation electrode exhibits differing amounts of irreversible capacity between the first discharge and first charge, and have different capacity fade rates as a function of cycle number and temperature.[19]

Performances of Li-ion batteries compared with other rechargeable technologies are shown in Table 6.1. These batteries possess higher specific energies and energy densities than Ni/Cd and Ni/MeH cells, and exhibit good rate capability and a long cycle life.

6.3.4 Space Allocation Powering Issues: Present Status

The Ni-Cd battery, well known for its high-power-rate capability and long-cycle life, was among the first technologies to be successfully used for space application purposes. More than 27,000 room-temperature cycles at a depth of discharge of about 25% have been achieved.[42] The system consists of nickel oxyhydroxide as the positive electrode; cadmium, the negative electrode; and potassium hydroxide, the electrolyte.

The output voltage of the cell is 1.3 V. Again, these cells are assembled in their discharged state and must be charged prior to their initial use. However, there are concerns with Ni-Cd cells:

- A weight penalty resulting in low specific energy (40 to 58 Wh/kg)
- Dendritic growth of the Cd at the negative electrode that can induce shorts
- Hydrolysis and degradation of the nylon membranes used to separate the electrodes in the battery that can lead to a serious decrease in the cell's lifetime
- Toxicity of Cd

Such concerns have prompted the study, design, and development of more efficient cells. These studies have led to improved Ni-Cd cells (cells that utilize a polypropylene rather than a nylon separator) but more importantly to the development and optimization of the Ni/H_2 batteries that are used aboard GEO and LEO satellites (Hubble space telescope).

Such Ni/H_2 cells[43] are similar to the Ni-Cd cell only in that they still contain the same positive electrode, the negative electrode being hydrogen gas. This technology is superior to the Ni-Cd in terms of energy density (50–80 Wh/kg) and cycle-life (for instance, 4000 cycles with a 75% depth of discharge compared with 800 for the Ni-Cd). Recently, an outstanding performance of 40,000

Table 6.1. Performance Comparison of a Variety of Rechargeable Battery Technologies

Quantity	Ni-Cd	Ni-MeH	$LiCoO_2$/C Li-ion (liquid)	Bellcore Liquid Li-ion	Bellcore Plastic Li-ion	Lead-acid
Average voltage (V)	1.2	1.2	3.6	3.8	3.8	2.0
Energy (Wh/kg)	58	70	90–130	100	100–140	30
Capacity (Ah/kg)	48	58	25–36	27	26–37	15
Cycle life	1000+	500	1000	1000	1000	200
Average cost ($/Wh)	0.88	1.50	2.00	1.20[a]	0.50[a]	0.30
Operating temperature range	varies within −20 and 50°C	varies within −20 and 40°C	−20⇔60°C	−20⇔60°C	−20⇔60°C	varies within −20 and 50°C
Toxicity	yes	yes	yes	no	no	yes
Memory effect	yes	varies with processing	no	no	no	no

[a] Estimate based on material costs.

cycles with 40% depth of discharge was reported for Ni/H_2 batteries.[44] Finally, the Ni/H_2 technology is self-protected against accidental overcharge or overdischarge. The oxygen produced during overcharge at the positive electrode reacts with the hydrogen to give water. With this process, the Ni/H_2 batteries can tolerate extended overcharges 4 to 5 times greater than for the Ni-Cd technology.[45] In addition, Ni/H_2 cells exhibit outstanding overdischarge protection.

These cells are preloaded in hydrogen (e.g., there is always a net hydrogen pressure even in the discharged state). During overdischarge the following scenario occurs.[46] After the electrochemically active NiOOH positive electrode is completely reduced to $Ni(OH)_2$ (e.g., cell completely discharged), water is reduced on this electrode, resulting in hydrogen. This hydrogen is then reoxidized to give water at the negative electrode. The overall result is that during overdischarge the pressure and the water content within the cell remain constant, since once water is reduced at the positive electrode, water is recombined at the negative electrode. On the contrary, in Ni-Cd technology, an overdischarge (e.g., cell reversal) will lead to a significant internal cell pressure. Finally, another important advantage of the Ni/H_2 technology over the Ni-Cd technology is that the hydrogen pressure can be used as an indicator of the state of charge of these batteries. Consequently, the present Ni/H_2 technology, although significantly more expensive than the Ni-Cd technology, offers a greater transparency to harsh operating conditions. This is why the lifetime of Ni/H_2 batteries can exceed 5 years under extreme conditions.

The Ni/H_2 technology is, however, not the ultimate solution to space applications because of a major drawback of extremely high self-discharge. Indeed, a Ni/H_2 battery can lose as much as 50% of its initial capacity at 20°C in only 10 days. The direct interaction of hydrogen with the fully oxidized positive electrode material is at the origin of such a high self-discharge, and it has

been demonstrated that the self-discharge rate varies linearly with the hydrogen pressure (e.g., the more charged the Ni/H_2 cell, the larger the self-discharge).[47] A possible solution to this issue is to lower the hydrogen pressure within the cell (e.g., increasing the cell volume); however, since this can only be done at the expense of lowering the volumetric energy density, such a solution was not pursued.

Another tentative solution regarding this hydrogen pressure has led to the recently emerging low-pressure Ni-MeH technology that is supplanting the Ni-Cd technology for portable consumer applications.[48] Such cells are similar to the Ni-Cd cell, the main difference being that the cadmium electrode has been replaced with a "hydrogen electrode" in which the hydrogen is stored as a medium-cost metal-hydride. Current hydrogen storage materials (i.e., hydrogen "sponges") such as $LaNi_5$ can have hydrogen densities greater than that of liquid hydrogen. Ni-MeH cells have up to twice the capacity of Ni-Cd batteries, and significantly higher energy density per unit of volume. However, while these cells show less self-discharge than the Ni/H_2 ones and greater than the Ni-Cd ones (consistent with the kinetics of self-discharge being linked to hydrogen pressure), they exhibit limited cycle-life performance owing to the lack of alloys that exhibit perfect stability against corrosion in KOH media used in the electrolyte solution. Improved hydrogen-absorbing alloys with anticorrosion characteristics must be developed, either through new chemistry or new processing and manufacturing methods in order to improve the cycle life of the Ni-MeH technology before they can be seriously considered for space applications.

Although current Li-ion technology offers significant advantages in performance over the nickel-hydrogen technologies discussed above, there are intrinsic differences between the two technologies, which if not properly accounted for, can result in cell degradation. All nickel-hydrogen technologies make use of an aqueous-based electrolyte. In abuse conditions (overcharge, high-temperature storage, etc.), electrolyte breakdown (water decomposition) will occur, leading to high rates of self-discharge. Recombination reactions, however, make the electrolyte breakdown reversible, and performance can be largely recovered. This is not the case in Li-ion technology. Because of the large range of voltages used in Li-ion, lithium salts dissolved in non-aqueous organic solvents must be used. Under abuse conditions like those discussed in the nickel-hydrogen aqueous technologies, the non-aqueous electrolyte will also undergo decomposition, although to a much smaller degree. Unlike the aqueous technologies, no recombination mechanism exists, and all electrolyte breakdown is irreversible; therefore, although the Li-ion technology has performance advantages, care must be taken not to subject cells to long-term abuse. In light of this, care must be taken to keep long-term operation and storage temperatures as close to 20°C as possible in order to retain extended lifetimes on the order of 10 years. Such cells can operate well in the temperature region near 60°C for shorter times as commonly required by consumer electronics manufacturers. The 20°C temperature range has not been very difficult to maintain in traditional satellite design, but may pose greater difficulty in nanosatellite designs relying solely on passive thermal management schemes.

Attention must also be given to the implementation of good monitoring and charge controls, especially critical in Li-ion technologies using the layered positive electrode compounds of $LiCoO_2$ and $LiNiO_2$. In normal operation, approximately 0.5 Li is reversibly removed from these electrode compounds. If the cells are charged in excess (removing greater quantities of Li), this will have a deleterious effect on capacity retention. In addition, balance between the positive and negative electrode capacities will be altered. The excess Li will have no sites available for intercalation in the negative electrode and will plate, causing the formation of Li metal, which will cause a reduction of cell performance and safety. For the $LiMn_2O_4$ positive electrode, all the Li is always removed, so those dangers listed above for the layered compound are not as extreme.

However, even in the $LiMn_2O_4$ case, overcharge for extended periods will cause irreversible electrolyte oxidation and must be controlled. In light of these special considerations, considerable effort has already been expended to make reliable, low-cost monitoring and charge control electronics. This is particularly important when charging groups of cells are placed in series.

In commercial Li-ion batteries, many electronic monitors and controls of the charge/discharge cycles are built into the multicell battery pack. Most circuits include both electronic charge and discharge protection. Placing individual cells in series increases the difficulties with charge control. Although Li-ion battery cells are not charged individually, constant monitoring of each series element is incorporated to set the charge and discharge conditions relative to the weakest cell. This type of control eliminates undue stress on the remaining cells.

Determining the state of charge is an important factor in evaluating completion of the charge and discharge cycles and the overall health of the battery. Such determinations can be evaluated by monitoring both current and voltage. Current is usually monitored and integrated to calculate the coulombs of charge remaining in the battery, generally accomplished through the use of a low-resistance (0.05 Ω) sense resistor. Depending on the active electrodes used in their manufacture, Li-ion batteries can be designed to have either a sloping or a relatively flat discharge profile. At present, this is largely determined by the choice of negative electrode material (Fig. 6.7). For example, the use of graphite as a negative electrode gives a very flat discharge profile, where the voltage changes very little as the cell capacity is consumed during discharge. The flat discharge profile offers an advantage in overall higher energy density of the battery (retains consistently high V_{cell}) and allows a relatively constant current during the full discharge of the Li-ion cell when powering devices that work on a constant power basis. In the graphite case, the current does not have to increase to offset the loss in voltage that would be experienced in the sloping discharge profile for coke to retain a constant power output. However, the flat discharge profile makes charge status determination on the basis of voltage difficult. In the use of coke or "hard-carbon"-

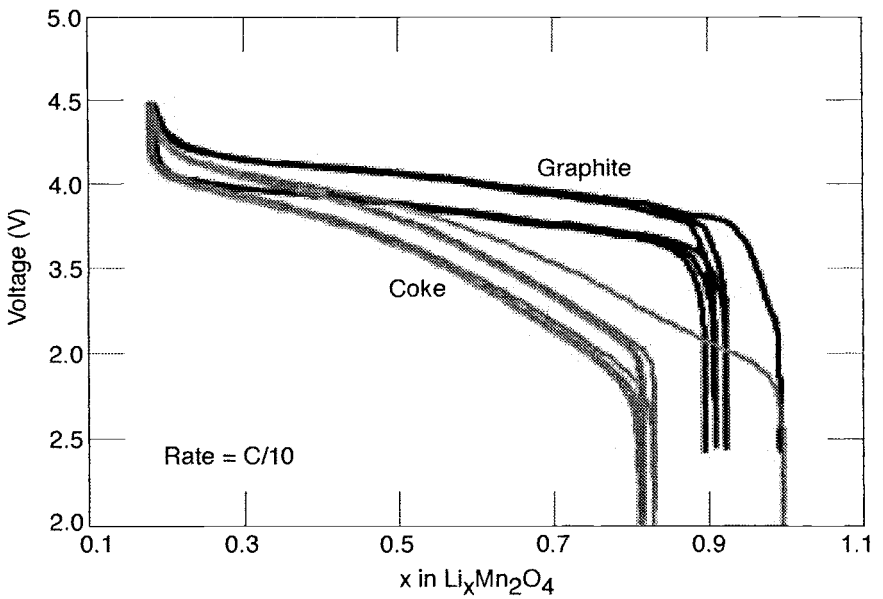


Fig. 6.7. Difference in voltage profiles vs Li content for graphite and coke-based carbon anodes in $LiMn_2O_4$ based Li-ion cells.

based anodes (such as Sony's), voltage profiles slope steeply, and it is much easier to make a determination of the charge status on a voltage basis alone.

In summary, the emerging rechargeable Li-ion technology by no means meets all requirements for powering a satellite, but seems to be the one, when compared with Ni/Cd, Ni-MeH and Ni/H_2, closest to the "wish list" for space applications outlined at the beginning of this chapter. The suitability of the Li-ion technology for nanosatellite applications is being explored by AeroAstro Corporation. AeroAstro originally developed the Bitsy nanosatellite for the U.S. Air Force and is now making it available for commercial applications. Bitsy is a fast turnaround satellite that can be designed and ready for flight within a few months from order, to be used in areas such as remote sensing, communications, space science, and astrophysics. The satellite is highly integrated and utilizes a single circuit-board design. It is in a very small 1-kg package of 15 × 15 × 5 cm. Thermal control is completely passive, although the option to incorporate active control may be provided. The powering system is based on an 8-V power bus. It is the first nanosatellite designed to use liquid Li-ion cells. These cells will offer 4 Wh of energy.

6.3.5 Projections for Future Development

Batteries with greater capacity for increasing mobility during use are in constant demand by consumers. Such demand has generated, and will continue to generate, much research toward the development of new materials. The satellite community will undoubtedly benefit from these efforts. Capacity improvements within a battery technology can result from improvement either in the chemistry of the system (e.g, better materials) or in its engineering (e.g., electrode forming and packing). Usually, engineering improvements are incremental for mature technologies such as Ni/Cd; whereas, material improvements can be drastic for emerging battery technologies. This is especially true for the Li-ion technology that incorporates materials for which only 50% of the total capacity is realized. With the surge of activities toward searching for better electrode materials, the discovery of a cathode that could intercalate/deintercalate one Li per transition metal (Mn, Co, Ni, etc.) must happen. Several directions are being pursued. Among them, one reported by Goodenough,[49] consists of adjusting the ionic/covalent character of the M-X bond by replacing the anion with a polyanion. These materials can now exchange one Li per transition metal at a potential lower than 4 V. However, volumetric capacity (capacity with respect to volume) improvements are less than the gravimetric (with respect to weight), and the diffusion being so slow, such materials cannot at this time be conceived for energy-hungry applications.

For the negative electrode, a common goal is to find materials with a slightly greater intercalation voltage (to completely eliminate the risks of Li plating) and a larger electroactive capacity than present-day carbons. Fuji recently disclosed the fruits of their long research effort employing metal-oxide: the discovery of a negative electrode based on Sn-Si-O, which has an intercalation voltage of about 0.5 V and a capacity twice that of carbon.[50,51] Li-ion cells with 50% improvement in the overall cell capacity were claimed by Fuji, which announced the production of such a new Li-ion cell by the end of 1998.

Li-ion technology, while the most attractive, is still in its infancy, and major breakthroughs in materials have yet to come. It is left to the creativity and innovation of the solid-state chemists designing and elaborating new intercalation electrodes to ensure that such an advantage endures over the next decades.

One recent advance that addresses the Li-ion battery as a whole is the plastic Li-ion battery. In addition to providing the performance of Li-ion liquid technology, the plastic Li-ion battery adds shape and performance design flexibility, enhances safety, and offers resistance to vibrations.

6.4 Plastic Li-ion Batteries

Current liquid-electrolyte Li-ion technology does not meet shape flexibility requirements dictated by portable electronics, and is limited to cylindrical or prismatic shapes. Rechargeable polymer Li batteries could provide shape flexibility, but after 20 years of research and development, the rechargeable polymer lithium battery technology still cannot function efficiently at room temperature, mostly because of the lack of polymer electrolytes with sufficient ionic conductivity.

Midway between rechargeable lithium batteries using liquid and those using pure polymer electrolytes are those using the hybrid electrolytes, for example, polymers swollen in liquid electrolytes (Fig. 6.8). These batteries combine the advantages of both liquid (high-power rate) and polymer electrolyte (no electrolyte leakage, easier scaleability) batteries. Reemerging during the last several years, the hybrid electrolyte concept was demonstrated by Feuillade and others[52] as early as 1974. In addition, the primary dry alkaline cells contain a liquid electrolyte that is "immobilized" in an elastic matrix.

The principle of electrolyte immobilization has also been applied to the recently popularized rechargeable lead-acid batteries known as VRLAs (valve regulated lead-acid batteries) in which the electrolyte is either in the form of a gel or absorbed in a silica mat separator. In the last several years, this concept has been widely applied in the field of rechargeable Li batteries to prepare a solid polymer electrolyte battery in which the electrolyte is immobilized in an appropriate polymer matrix. Hybrid polymer electrolyte films are usually made by dissolving a polymer matrix into a low boiling solvent (acetonitrile, tetrahydrofuran, etc.) together with a non-aqueous Li-salt electrolyte. The most popular polymer matrices[53] are polyethylene oxide (PEO) and its derivatives, such as polyacrylonitrile (PAN). Depending on the voltage range, a variety of liquid electrolytes were tried, with propylene carbonate (PC)-ethylene carbonate (EC)/Li-based salts being the most popular. The resulting viscous solution, consisting of the polymer matrix, low boiling solvent, and liquid electrolyte, is cast, usually resulting in tacky and mechanically weak films.

While many hybrid electrolytes reported in the literature exhibit high ionic conductivities, most exhibit various deficiencies preventing them from being used in practical cells. For example, their mechanical properties are often very poor, and the films must be hardened by either chemical or physical (high-energy radiation) curing. Besides the need for cross-linking, the main drawback

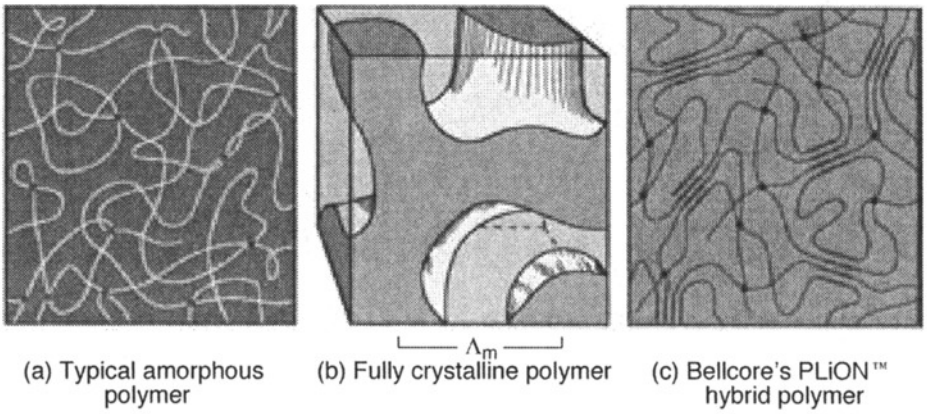


Fig. 6.8. Representation of three different approaches toward the polymer incorporation in Li-ion batteries: (a) amorphous swollen polymer; (b) fully crystalline polymer, which does not entrap the electrolyte; (c) Bellcore hybrid polymer, which combines the strength of the pure polymer and the electrolyte entrapment of the amorphous polymer.

of the above process is that it has to be carried out in a completely moisture-free atmosphere because the moisture-sensitive Li-salt is present at the initial stage. The manufacturing of such batteries requires advanced technologies for providing a low-humidity environment, thereby resulting in high processing costs.

Aiming to combine the recent commercial success enjoyed by liquid Li-ion batteries with the manufacturing advantages presented by polymer technology, Bellcore researchers introduced polymeric electrolytes in the liquid Li-ion system and developed the first reliable and practical rechargeable Li-ion plastic battery (Fig. 6. 9).[54-56]

A plastic is commonly defined as a combination of various material components, such as polymer matrix, plasticizers, fillers, stabilizers, and so on. These components are chosen in proportions, depending on the targeted application, to produce solid but elastic structures. Bellcore's plastic electrolyte complies with the above definition, the liquid electrolyte acting as the plasticizer. The correct choice of a polymer matrix is very important in the cell's manufacturability and longevity under extraneous operating conditions. The PLiON™ (Plastic Li-iON) battery uses a poly (vinylidene fluoride)-hexafluoropropylene (PVDF-HFP) copolymer (8–18% HFP). The Li^+ conduction is carried out by the liquid electrolyte embedded in the separator and the electrodes. The polymer remains chemically and electrochemically inactive. The polymer used as the matrix possesses several important characteristics:

- Chemically inert, with respect to electrode active materials and components of the plasticizer
- Electrochemically stable within the voltage range of interest (0–5 V vs Li)
- High melting/softening point, providing robustness in the final product
- Good mechanical stability for ease of processing and battery manufacturing
- Sufficient liquid electrolyte to provide good ionic conductivity
- Commercially available in large quantities at a low cost

PVDF-HFP copolymer consists of mixed amorphous and crystalline phases. Amorphous regions hold large amounts of liquid for high ionic conductivity, and crystalline domain provides the mechanical strength in the polymer film. Having the same polymer matrix throughout all the cell layers allows the fusion of the layers (Figs. 6.10 and 6.11) when laminated via heat and pressure, adding an unprecedented robustness to the cells, thereby eliminating the need for cross-linking. In addition, because of the fusion of the electrodes and separator, external stack pressure is not needed to ensure contact between the layers. Therefore, the cells can be assembled and packaged in a thin and soft metal laminate, providing a hermetic seal.[57]

Fabrication of a typical cell is as follows. The positive electrode composite is prepared by casting a slurry consisting of the transition metal positive electrode, conductive carbon black, PVDF-

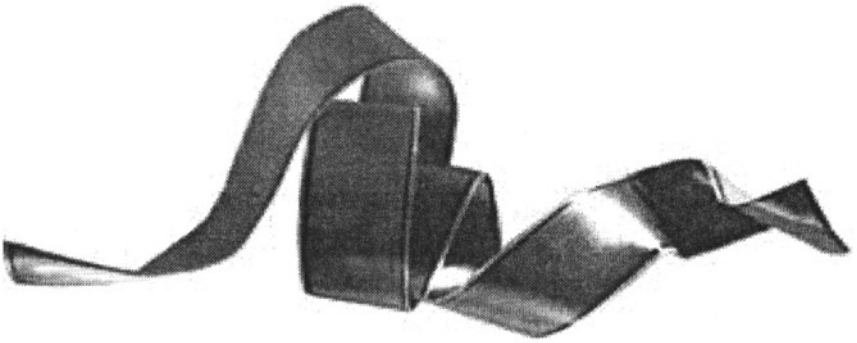


Fig. 6.9. The unpackaged PLiON™ battery.

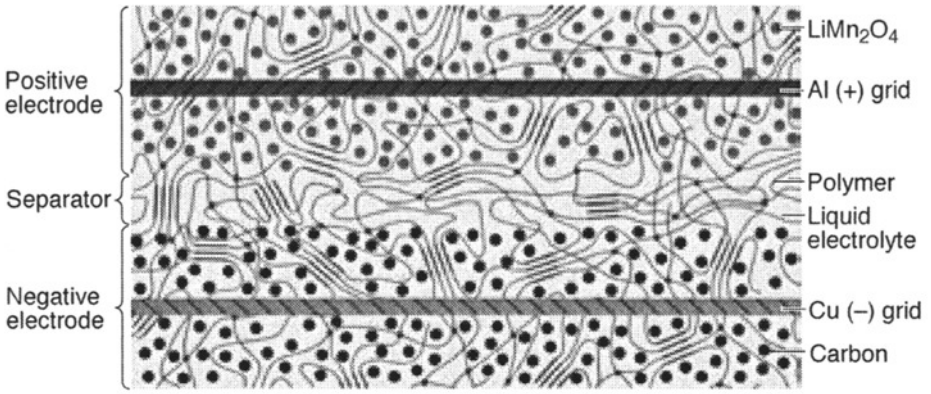


Fig. 6.10. Cross section of the PLiON™ battery. PVDF-HFP polymer and liquid electrolyte are dispersed throughout the entire battery assembly and are represented by lines and gray field, respectively.

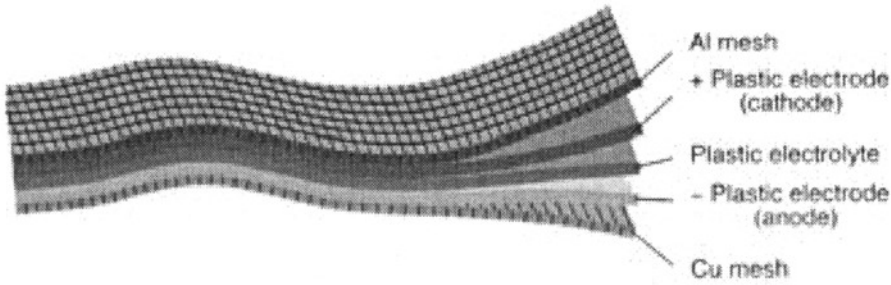


Fig. 6.11. The PLiON™ battery laminate showing the individual layers, which are laminated into one interface free structure.

HFP copolymer, dibutyle pthalate (DBP) plasticizer, and suitable solvent. The slurry is dried, and the resulting tape is laminated to an aluminum positive electrode grid. The negative electrode composite is prepared in very much the same way, except the transition metal positive electrode is replaced with a carbon-based negative electrode. The resulting negative tape is then laminated to a copper grid current collector. A separator is also cast, consisting of the PVDF-HFP copolymer, a filler, and DBP.

The positive laminate, the separator, and the negative laminate are then laminated between hot rollers to produce a bonded cell assembly with a continuous matrix of PVDF-HFP copolymer. The laminates can be made in any size, thickness, or shape. The completed cell is then placed in a bath of solvent to extract the DBP, which leaves a very fine continuous porosity (20 nm) throughout the polymer. Such a process lends itself to continuous manufacturing (Fig. 6.12).

All the fabrication steps to this point, unlike traditional plastic Li-ion batteries, are done under ambient conditions. To bypass the burden of assembling the cell in a moisture-free environment, the plastic laminate is activated at the very last stage through an extraction/activation step. The last step, cell activation, is performed in an atmosphere-controlled dry box. In this step the nonaqueous electrolyte is introduced into the cell. Since the porosity is very fine, capillary action soaks in the electrolyte without the need for degassing. The most important aspect of the PLiON™ battery is that the plastic laminate does not need to be made in a moisture-free environment. The only fabrication steps that require careful moisture control are the activation and packaging when

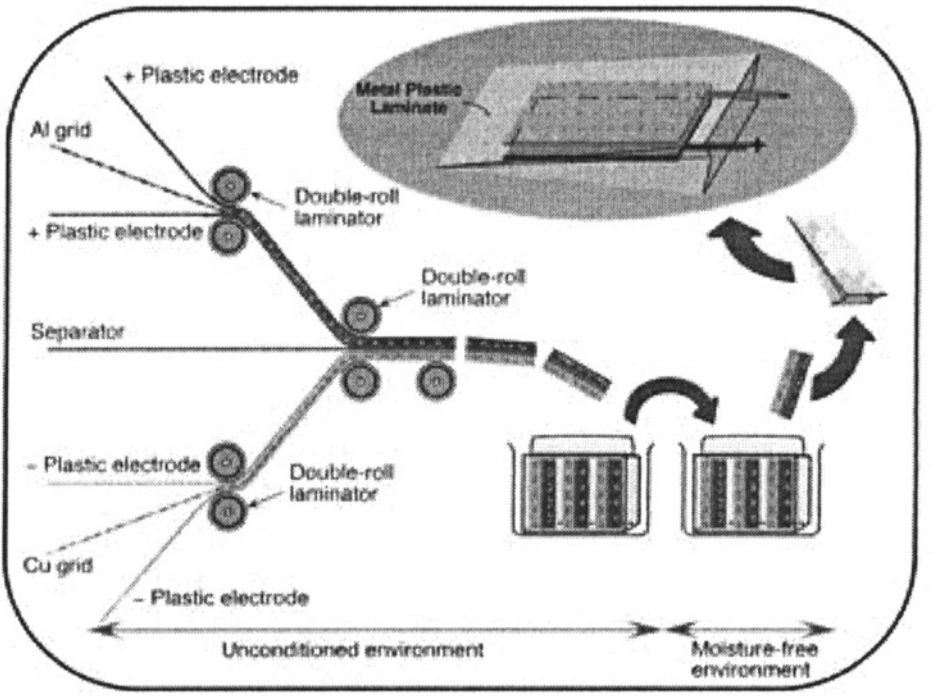


Fig. 6.12. Graphical representation of the PLiON™ battery fabrication process from initial positive electrode, negative electrode, and separator tapes to activated and packaged cell.

the moisture-sensitive lithium salt is present. It is important to remember that the PLiON™ battery performance is dictated by the Li-ion chemistry that is entrapped within the polymer matrix. Any combination of liquid electrolyte, positive electrode, and negative electrodes can be used within the PLiON™ battery. Figure 6.13 gives an example of the flexibility in chemistry by showing the performance of the PLiON™ battery using a number of different positive electrode materials.

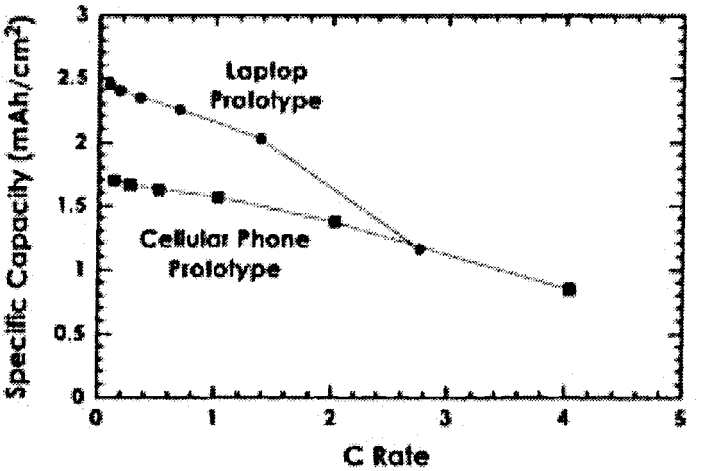


Fig. 6.13. Specific capacity vs C rate for PLiON™ batteries fabricated with thick and thin electrodes, showing the intrinsic performance flexibility of the technology based on application.

A unique and important feature of PLiON™ cells is that they can be designed for specific applications that have specific rate requirements. In a hybrid system, such as the PLiON™ battery, separator thickness plays a minor role in determining how fast the cells can charge or discharge (Fig. 6.14). The electrode thickness defines the initial cell capacity as well as the rate capability of each cell; this is a component that cannot be varied too much of a degree in today's liquid Li-ion technology. Therefore, PLiON™ cells can be designed to either optimize energy densities or rate capabilities, depending on the application.

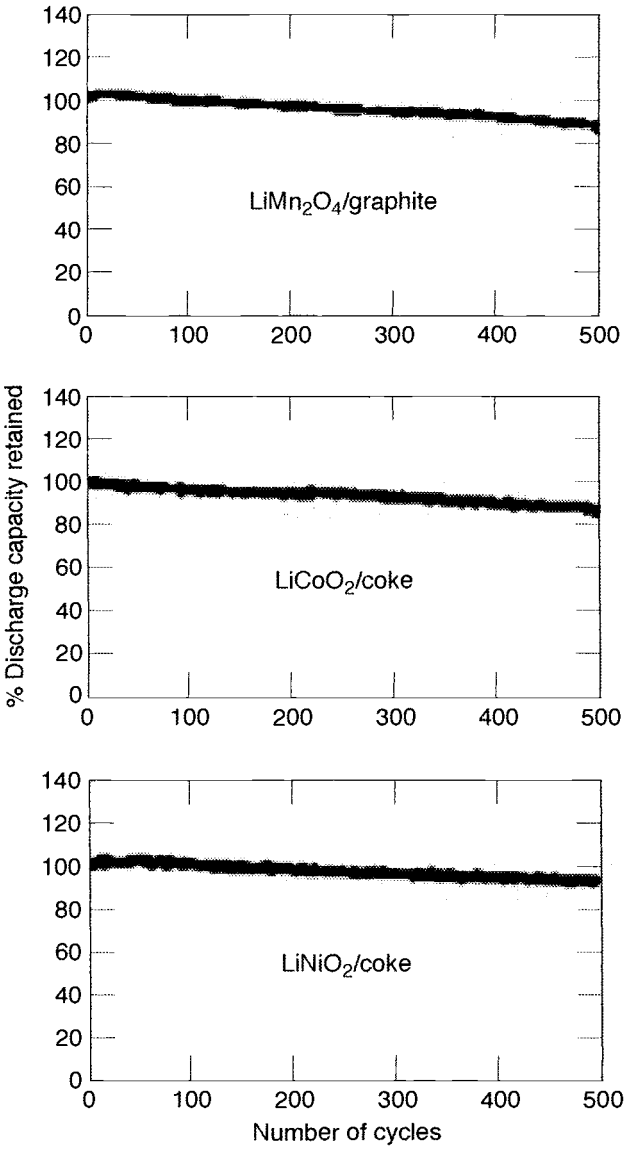


Fig. 6.14. Percent capacity as a function of cycle number showing the transparency of PLiON™ technology toward the incorporation of various positive electrode materials. PLiON™ technology is electrochemically inert and allows the incorporation of any variation of Li-ion chemistry.

PLiON™ technology combines the performance of the Li-ion batteries, which contain free liquid electrolyte, with the advantages of a polymer matrix. These batteries can be fabricated as flexible sheets or ribbons of different capacity of dimensions. The cells are thin, flat, flexible, shock- and vibration-proof. Unlike Li-ion batteries based on liquid technology, the main components of the PLiON™ battery are immobile and laminated into one solid composite. This offers advantage in situations involving large quantities of mechanically induced vibrations such as those developed during launch. The composite will be less prone to internal electrical disconnect and delamination. PLiON™ battery has high energy density, low weight, low self-discharge, long-cycle life both at room temperature and 55°C, and is environmentally benign (see Table 6.1 in Subsec. 8.3.3). In addition, these cells have low internal resistance, demonstrating low voltage drop at pulsed discharges of up to 3 C rate, maximizing the usable capacity at high rates. This performance is compared with that of high-rate NiCd technology in Fig. 6.15 for a consumer portable-phone application. In addition, the figures of merit of the plastic technology with its liquid counterparts and the Ni-based rechargeable systems are summarized in Table 6.1.

6.4.1 The Plastic Battery and the Nanosatellite

Since Li-ion technology has not been proved during tens of thousands of cycles and publicized accelerated testing results are basically nonexistent, its implementation in space applications must first be targeted to GEO applications because of the limited number of eclipses and the relatively low number of cycles required from the battery.

The real advantage of the plastic technology resides in its design flexibility based on a plate type architecture. Current Li-ion technology relies on jellyrolling the three main components: positive and negative electrode and a separator. After this jellyroll is rolled under tension, the cell is placed into a cylindrical can, evacuated, and then filled with the non-aqueous electrolyte. Another derivative of the jellyroll is found in the liquid-based prismatic cells. In the construction of these small "brick-like" cells, an oval jellyroll is fabricated and fitted into prismatic cell housings.

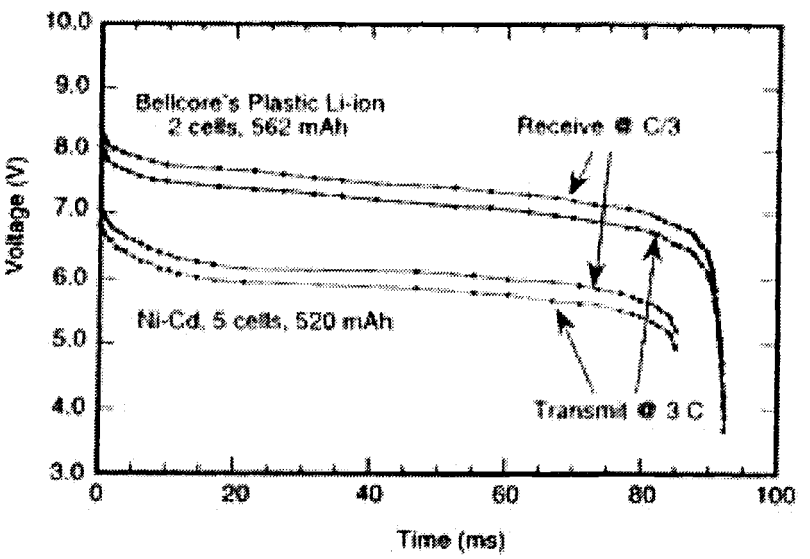


Fig. 6.15. Comparison of 3 C pulse rate performance of comparable PLiON™ and NiCd batteries.

Although such cells pack efficiently in a multiple cell configuration, the active components do not pack very efficiently inside the prismatic cell, as they are ovally wound jellyrolls inside rectangular, prismatic cans. The plate-like PLiON™ technology allows excellent packing efficiency, since multiple plates can be densely packaged in parallel within one cell. In addition, these multiple cells can then be placed in series or in parallel.

In specialty applications such as the nanosatellite, the plastic Li-ion battery can be designed in custom shapes that would maximize its use efficiency in a satellite of such design density. If powering requirements for the nanosatellite do not require voltages in excess of 4 V, single cells can be used in the application of Li-ion to the satellite, greatly simplifying the electronics and cell construction. The placement of such cells could be extreme, such as under solar cells, depending on the thermal conditions on which the cell would be exposed. Although these cells offer excellent design flexibility and performance, care must be taken to ensure proper environmental controls.

One of the first nanosatellites for planned launch is the ASUSat1, designed and developed by Arizona State University with the Aerospace Research Center.[58] The 10-lb nanosatellite will have a LEO and perform Earth imagery and AMSAT operations over its lifetime of 2 to 4 years. NiCd is the battery technology. Two battery packs are used; each consisting of six cells, presumably set in series to give an output voltage of 7.4 V. These cells are mounted in a rectangular foam box and mounted to an inside panel.

Using this as a case study, the adoption of the Li-ion plastic technology could result in significant advantages. Li-ion cells would occupy half the volume and one-third the weight of their Ni/Cd counterparts. This would free valuable space and weight for the implementation of additional analytical equipment in support of the satellite mission. Instead of having six individual Ni/Cd cells in series, which requires careful cell matching, the Li-ion cells would be able to deliver the required voltage using two cells. The Li-ion would not require any conditioning after partial discharge, as this technology exhibits no memory effect. The plastic cells offer exceptional design flexibility, allowing the most effective and efficient use of the volume of the satellite, as these cells can be manufactured in almost any 2D shape. The design flexibility offers an added benefit in thermal management. Passive thermal management is almost always maximized in satellite design, which is especially true in nanosatellite technology where volume and weight are at a premium. Therefore, the cell plates can be arranged and packaged in a material that would facilitate thermal management to a greater degree than traditional battery technology. Plates can be arranged so that thermal conduits can be placed between the high surface area stacks. These conduits can then eventually lead to a heat sink or exchanger.

The application of the PLiON™ battery also goes beyond the nanosatellite proper in support of the general nanosatellite program. One of the intents of the nanosatellite is to be independently controlled through remote ground links with devices such as personal laptops. The PLiON™ technology can be easily extended to these systems to offer lightweight, high-energy density powering systems. In fact such laptop batteries can be expected to enter the consumer market by early 1998. The cells will offer the performance advantages intrinsic to the Li-ion chemistry, with the added design flexibility of plastic.

6.5 Conclusion

Li-ion rechargeable battery technology possesses many beneficial attributes with respect to satellite powering issues. Although not perfect, Li-ion technology is superior to the current aqueous nickel technologies in both gravimetric and volumetric energy densities. In addition, this battery chemistry is relatively young and, in time, will show significant improvements in energy density.

Li-ion battery technology coupled with plastic Li-ion configuration (PLiON™) allows unprecedented flexibility in both physical and performance design. Such benefits in terms of weight, size, and design flexibility have much potential for design engineers striving to develop efficient, economical micro and nanosatellite technologies.

6.6 Acknowledgments

Drs. N. Sac-Epee and A. Delahaye are warmly acknowledged for their comments and helpful discussions about this review.

6.7 References

1. P. P. K. Chetty, *Satellite Technology and Its Applications*, 2nd ed. (TAB Professional and Reference Books, Blue Ridge Summit, PA, 1991).
2. D. W. Murphy and P. A. Christian, "Solid State Electrodes for High Energy Batteries," *Science* **205**, 651–656 (1979).
3. D. Linden, ed., *Handbook of Batteries*, 2nd ed. (McGraw-Hill, 1995)
4. J. P. Gabano, ed., in *Lithium Batteries* (Academic Press, London, 1983), pp. 1–11.
5. M. Armand, "Intercalation Electrodes," in *Materials for Advanced Batteries*, edited by D. W. Murphy, J. Broadhead, and B. C. H. Steele (Plenum, New York, 1980), pp. 145–161.
6. B. C. H. Steele, "Materials Requirements for High Performance Batteries," *Electric Vehicle Developments* (Wolfson Unit for Solid State Ionics, London, UK, March 1979), pp. 10–12.
7. W. R. McKinnon and R. E. Haering, "Physical Mechanisms of Intercalation," in *Modern Aspects of Electrochemistry*, no. 15, edited by R. E. White, J. M. Bockris and B. E. Conway (Plenum, New York, 1983), pp. 235–304.
8. B. C. H. Steele, "Characterisation and Performance of Solid Solution Electrodes," *Proceedings of the Meeting on Prospects for Battery Applications and Subsequent R& D Requirements* (Brussels, 17 Jan. 1979), pp. 201–221.
9. J. Dudley, D. Wilkinson, G. Thomas, *et al.*, "Conductivity of Electrolytes for Rechargeable Lithium Batteries," *J. of Power Sources* **35**, 59–82 (1991).
10. M. Armand, J. M. Chabagno, and M. J. Duclot, "Poly-ethers as Solid Electrolytes," in *Fast Ion Transport in Solids*, edited by P. Vashishta, J. M. Mundy and G. K. Shenoy (North-Holland, Amsterdam, 1979), pp. 131–136.
11. Y. Choquette, M. Gauthier, A. Belanger and B. Kapfer, *Sixth Intern. Meeting on Lithium Batteries*, Muenster, Germany, 10-15 May 1992, abstract TUE 06.
12. D. W. Murphy, F. J. DiSalvo, J. N. Carides and J. V. Waszczak, "Topochemical Reactions of Rutile Related Structures with Lithium," *Mater. Res. Bull.* **13**, 1395–1402 (1978).
13. M. Lazzari and B. Scrosati, "A Cyclable Lithium Organic Electrolyte Cell Based on Two Intercalation Electrodes," *Mater. Res. Bull.* **127**, 773–774 (1980).
14. K. Mizushima, P. C. Jones, P. J. Wiseman, and J. B. Goodenough, "Li//xCo0//2 (0 < x less than equivalent to 1/): A New Cathode Material for Batteries of High Energy Density," *Mater. Res. Bull.* **15**, 783–789 (1980).
15. A. D. Little, "Power '94," Santa Clara, CA, 1994.
16. D. Guyomard and J. M. Tarascon, "The Carbon $Li_{1+x}Mn_2O_4$ System," *Solid State Ionics* **69**, 222–237 (1994).
17. T. Nagaura and K. Tozawa, "Lithium Ion Rechargeable Battery," *Progress in Batteries and Solar Cells* **9**, 209–217 (1990).
18. J. R. Dahn, U. Von Sacken, M. R. Juzkow and H. Al-Janaby, "Rechargeable $LiNiO_2$/Carbon Cells," *J. Electrochem. Soc.* **138**, 2207–2211 (1991).
19. D. Guyomard and J. M. Tarascon, "Li Metal-Free Rechargeable $LiMn_2O_4$/Carbon Cells: Their Understanding and Optimization," *J. Electrochem. Soc.* **139**, 937–948 (1992).

20. K. W. Semkow and A. F. Sammells, "Secondary Solid-State SPE Cells," *J. Electrochem. Soc.* **134**, 766–767 (1987).
21. M. Lazzari and B. Scrosati, "A Cyclable Lithium Organic Electrolyte Cell Based on Two Intercalation Electrodes," *J. Electrochem. Soc.* **127**, 773–774 (1980).
22. J. M. Tarascon, D. Guyomard, and G. L. Baker, "An Update of the Li Metal-Free Rechargeable Battery Based on $Li_{1+x}Mn_2O_4$ Cathodes and Carbon Anodes," *J. Power Sources* **43–44**, 689–700 (1993).
23. B. di Pietro, M. Patriarca and B. Scrosati, "On the Use of Rocking Chair Configurations for Cyclable Lithium Organic Electrolyte Batteries," *J. Power Sources* **8**, 289–299, (1982).
24. J. M. Tarascon, "Mo_6Se_6: A New Solid-State Electrode for Secondary Lithium Batteries," *J. Electrochem. Soc.* **132**, 2089–2093 (1985).
25. J. J. Auborn and Y. L. Barberio, "Lithium Intercalation Cells Without Metallic Lithium," *J. Electrochem. Soc.* **134**, 638–641 (1987).
26. E. J. Plichta, W.K. Behl, D. Vujic, W. H. S. Chang, and D. M. Schleich, "The Rechargeable Li_xTiS_2/ $LiAlCl_4/Li_{1-x}/CoO_2$/ Solid-State Cell," *J. Electrochem. Soc.* **139**, 1509–1513 (1992).
27. R. Kanno, Y. Takeda, T. Ichikawa, K. Nakanishi, and O. Yamamoto, "Carbon as Negative Electrodes in Lithium Secondary Cells," *J. Power Sources* **26**, 535–543 (1989).
28. M. Mohri, N. Yanagisawa, Y. Tajima, H. Tanaka, T. Mitate, *et al.*, "Rechargeable Lithium Battery Based on Pyrolytic Carbon as a Negative Electrode," *J. Power Sources* **26**, 545–551 (1989).
29. T. Ohzuku, M. Kitagawa and T. Hirai, "Electrochemistry of Manganese Dioxide in Lithium Nonaqueous Cell," *J. Electrochem. Soc.* **137**, 769–775 (1990).
30. J. M. Tarascon and D. Guyomard, "Li Metal-Free Rechargeable Batteries Based on $Li_{1+x}Mn_2O_4$ Cathodes ($0 \leq x \leq 1$) and Carbon Anodes," *J. Electrochem. Soc* **138**, 2864–2868 (1991)
31. J. M. Tarascon and D. Guyomard, "The $Li_{1+x}Mn_2O_4$/C Rocking-Chair System—A Review," *Electrochimica Acta* **38**, 1221–1231 (1993).
32. D. Guyomard and J. M. Tarascon, "Li Metal-Free Rechargeable $LiMn_2O_4$/Carbon Cells: Their Understanding and Optimization," J. *Electrochem. Soc.* **139**, 937–948 (1992).
33. B. Scrosati, "Lithium Rocking Chair Batteries: An Old Concept?" *J. Electrochem. Soc.* **139**, 2776–2781 (1992).
34. M. M. Thackeray, P. J. Johnson, L. A. De Picciotto, P. G. Bruce, and J. B. Goodenough, "Electrochemical Extraction of Lithium from LiMn//20//4," *Mater. Res. Bull.* **19**, 179–187 (1984).
35. M. H. Rossow, A. De Kock, L. A. de Picciotto, M. M. Thackeray, W. I. F. David, and R. M. Iberson, "Structural Aspects of Lithium-Manganese-Oxide Electrodes for Rechargeable Lithium Batteries," *Mater. Res. Bull.* **25**, 173–182 (1990).
36. U. von Sacken, E. Nodwell, A. Sundher and J. R. Dahn, "Comparative Thermal Stability of Carbon Intercalation Anodes and Lithium Metal Anodes for Rechargeable Lithium Batteries," *Solid State Ionics* **69**, 284–290 (1994).
37. T. Zheng, J. S. Xue, and J. R. Dahn, "Lithium Insertion in Hydrogen-Containing Carbonaceous Materials," *Chem. Mater.* **8**, 389–393 (1996).
38. F. Disma, L. Aymard, L. Dupont, and J. M. Tarascon, "Effect of Mechanical Grinding on the Lithium Intercalation Process in Graphites and Soft Carbons," *J. Electrochem. Soc.* **143**, 3959–3972 (1996).
39. I. Yoshio, European Patent No. 0567749AI (23 April 1993).
40. C. Sigala, D. Guyomard, Y. Piffard, and M. Tournoux, "Synthesis and Performances of New Negative Electrode Materials for 'Rocking-Chair' Lithium Batteries," *C. R. Acad. Sci. Paris* **320**, Ser. II, 523–529 (1995).
41. J. M. Tarascon and D. Guyomard, "New Electrolyte Compositions Stable Over the 0 to 5 V Voltage Range and Compatible with the $Li_{1+x}Mn_2O_4$ Carbon Li-ion Cells," *Solid State Ionics* **69**, 293–305 (1994).
42. H. S. Lim and S. A. Verzwyvelt in *Proceedings of the Symposium on Nickel Hydroxide Electrodes*, edited by D. Corrigan and A. Zimmerman, (The Electrochemical Society, Pennington, NJ, 1990), p. 341.

43. J. Giner and J.D. Dunlop, "The Sealed Nickel-Hydrogen Secondary Cell," *J. Electrochem. Soc.* **122**, 4–11 (1975).
44. J. D. Dunlop, G. M. Rao, and T. Y. Yi, in *NASA Handbook for Nickel-Hydrogen Batteries*, NASA Reference Publication **1314** (1993).
45. P. R. K. Chetty, in *Satellite Technology and its Applications*, 2nd ed. (TAB Professional and Reference Books, Blue Ridge Summit, PA, 1991).
46. H. S. Lim and S. J. Stadnick, "Effect of Precharge on Nickel-Hydrogen Cell Storage Capacity," *J. Power Sources* **27**, 69–79 (1989).
47. A. H. Zimmerman, "Mechanisms for Capacity Fading in the NiH_2 Cell and Its Effects on Cycle Life," *The 1992 NASA Aerospace Battery Workshop*, 15–19 Nov. 1992, Huntsville, AL, (1993), pp. 153–175.
48. A. Visintin, A. Anani, S. Srinivasan, A. J. Appleby, and H. S. Lim, "Microcalorimetry Study of Ni/H_2 Battery Self-Discharge Mechanism," *J. Electrochem. Soc.* **139**, 985–988, (1992).
49. A. K. Padhi, K. S. Nanjundaswamy, C. Masquelier, S. Okada, and J. B.Goodenough, "Effect of Structure on the Fe+3/Fe+2 Redox Couple in Iron Phosphates," *J. Electrochem. Soc.* **144**, 1609-1613 (1997).
50. Y. Idota *et al.*, U.S. Patent No. 5,478,671 (1995).
51. Y. Idota, T. Kubota, A. Matsufuji, Y. Maekawa, and T. Miyasaka, "Tin-Based Amorphous Oxide: A High-Capacity Lithium-Ion-Storage Material," *Science* **276**, 1395–1397 (1997).
52. G. Feuillade and J. Perche, "Ion-Conductive Macromolecular Gels and Membranes for Solid Lithium Cells," *J. Appl. Electrochem.* **5**, 63-69 (1975).
53. J. MacCallum and C. Vincent, ed., *Polymer Electrolyte Reviews* (Elsevier, London, 1987), Vol. 1. See also F. Gray, *Solid Polymer Electrolytes. Fundamentals and Technological Applications* (VCH, New York, NY, 1991).
54. T. Gozdz, C. Schmutz, and J. M. Tarascon, U.S. Patent No. 5,296,318 (22 March 1994).
55. T. Gozdz, C. Schmutz, and J. M. Tarascon, P. Warren, U.S. Patent No. 5,418.091 (23 May 1995).
56. T. Gozdz, C. Schmutz, and J. M. Tarascon, P. Warren, U.S. Patent No. 5,456.000 (10 Oct. 1995).
57. J. M. Tarascon, A. S. Gozdz, C. Schmutz, F. Shokoohi, and P.C. Warren, "Performance of Bellcore's Plastic Rechargeable Li-ion Batteries," *Solid State Ionics* **86–88**, 49–54 (1996).
58. J. Rademacher, H. Reed, and J. Paig-Suari, "ASUSAT: An Example of Low-Cost Nanosatellite Development," *Acta Astronautica* **39**, 189–196 (1996).

7

A Systems Approach to Microsystems Development

J. J. Simonne*

7.1 Introduction

The microsystem concept emerged at the end of the 1980s, introduced by the work of Richard S. Muller and coworkers at the Berkeley Sensor and Actuator Center at the University of California, Berkeley. The concept opened the way to overall monolithic integration of sensors, actuators, and signal-processing functions. Microsystems rapidly became an important aspect in several research programs worldwide, with some countries concentrating on slightly different objectives. For instance, the efforts in the United States were in microelectromechanical systems (MEMS), focusing on the microstructures and highlighting the interplay between electrical input and mechanical actuation. Efforts in Japan focused on micromachines to take advantage of the progress in precision mechanical tooling. In Europe, efforts were directed more toward an integrated systems approach, identified as microsystem technology (MST), which combined the advantages of the microstructure world and the progress in interconnection and assembly technologies.

In this chapter a microsystem is defined to be a microstructure, interacting through sensors and actuators with the nonelectric world and providing information and communication with the outer world or with other microsystems, typically sizes smaller than 1 cm^3 through on-board "smart" processing.

Development of the microsystem concept has been driven by two factors:

- **Scientific and technical**: Device development requires a multidisciplinary approach, which introduces the technology into many scientific fields at once, thereby accelerating its growth.
- **Economic and industrial**: Success in providing enhanced functionality of integrated circuits (IC) at reasonable cost is now dictated more by peripheral elements (e.g., packaging) than by the active electronic components.

Batch-processing techniques to fabricate microsystems make them suitable for mass markets (>1 million/year). Examples include pressure sensors, accelerometers, gyroscopes, etc., dedicated to automotive (sensors and fuel injection), ink jet, display (digital micromirrors), information storage, and biomedical application technologies. Large markets for microsystems application, however, are currently limited to these examples, which has forced research-and-development efforts toward servicing lower volume markets (<100,000 units/year), where each unit is a high-value item. Examples include, among others, advanced instrumentation, industrial systems monitoring, defense, aeronautic and space systems applications.

For space applications, MST might, at least in the midterm, permit the reduction of size in select satellite modules, thereby enhancing performance. In the far term, MST may even allow the reduction of all satellite subsystems, as in the proposed Aerospace Corporation design "all-silicon satellite."[1] The concept of stepwise reduction of satellite size from macro to micro, to nano, to picosatellite is a subject for a space systems engineer. An aspect of this is also the center of this

Laboratoire d'Analyse et d'Architecture des Systemes, LAAS-CNRS, France.

present analysis: How does one design and fabricate a complex multifunctional system that is fully integrated and assembled using heterogeneous materials?

This chapter will describe the novelty imposed by the microsystem design on both simulation and microtechnologies. The following issues will be examined.

- Microelectronics has played a pioneer role in microsystem simulation, e.g., rediscovery of the working principle of tools like SPICE (a general-purpose analog circuit simulator), which has enabled simulation of sophisticated semiconductor circuits worldwide and is now used in microsystems design.
- These tools could be extended to the microsystem that includes mechanics, optics, and thermal control; microelectronics becomes one technology in a complex system like a microsystem.
- Standardized microsystem design is still in its infancy as compared with the maturity of the electronic circuit design. How should the standardization proceed? This chapter attempts to lay out a framework for this process.

The argument is made for the need to develop simulation software for building a microsystem. This argument is predicated on gaining some insight into microtechnologies and advanced packaging: examples of microtechnologies include microelectronics techniques, microsystems sensors, actuators, and signal processing; advanced packaging includes monolithic and hybrid. The issues for monolithic type packaging are the use of heterogeneous materials and implementation of postprocessing steps. For hybrid type packaging, where sensor and actuator parts are processed separately, issues are the specific techniques, such as, micromachining, anisotropic etching, and LIGA (lithography, electroforming, and injection molding). Two examples of these technologies developed at LAAS (*Laboratoire d'Analyse et d'Architecture des Systemes*), in the context of European programs, will be illustrated. The final part of the chapter will focus on the peculiarity of microsystems in space applications: the role they can play in the space industry and how emerging microtechniques will significantly modify the design of payloads.

7.2 Simulation Software

7.2.1 Simulation of Semiconductor Circuits: SPICE

The objective of a simulation is twofold: either to check the working design and performance of a circuit not yet processed, or to design a circuit from a library list of specifications and minimize the prototype phase. To understand how such a tool works, let us look at the example dealing with the simulation of electronic circuits.[2]

The behavior of any component in an electronic circuit is monitored by a physical relationship between the voltage V across it and the current I through it. For example, $V = RI$ in a resistor, $I = I_S[e^{-(qV/Kt)} - 1]$ in a diode, and $I = CdV/dt$ in a capacitor. In other words, each component can be represented by a matrix.

The manipulation of these matrix components permits the solving of the implicit equations to determine the value of the I-V for the interconnected circuit as a whole when primary signals, or vectors, are applied at the input. An analog simulator is a computer program dedicated to solving these equations through iterative (or even other) methods. The computer analysis can be managed in the dc mode, to determine a static operating point, or in the transient mode, repeating the procedure for each value of time. Other types of analysis also calculated are noise, frequency response, and small-signal gain.

In SPICE (e.g., P-SPICE, M-SPICE, H-SPICE), which is one of the largest simulators, a list of the primitive elements (i.e., electrical components) to be interconnected in the design layout of the electronic circuit is indexed in a library and associated to the simulator engine. This

interconnection between the elements generates a block and is commonly called a "structural" model, or structural description of the design. Each element is considered as a black box provided with the functions representing the interactions between access pins (or terminals) of the black box. The inner functions within the black box have the form of differential, linear, or nonlinear equations. These equations are parametrized to describe the element as realistically as possible. In fact, these parameters, also available in the libraries, can be defined and are coefficients of the equations. The functions are also associated to an algorithmic (or conditional) structure to define the behavior of the element in the circuit more accurately. Taking the example of diodes, the I_S parameter will vary according to the technology used to make the diode (e.g., size, process type, temperature coefficient, access resistance). The algorithmic representation of a forward biased diode is written as:

If V > 0 (If the condition V > 0 is met),
then "do" ⟶ $I = I_S e^{(qV/kT)}$;
otherwise "do" ⟶ $I = -I_S$ (the diode is backward biased).

The description of an element, through its functions, parameters, and algorithms, as previously defined, is called the "behavioral" model and is associated to the element.

7.2.2 Simulation of VLSI circuits

In the most accurate simulation in VLSI (very-large-scale integration), all elements are represented at the lowest level of abstraction, also called the physical level. The result is a more precise functional description of the circuit. Such circuits are described in terms of basic gates, switches, and more generally, in interconnected electronic cell units.

This procedure provides the highest level of accuracy to the designer. However, working with schematics at the gate level involves tens of thousands of components. This becomes impractical and requires considerable time. Therefore, for VLSI circuits, a new approach to simulation is needed. To reach shorter time scales, new elements should be created at a higher functional level. However, this decreases the number of elements and the accuracy of each, which reduces the overall design accuracy.

In the analog simulation world, simulators are provided with libraries rich enough to start simulation without the need to create all models (or primitives or templates). The basic principle is to associate several models available in the simulator to build larger macromodels.[3] But at the same time, a language is available in the simulator allowing the user to enhance the library by creating user models.

A simulator for the design of complex circuits makes possible several levels of abstraction. Starting from the physical level as already presented, there is an upper level, where subfunctions with larger sizes constitute the new basic entities of the level, followed by a functional level with extended size entities also indexed in the library, and finally the top level, the simplest architectural level, which constitutes the highest abstraction level in the simulator. Structural and behavioral modelings exist at any level and are managed according to the method followed by the designer.

7.2.3 Top-Down Methodology

The methodology of the circuit designer is opposite to this "bottom-up" simulator structure, where at each level, each element is designed and optimized separately. The typical design methodology uses more a "top-down" approach, where each subfunction is optimized with respect to its interaction with the others. This approach has been formerly applied to the design of complex digital systems.

The designer starts from the digital system specifications expressed in clear writing and translates them into an architectural organization comprising hierarchical levels. At each level, functional blocks and sub-blocks are defined and interconnected; the logic synthesis is achieved automatically from this hierarchical decomposition and from the behavioral model of each element. A synopsis of the various simulation levels of the digital circuits simulation in the top-down methodology is shown in Table 7.1. The behavioral description as presented in the form of abstract expressions (algorithms, logic expressions, etc.) will, in this example of logic synthesis, automatically be translated into basic elements organized as the circuit layout.

Therefore, there is a need to develop a language that interprets the structural and behavioral descriptions at any hierarchical level. If such a language were to be created, it offers the advantage of a common language between designers, manufacturers, and consumers, as well as between various simulators.

7.2.4 VHDL Language (VHSICHDL: Very High Speed Integrated Circuit Hardware Description Language)[4]

HDL is associated with the top-down methodology and yields the definition of the digital circuit layout. Such a language, when associated to a simulator, also permits systematic verification of the design at each abstraction level, such as stray thermal, vibration, and radiation phenomena and overall performance. This verification can be done well before the selection of the final technology. Tools, such as SPICE and the ELDO[5,6] simulators, using an HDL were introduced in the beginning of the 1980s by IBM, Texas Instruments, Intermetrics, and Analogy to offer a foundation for the development of a global functional simulation program. Specifically dedicated to digital electronic silicon systems, the high-speed version called VHDL was developed step by step to improve a standard communication format whose usefulness has already been mentioned. This language is standardized under the IEEE (Institute of Electrical and Electronics Engineers) norm 1076 and is regularly updated.

7.2.5 Extension of the Simulation to Other Fields

If we wish to consider how a digital simulator works, we would say that its basic principle relies on statistical or algebraic rules applied to signals that are quantized in amplitude and discrete in time. An analog simulator, on the other hand, relies on energy conservation rules on actual variables applied to analog waveforms being continuous in amplitude and in time. A mixed-mode simulator operates through use of synchronous algorithms and can be considered a mixed simulator having both analog and digital kernels. However for microsystems, which must integrate electronic, mechanical, optic, and fluidic elements, a new challenge arises in integrating all the various property simulators. Once again, the top-down microsystem design simulation tool must adopt a common behavioral language that is able to integrate the necessary information from the multisciences.

Table 7.1. Synopsis of the Various Simulation Levels of the Digital Circuits Simulation in the Top-Down Methodology.

Level	Description	Simulation
System	Specification language	Behavioral
Blocks—registers	Behavioral language	Logic behavioral
Gates—switches	Boolean equations	Logic
Circuit	Differential equation	Analog

7.2.6 Modeling Languages for Microsystems

A language dealing with microsystems must support analog and mixed-analog signals as VHDL does for the description and simulation of digital systems, but it must also extend its scope to support those nonelectrical systems that can be made analogous to electrical systems in their modeling. If one makes VHDL the basis for this common description language, for example, VHDL-A (A for analog),[7] it should handle both discrete- and continuous-time-element description, and should support mixed structural and behavioral specifications. It should also provide mechanisms that allow the digital and analog parts of a mixed description to interact with each other.

There is no proposal to make VHDL-A the standard. As a matter of fact, very few simulators are able to support mixed, structural, and hierarchical languages and to account for the various types of physical phenomena (multiscience modeling). Although standardization on VHDL-AMS (analog mixed signals) is progressing, the likelihood of implementation is postponed until 1999.

Designers, cooperating with manufacturers and consumers, are pushing efforts to see the norm VHDL-A set up successfully. Meanwhile, other languages with equivalent advantages are used.

- ELDO[5] (fast electrical VHDL-A-based simulator designed for simulation of analog, digital, and mixed circuits of various technologies,) accepts both mixed and behavioral descriptions of multiscience elements, but not a structural description.
- MAST,[8] a language based on SABER[6] (designed for simulation of analog, digital, event-driven analog, and mixed-mode systems), allows expression of behavioral models for multiscience elements, structural, and hierarchical descriptions.

Other behavioral modeling languages are proposed based on simulators such as INTOSOFT SPICE, SPECTRE HDL, SMASH, DAULPHIN, and AMESIM.

Today, simulator users can construct their own element libraries specifically oriented to their fields. These elements, included in a layout description, will permit simulations to be performed. However, simulators based on different languages cannot combine their libraries directly until there is agreement on a standard language basis. Current electronic industries use the top-down methodology to design new circuits. Models and simulations of specific integrated circuits (up to a few million components) are designed and tested within several months, resulting in a final product with a first-run fabrication success rate close to 100%. This success level can be maintained as long as there is a high level of confidence in the representative models and simulation tools. This methodology is progressively being applied to microsystems since it cuts down time-consumption in the design and fabrication phases and allows the processing of a limited series at an affordable cost. But this approach implies a complementary parallel effort in infrastructure technology to develop components, processing technology, and simulation tools. In short, we recommend application of the silicon microelectronic methodology to microsystems design.

7.2.7 General Procedure for Microsystem Simulation

We know now that two types of modeling descriptions are available in the simulation procedure: the physical (structural) and behavioral. Outcomes from these modeling studies must be integrated to design a microsystem. Following is an approach to this integration.

- The various component designers describe the components at the technology level with FEM (finite element method) simulators, using a structural modeling language. The different models consist of mathematical representations of physical phenomena acting on components and are expressed in various languages. To be able to create a behavioral description of each element in VHDL-A format, a model translator is used to store the different transient-domain simulation models in a VHDL-A library, which is made available to global access.

- The microelectronics designer will extend the standard IC design framework to encompass microsystem technologies and generate the layout generation of the whole microsystem (electronic, mechanical, optical, thermal) and verify a design rule. The result is linked to available packaging technologies either monolithic or hybrid. A parameter extraction tool (as found in microelectronics) from the layout level enables consideration of the effect of packaging on the behavior of the global system. As an example, the mixed-mode multilevel simulator ELDO/VHDL-A allows a full verification of the design functionality (Fig. 7.1).

7.2.8 Space-based Microsystems Specifics

Top-down methodology could potentially reduce design and fabrication costs of manufacturing in small lots, making it very attractive to the space industry. Moreover, any mass and volume reductions brought about by use of microsystems are added benefits in a satellite. By reducing mass and volume, smaller launchers can be used. On the other hand, if mass and volume are retained, added functionality can be given to the satellite, or the mission life can be extended by increasing the attitude control propellant. It should be mentioned that an orbit correction may use up to 20 kg of propellant per ton of satellite mass. These actions effectively limit the lifetime of a geostationary satellite. Therefore, solutions that favor low cost may be important, but those that favor lower weight and volume, as in microsystems, may be more paramount.

Another feature of the top-down methodology is the ability to simulate environmental effects on microsystem design, such as assessing stress and thermal design.[9] Given that the space application environment is harsh (e.g., radiation, hot/cold thermal cycling), inclusion of stress analysis in the microsystem design would be necessary for any space application. Currently, inclusion of global stress modeling is not possible for a microsystem. However, adopting a deterministic methodology approach could provide a means for making an assessment for space reliability.

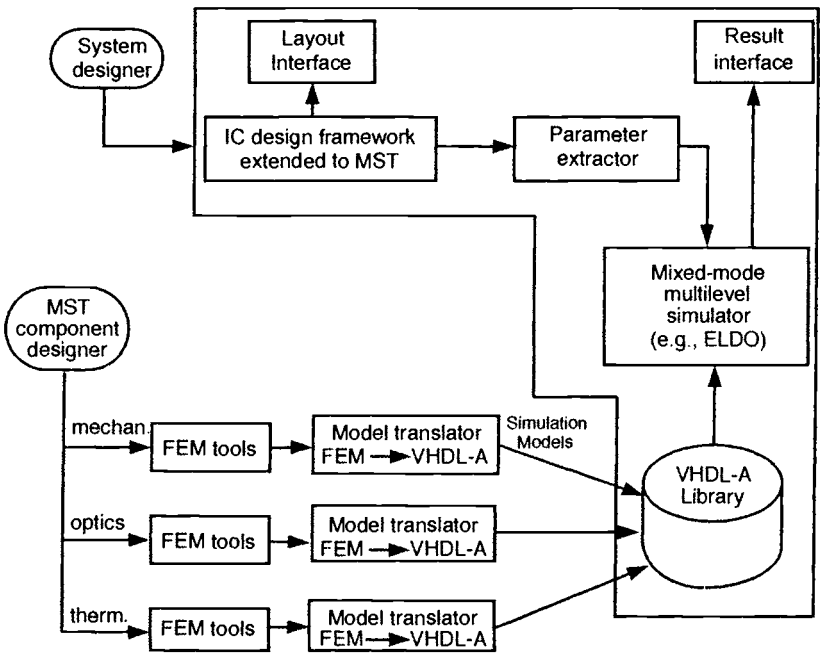


Fig. 7.1. CAD tools for microsystem design.

7.2.9 Deterministic Methodology

Space systems reliability has for a long time been based on statistical methods. The procedure is to fabricate and test samples, from which statistics are derived to assess the defect probability; for instance, a good breakdown rate could be $\tau_B = 10^{-6}$ a year. Strictly speaking, this would mean that at least 10^6 samples are needed to determine the breakdown rate. To overcome this impracticable method, simple procedures are used to accelerate the breakdown rate, such as increasing the temperature during tests. According to the Arrhenius law, the breakdown rate τ at temperature T becomes $\tau = \tau_B e^{(E/(kT))}$, where E is the activation energy, k the Boltzmann constant, and T the temperature. For example, τ is reduced from 10^{-6} at ambient temperature down to 10^{-3} at $T = 1000$. Nevertheless, the statistical methodology still requires large-scale arrays to be fabricated and tested.

The deterministic methodology, on the other hand, relies on the simulation of a fabrication process (e.g., SUPREM), formerly developed for VLSI and extended to microsystems design. It includes assembly and packaging procedures. From the combined physical models of these three steps in the complete integrated description, time evolution of the operating system parameters is deduced. As a consequence, a few samples need to be tested.

This methodology becomes, for obvious reasons, more attractive and even more accurate: more attractive because it is based on simulation methods, more accurate because the breakdown rate can be related to one or more physical mechanisms. For example, consider the behavior of a system embedded in resin packaging and its failure as a result of water absorbance: the physical model will include the water diffusion coefficient based on particular resin absorption properties. Even though the deterministic method is already available to microelectronics systems, simulators for complex microsystems analysis are still under development. In the long term, the deterministic methodology should definitely ensure cost savings.

The preceding analysis has summarized the variety of simulation tools that are available to microsystem designers or that have to be improved to reach the current level of electronic circuit design. The second aspect of this discussion is the technology tools that can be offered to customers to implement their projects. In the same way that the microsystem design is an extension of electronic circuit design with the addition of new tools, microsystem technology is an extension of microelectronic processes with the incorporation of new procedures. New procedures include

- Silicon and polysilicon surface micromachining, making possible fabrication of microcavities and free-standing cantilevers and moving parts by etching away the underlying silicon-dioxide films
- Front- and back-side anisotropic etching, making possible silicon membranes
- The LIGA technique,[10,11] making possible mass-replication of high-aspect ratio structures out of metals, polymers and ceramics, and the ability to assemble microdevices with a high degree of accuracy.

As regards packaging of microsystems, two approaches are proposed: integrate sensors and actuators with the electronics (i.e., monolithic packaging) with postprocessing steps, or fabricate the sensors and actuators as separate die and package with the electronic die via a hybrid packaging approach (e.g., multichip module).

7.3 Microsystem Fabrication Technologies

Developing a viable microsystem requires examining not only each functional unit, but also material compatibilities between the various elements, their sizes, and the interfaces between them and with the environment, regardless of the fabrication technique adopted. Much information can be drawn directly from microelectronic technology: thermal, mechanic, electric, and magnetic.

However, there are areas that are important to microsystems but are not in the traditional purview of microelectronics, specifically optical, electromechanical, and fluidic functions. Appropriate interfaces with these new functions must be dealt with.

In contrast to electronic devices, no standard yet exists for microsystems, which often require assembly of several layers of various materials. In addition, microinstruments may include cavities, moving parts of various dimensions and geometries, as well as substrates providing optical, fluidic, and electromechanical links. Packaging microsystems has problems similar to those already encountered in hybrid packaging of microelectronics.

As already stated in the introduction of this chapter, the space industry is not a real driving force for mass production of microsystems. However, the space industry must make the best possible use of new technology to meet the requirement for high-reliability components and systems. One solution is the development of a space-qualified, microinstrument-design technology to take advantage of components, chips, and associated technologies used in terrestrial and commercial applications.

There are various packaging schemes under consideration for microsystems: monolithic microsystem integration (MMI), the hybrid multichip module (MCM), and the new concept labeled ultra-thin-chip-stacking (UTCS) technology.

7.3.1 Monolithic Microsystem Integration (MMI)

In this context, MMI is understood as an extension of IC processing to fabricate sensor arrays with corresponding signal-processing circuits. The microsystem includes heterogeneous materials such as the sensing elements, but the integration process is monolithic in the packaging context.

The procedure starts with substrate conditioning: a standard silicon substrate is prepared in the same way as for microelectronics. Following substrate preparation, the fabrication scheme follows the classic IC methodology, which consists of several operations starting from the highest temperature down to metallization at 400°C. An exception comes if chemical sensors[12] (ISFET: Ion Sensitive Field Effect Transistor) are to be fabricated. The metallic gate is replaced by an ion-sensitive membrane, upon which resides the electrolyte. The device is completed by incorporating a reference electrode. In this particular case no special postprocessing is required for obtaining an MMI category microsystem. After metallization, the postprocessing steps follow the same IC methodology requirements for temperature processing. The process steps must be ranked in decreasing steps of temperature to prevent modification on the preceding operations.

These low-temperature, postprocessing steps include deposition of heterogeneous materials, like polymers, which are processed to get sensor arrays, and anisotropic etching of front and back side, to fabricate a variety of sensing elements in the substrate area (Fig. 7.2).

The main drawback of this packaging approach is that the electronic circuits, which are processed first, must survive all postprocessing steps required by the sensor implementation. A good example of MMI is the European DEMAC project, where the goal is to define and implement a design and technology methodology for smart monolithic integrated sensors and microsystems, and to make it available for system designers through a standard silicon foundry. This successful project could be compared with MOSIS (Metal Oxide Semiconductor Implementation System), a low-cost prototyping and small-volume production service for custom and semicustom VLSI circuit development at the University of Southern California Information Science Institute. In comparison with the MOSIS service where only the postprocess specifications are given to the user, DEMAC intends to offer a postprocessing service more like that of *D+T Microelectronica, Centro Nacional de Microelectronica*, Bellaterra, Barcelona, Spain.

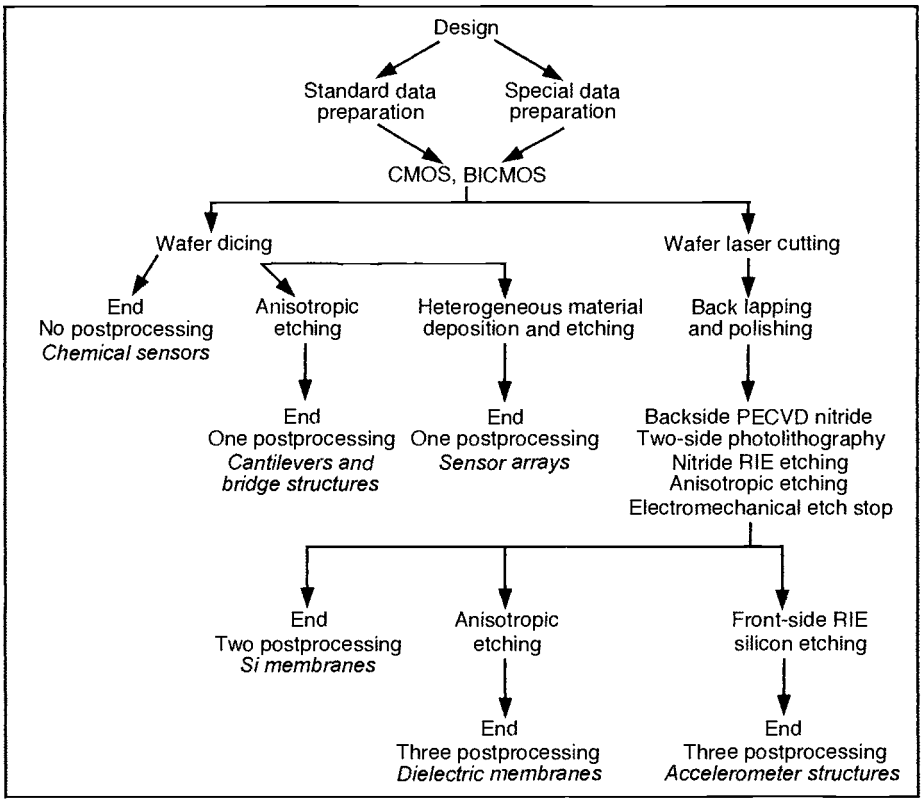


Fig. 7.2. Monolithic microsystem technology: various options according to postprocessing steps.

7.3.2 Multichip Modules (MCMs)

According to standard definition, an MCM is a structure where bare chips are interconnected on a common supporting substrate.[13] Considering its hybrid characteristics, an MCM could more explicitly be defined as a functional electronic module, composed predominantly of bare die with high-density interconnections on a substrate that is either ceramic, silicon, organic, or metallic, and an interconnection network of successive isolation and conductive layers.[14] All available bonding technologies can be used to interconnect the chips: wire bonding, which includes the basic ball- and wedge-bonding; tape automated bonding (TAB), complementary to the preceding method where the dice is already bonded and transferred onto a support film; flip chip, a technology introduced by IBM and achieved through solder balls often distributed as a matrix array on its surface.[15] There are other techniques called chip-scale package, thin-small-outline package, and chip on board.[13,16] No one bonding technology has emerged as leader, and all MCM technologies are still in an evolutionary mode.

Several approaches[16] to MCM packaging have been developed. They are labeled MCM-C for "ceramic," MCM-D for "deposited," and MCM-L for "laminate." A brief review of the characteristics of these two-dimensional (2D) packaging technologies will be followed by another technology called MCM-V, for "vertical." MCM-V is a three-dimensional (3D) packaging approach, which gives the highest degree of compactness and, as a result, could become the most attractive for space application.

7.3.2.1 Multichip Modules 2D-Packaging Technologies

MCMs are usually ranked according to their substrate technologies and the density level of interconnections. The 2D-MCM families[16] are the following.

- MCM-L is an extension of chip-on-board (COB) technologies. Substrates can be either a conventional rigid printed circuit board (PCB) or flexible circuits.
- MCM-C is an extension of thick-film technologies. Substrates are multilayer ceramic cofired (1600°C–1800°C), often alumina, where bare chips are mounted on, and interconnected with, screen-printed conductors.
- MCM-D is the solution providing the highest interconnection density owing to very fine-line capabilities. Substrates are mainly silicon or ceramic supports, on which polymers or SiO_2 are deposited. Interconnections are processed through metallic thin-film deposition techniques using IC processing.

These hybrid 2D-packaging approaches, where the interconnection densities are close to those of wafer-scale-integration (WSI)[16] technology (the construction of very large ICs that completely occupy the space of a silicon wafer), already give a solution to packaging of complex systems where costs are a concern.

7.3.2.2 3D-Assembly Approaches

In the 1990s, the need to develop large memory modules with a low bit-transfer transit time in a package of reasonable weight and volume size promoted the development of 3D MCMs. With this technology that uses vertical interconnection schemes, the interconnection densities increased by several hundred percent as compared to the best MCMs-L/C/D and WSI approaches.

Increasing the number of I/Os in a 2D system[17,18] requires bringing the die closer, which often makes the floor plan more difficult. The 3D approach offers a solution to this problem and also shrinks the signal propagation time. This is a major improvement as clock rates in microprocessors continue to increase (currently 400 Hz). The 3D-integration scheme allows a data transfer between two die in just one clock cycle.

Several methods of die and stacking are used throughout the world, although mainly restricted to the packaging of identical elements.[19,20] In the United States, the 3D-packaging scheme has been primarily used for space or military products dedicated to specific programs. The MCM-V family is basically dedicated to bringing the 3D-packaging concept to the consumer.

7.3.2.3 3D MCM Technology: MCM-V

MCM-V packaging involves vertical stacking of substrates with interconnected die organized to form a compact assembly of parallel layers. Each layer of stacking is attached to each other by a molding, resulting in an overall rigid mechanical structure. This option of a nonseparable 3D assembly, developed by Thomson-CSF and its cousin $3D^+$ (European TRIMOD project), is recommended for space applications.

The 3D-MCM approach results in a reduction of size and weight by a factor of approximately 7 if we compare[21] a very sophisticated MCM-C package with an MCM-V package (see Table 7.2). A cost comparison between these two approaches is not yet possible because of limited production.

7.3.3 The 3D MCM-V Stacking Technique

This approach, which enables nonidentical chips to be stacked, involves these steps:

1. Microassembly of chip-on-tape, using a flexible polymer film or, even better, silicon, which can become "smart" when electronic circuits are deposited
2. Burn-in and electrical testing, a technique commonly used in VLSI

Table 7.2. Comparison Between Memory Modules in MCM-C and MCM-V Technologies

Quantity	MCM-C Double-Side Flip-Chip Memory	MCM-V
Area (mm^2)	4053.0	214.5
Weight (g)	64.5	9.1
Volume (mm^3)	26625.0	4075.0
Density ($Gbits/dm^3$)	13.2	62.9

3. Stacking substrates and molding, the result being a cube a few millimeters larger than the final volume block. Details are given in Sec. 7.4.9.
4. Sawing the cube with a diamond saw
5. Side-plating procedure and laser writing; the vertical connections are patterned using a YAG (yttrium aluminum garnet) laser. The result is no internal connection between each level of the stack

A schematic view of the process is displayed in Fig. 7.3. This method until now had been an industry experiment, using only electronic elements (e.g., block memory with 16 DRAM memory chips each of 16 Mbits, giving a total of 256 Mbits, stacked and designed by Alcatel Espace,)[19,20] or the development of a microcamera, which assembles 50 components in a volume of 1.5 cm^3 by $3D^+$ packaging. Recently a new technology development has been prototyped, which introduces a fluidic element (a micropump) into the stack, the European BARMINT (Basic Research of Microsystem Integration) project,[22] developed by LAAS and partners.

7.3.4 Ultra Compact Assembly Perspectives

An alternative approach is the ultra-thin-chip-stacking (UTCS) technology, which is still in the research phase (1996 European ESPRIT program) but appears promising and could announce the next generation of compact assemblies. The basic principle of this novel technology is to build up a 3D stack according to the following procedure. Chips, on silicon and even on III-V semiconductors, coming from standard process lines, are thinned down to 5–10 μm. They are mounted on a silicon substrate that is metallized if a metallization pattern is required at this level. A dielectric layer is deposited and then the substrate is planarized. Vias are drilled into the dielectric to allow contacts with the buried chips, and a new metallization pattern on top of the dielectric is patterned. More chips are then mounted on the dielectric, and the procedure is repeated.This procedure can be extended to several layers. The final stack has a thickness of the same order of magnitude as a conventional silicon chip but has bonding pads for I/O. A schematic drawing is shown in Fig. 7.4

This new approach to miniaturization of electronic systems packaging requires mastering various technologies: the thinning technology must be achieved with 1- or 2-μm accuracy; and transport, attachment, and bonding of very thin chips are difficulties to be overcome. Certainly to be considered are thermal problems. However, UTCS, compared with monolithic packaging (WSI), has many more advantages; for example, different semiconductor technologies can be mixed without incompatibility problems. This project is at its initial stages. This innovative solution is of interest to all electronics industries where size and weight are of prime importance, which includes all portable applications like telecom terminals, personal digital assistants, portable computers, and in aerospace systems, avionics and satellite applications. A rough estimate of the size reduction given by UTCS was evaluated in the TRIMOD project on a 16 × 16-Mbit DRAM chip in comparison with MCM-C and MCM-V technologies. The result is shown in Table 7.3.

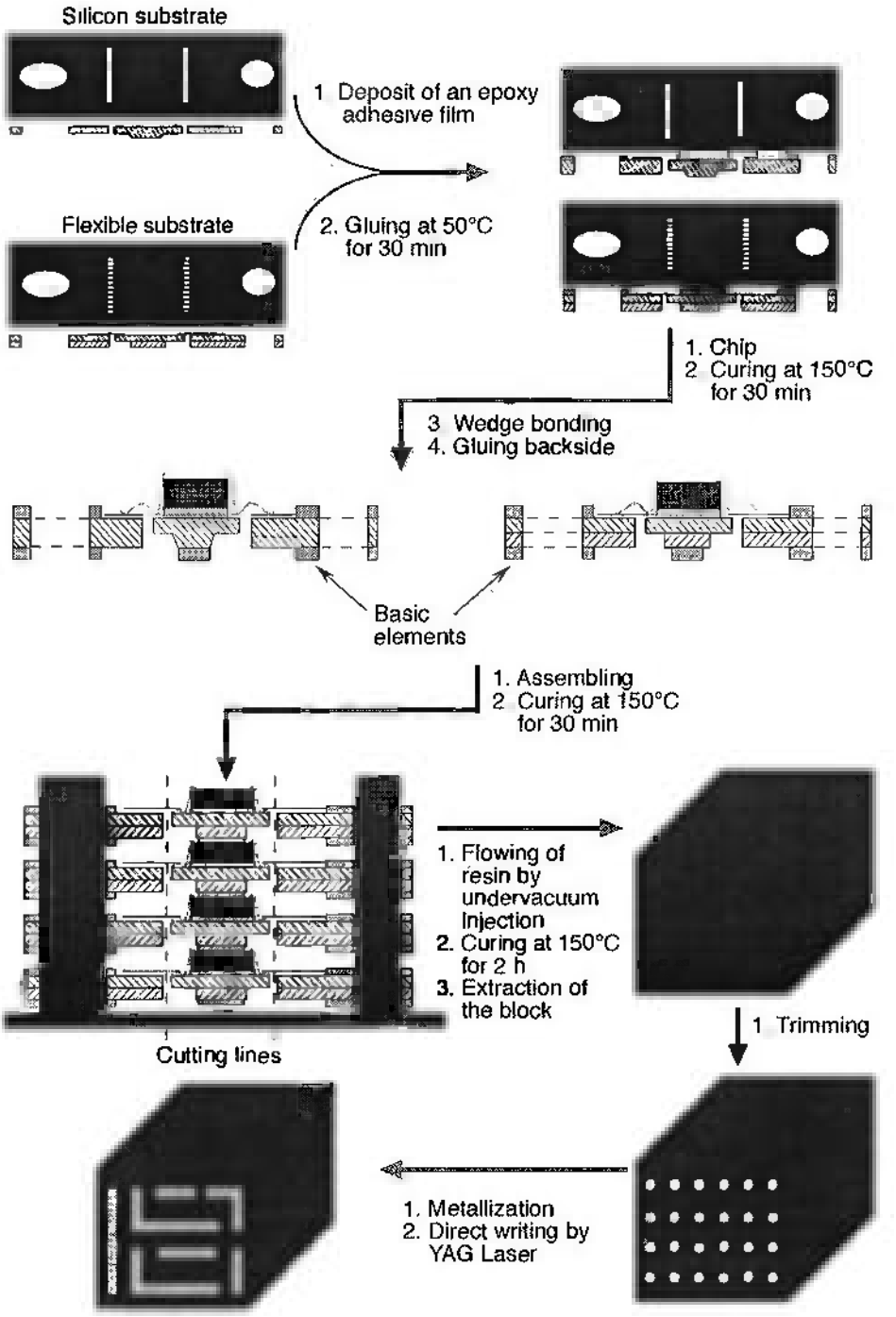


Fig. 7.3. The 3D stacking process.

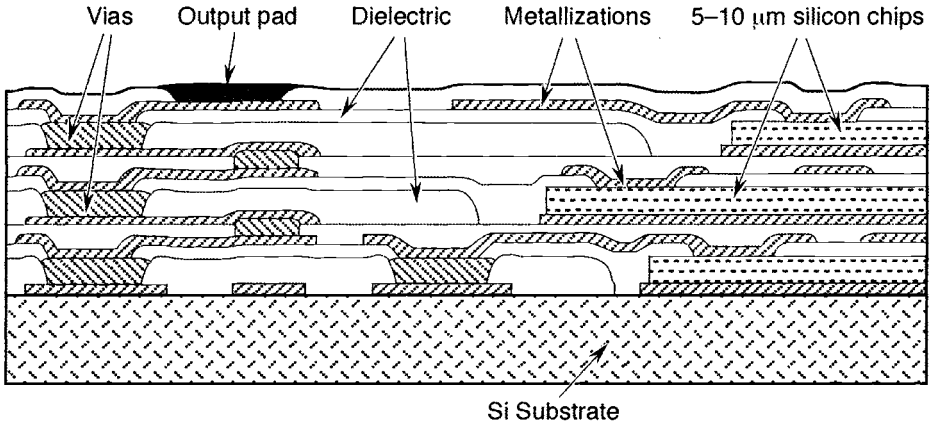


Fig. 7.4. Schematic cross-section of the ultra-thin-chip stack (courtesy Alcatel Espace).

Table 7.3. Comparison Between DRAM Chips in MCM-C, MCM-V, and UTCS Technologies

Quantity	MCM-C	MCM-V (3D Stack)	UTCS
Weight (g)	52	7	0.5
Size (mm)	63 × 63 × 5	19.5 ×19 × 11	20 × 20 × 0.5
Volume (cm3)	20	4.7	0.2

7.4 Two Microsystem Investigation Examples at LAAS

This section outlines two European programs spearheaded by LAAS. The development of the PROMETHEUS-PROCHIP[23] project dedicated to a passive infrared obstacle detection focal plane array (FPA), and the MCM-V/BARMINT project dedicated to the design and the realization of a 3D stack that includes a micropump.

7.4.1 Uncooled Infrared VDF-TrFE Staring Array

LAAS has developed a 32 × 32 100-μm pitch copolymer VDF-TrFE (vinylidene fluoride-trifluoroethylene copolymer) staring array, with sufficient definition to be used in imagery applications. The light detection is based on the pyroelectric effect in the copolymer. Figures 7.5 and 7.6

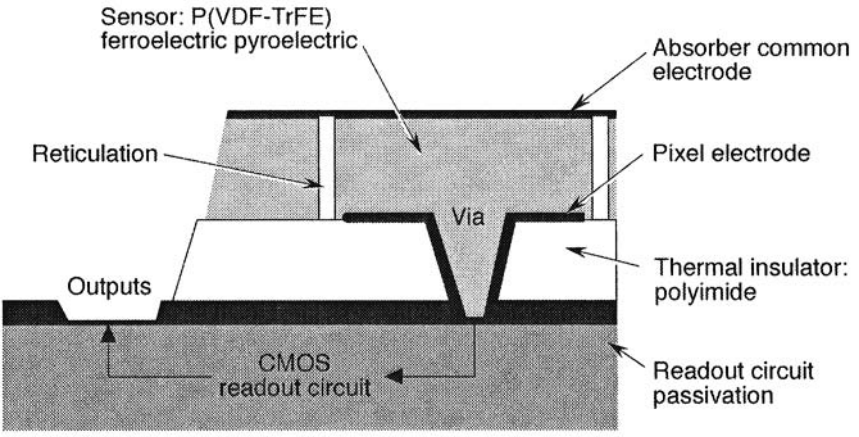


Fig. 7.5. Detector array: pixel cross-section.

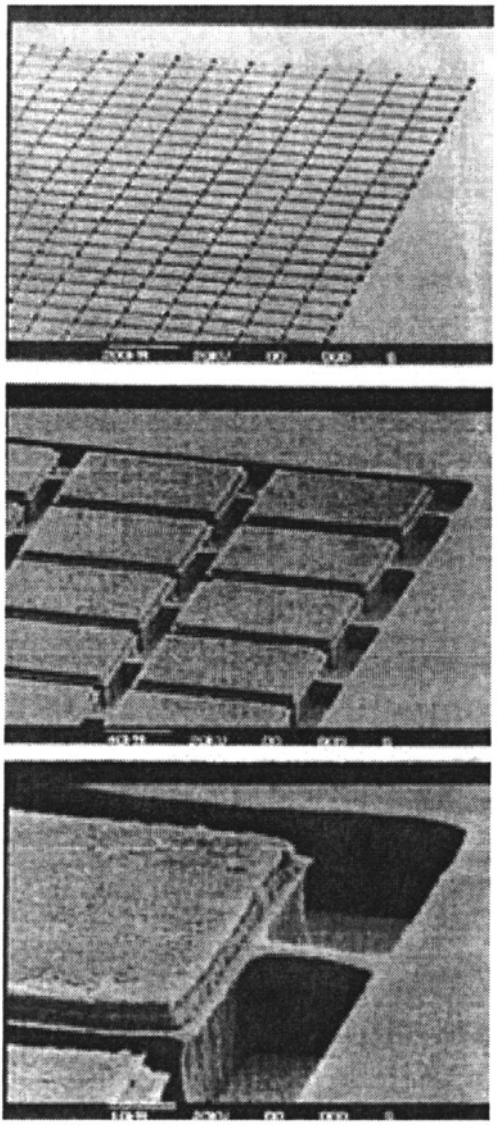


Fig. 7.6. SEM view of the detector array reticulated by dry etching.

present the technology and a SEM (scanning electron micrograph) image of the array. The input infrared photon flux is converted by the absorbing layer into thermal energy, which heats the copolymer. The temperature change induces an electrical charge on the metallized interface. The detecting layer is thermally isolated from the silicon by a specific layer that optimizes temperature signal change in the copolymer. The copolymer is also reticulated to reduce cross-talk between the pixels. The array is processed on a CMOS (complementary metal oxide semiconductor) readout circuit that filters, amplifies, and multiplexes the output signal.[24] This pyroelectric FPA is ac coupled and requires a chopper.

Another microsystem, which is monolithic in uncooled packaging and provides imagery, is the resistive bolometer technology developed by Honeywell[25] and licensed to Loral, Boeing, Hughes-SBRC (Santa Barbara Research Center) and Amber. The microdevice has 320 × 240-array microbridges, where each microbridge has a thermoresistive element (Fig. 7.7). Measurements of the array NETD (noise equivalent temperature difference) performance with an f/1 lens have reached 40 mK at 30 Hz. The sensor is dc coupled and therefore does not require a chopper. Only the lens and the pixel geometry limit the camera MTF (modulation transfer function): the microbolometer pixels have no thermal coupling between nearest neighbor pixels because of the excellent thermal isolation. The focal plane is packaged with a temperature stabilizer operating at 25°C.

7.4.1.1 The Staring Array

The successive processing steps for the development of the LAAS staring array are displayed in Fig. 7.8. The copolymer, which crystallizes in a beta polar phase, is available under a liquid form. The liquid copolymer is directly deposited by spin-coating upon its support: the CMOS readout circuit. The readout circuit is fabricated using a standard CMOS process with 1.2 μm geometry. Once the complete detector array has been processed, the material is then poled to give the copolymer its ferroelectric and pyroelectric properties. The poling procedure is a difficult process to perform across the whole structure without destroying the readout circuits. The method first used and developed by ISL (German-French Saint-Louis Research Institute),[26] consists of applying a low-frequency electric field (0.2 Hz) at room temperature. This allows a hysteresis loop to be generated while the magnitude of the field is continuously increased to nearly the material breakdown voltage limit. Reproducible values for the remnant polarization equal or higher than 8 $\mu C/cm^2$ are commonly reached. Finally, all circuits are tested on the wafer before packaging. The whole development process is fully compatible with microelectronics processing technology. The complexity of the staring array technology is comparable to that of BiCMOS, and meets the "low-cost" fabrication requirements, which is an advantage over the costlier quantum-detection, cooled-semiconductor technology.

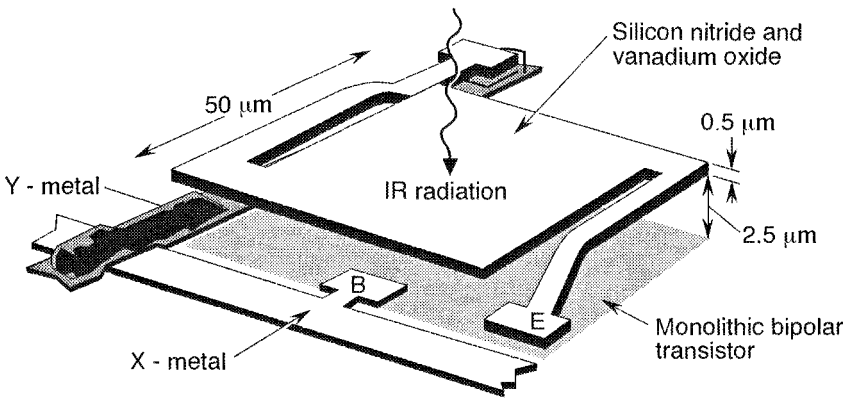


Fig. 7.7. Thermal isolation structure employed in the uncooled monolithic thin-film 2D resistive bolometer array (from R. A. Wood *et al.*, Honeywell Technology Center, reprinted courtesy Honeywell).

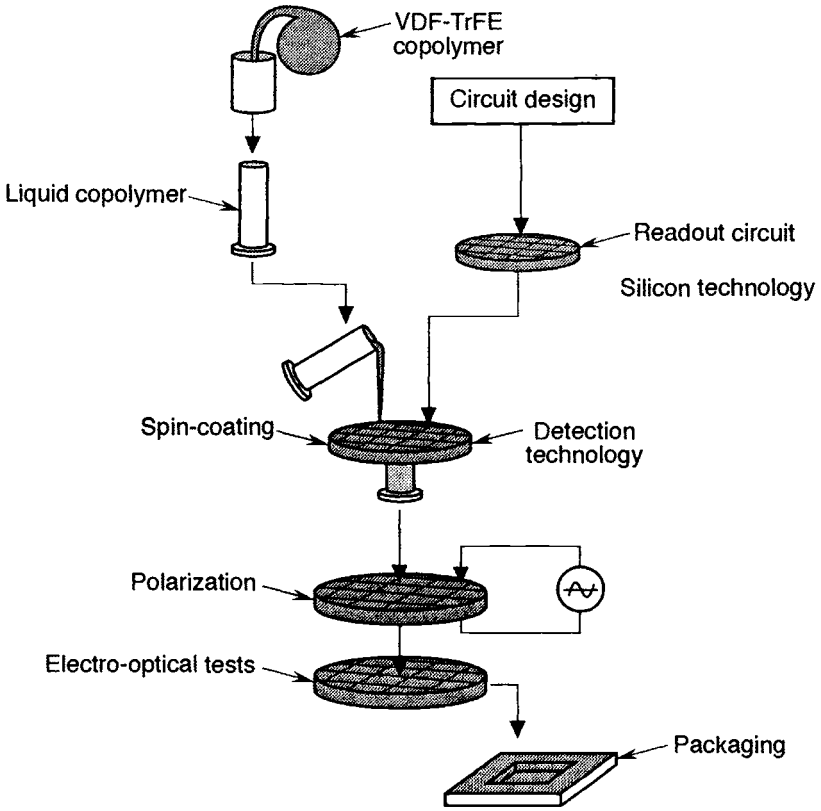


Fig. 7.8. Copolymer array technology.

7.4.1.2 Sensor Modeling

In order to optimize the performance of the sensor, theoretical analysis of the response has been investigated using a classic thermal diffusion model, which is adapted to the heterogeneous monolithic structure. The general thermal diffusion equation shown as Eq. (7.1) has been computed to predict thermal diffusion through the heterostructure.

$$\frac{\partial T(x,y,z,t)}{\partial t} = D_T\left(\frac{\partial^2 T(x,y,z,t)}{\partial x^2} + \frac{\partial^2 T(x,y,z,t)}{\partial x^2} + \frac{\partial^2 T(x,y,z,t)}{\partial x^2}\right) , \tag{7.1}$$

where $T(x,y,z,t)$ is the temperature in the copolymer and D_T is thermal diffusivity of the layer (m^2s^{-1}).

Figure 7.9 shows a static simulation result when only one pixel is heated. This model allows evaluation of the mean temperature in the pyroelectric layer and thus the pixel response V_{pixel} when the pyroelectric coefficient of the copolymer is known. Eq. (7.2) relates the temperature distribution with the corresponding voltage response.

$$\varepsilon \nabla^2 V_{pixel} = \rho\left(\frac{\partial(T(x,y,z,t) - T_{amb})}{\partial z}\right) , \tag{7.2}$$

where ε is permittivity, ρ the pyroelectric coefficient, and T_{amb} ambient temperature.

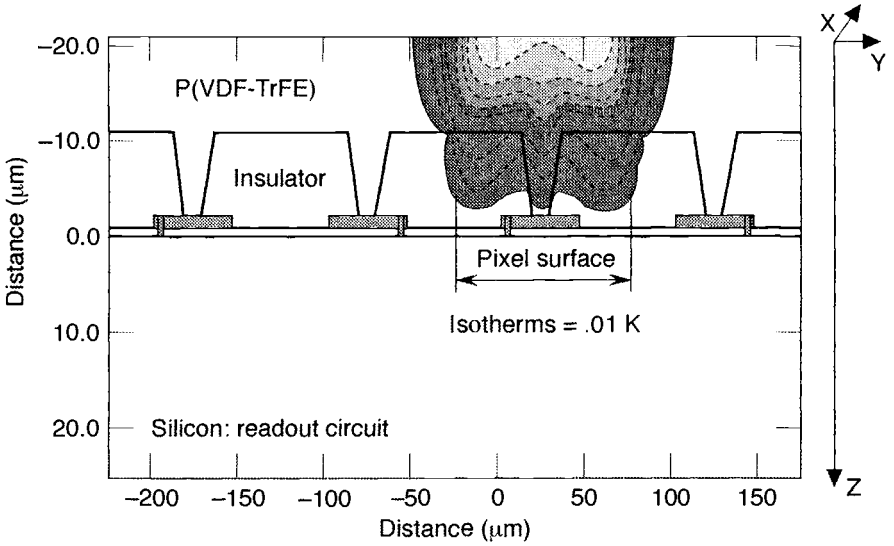


Fig. 7.9. Thermal diffusion through the pyroelectric structure.

7.4.1.3 Sensor Noise

Noise of the copolymer has been calculated assuming the copolymer as a dielectric material. Assuming that the dielectric loss can be represented by a loss resistance of the form $R_p = (C\omega\tan\delta)^{-1}$ in parallel with the copolymer capacitance C, Eq. (7.3) gives the noise input voltage e_b due to the copolymer, after transformation of the R_p–C parallel circuit into an R_S–C_S series circuit, where $C_S = C$, and $R_S = (\sin(2\delta)/2\omega C)$

$$(e_b)^2 = 4kTR_s, \text{ or } e_b = \sqrt{\frac{(2kT\sin(2\delta)}{\omega C}}, \tag{7.3}$$

where δ is the loss angle, ω the pulsation frequency, and k the Boltzmann constant.

Sensor performances are typically characterized by the NETD (noise equivalent temperature difference) parameter. The NETD represents the minimum contrast temperature, which gives the sensor a response equal to its RMS noise level. NETD defines the thermal resolution with a signal-to-noise ratio of one. This parameter can be calculated once the measured responsivity (mV/K) and the total output noise (mV) of the sensor is known. Figure 7.10 shows typical electrooptical characterizations with an f/1 lens at a chopper frequency of 10 Hz. A mean value of NETD of 0.4 K has been measured; however, with an ideal VDF-TrFE sensor, which is only limited by the copolymer noise, a theoretical limit of 160 mK should be reached. In the data shown in Fig. 7.10, the sensor performance is actually limited by the readout circuit noise.

The Honeywell microbolometer and the 128 × 128-array copolymer sensor distributed by Thomson-CSF are available on the market. Performances of the microbolometer are only slightly better. But the microbolometer approach has a technological advantage in that the thermal insulation provided by the air film is 10 times better than that provided by the polyimide film.

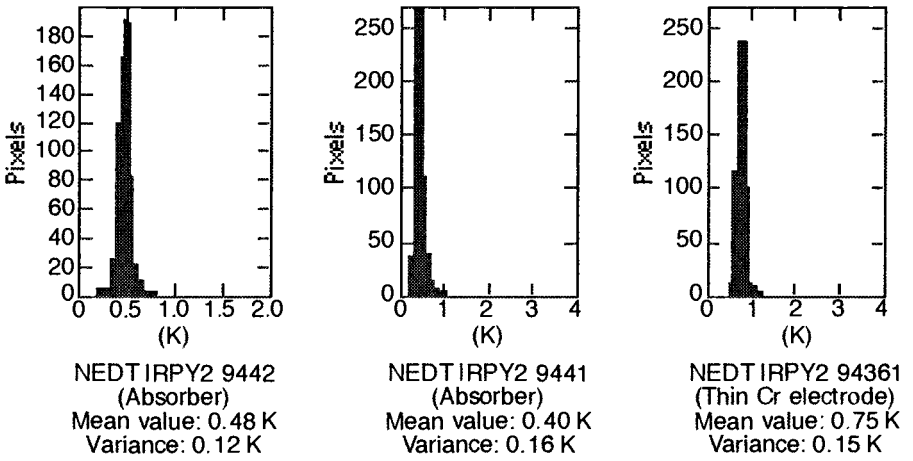


Fig. 7.10. NETD histograms of detector arrays.

7.4.2 BARMINT Project

The objectives of the BARMINT (Basic Research of Microsystem Integration) project are to set up a methodology, to define simulation tools, and to implement generic techniques requisite to a microsystem integration. Design approaches are validated and a 3D-stacked microsystem demonstrator is fabricated to check the silicon processing techniques, the design tools, and the general assembly process. The goal is to integrate a micropump actuated by a resistance heater, pressure sensors, chemical ISFET (ion-selective field effect transistor) sensors, temperature sensors, IC for signal processing, an optical module for the power source, and a test module that includes strain gauges and temperature sensors to give information on stresses during the annealing phases. The whole package is to be molded in a compact plastic cube of 1 cm^3 (Fig. 7.11). This basic research demonstrator project has been implemented to check the design and the technology approach, the functionalities and the behavior of sensors, and more especially, the integration of a micropump. The results of this project are expected to determine whether this methodology could be applied for medical and space applications.

The aim is to create microsystem structures in a monolithic package, based on existing VLSI approaches (CMOS, BiCMOS, etc.) with postprocessing techniques that involve single- and double-face micromachining, etching, and active layer deposition, in order to achieve a multisensor system, including a fluidic element, coupled with its basic integrated electronics. The BARMINT project will use existing computer-aided-design tools, such as the SABER simulator associated with the MAST language. Mechanical aspects of the microinstrument will be designed by using ANSYS software, which can do 3D calculation of mechanical parts. In addition, the MCM-V procedure will be validated on this demonstrator project. The specific modules to be fabricated are the micropump, the photovoltaic converter, and the multisensor package.

7.4.2.1 Micropump

The micropump consists of two substrates bound together: one is alumina with a deposited Ta_2N resistor (the thermal actuator of the micropump); the other is a micromachined silicon substrate having one cavity, two holes for liquid inlet and outlet, and two additional polysilicon valves. Upon fusing the two substrates, the cavity region is sealed. A current through the Ta_2N resistor raises the temperature and induces flexure of the silicon membrane, which moves the liquid

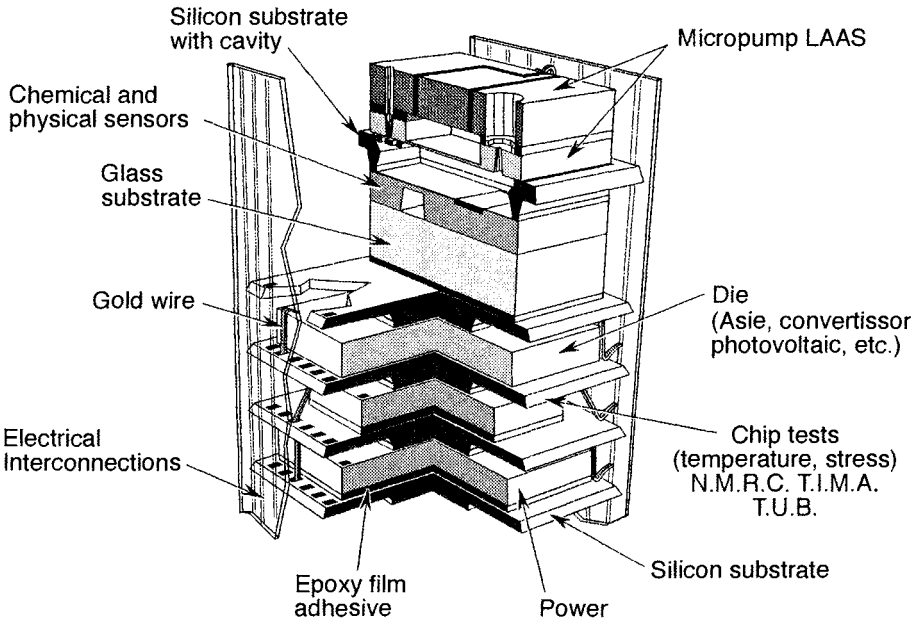


Fig. 7.11. The BARMINT microsystem schematic view.

through the physical and chemical sensors integrated in the bottom parts of the micropump. This operation can be repeated by applying a periodic voltage across the resistance to produce the liquid flow. The liquid flow direction is determined by the geometrical shapes of two microvalves and the pressure difference across the valve popette (Fig. 7.12). The polysilicon microvalve is made of a polysilicon valve lid and supported by four polysilicon supports. The valve was processed by CVD (chemical vapor deposition). The valve is designed for one-way operation. If the liquid pressure under the valve lid is high enough to deflect the four arms of the valve, the liquid

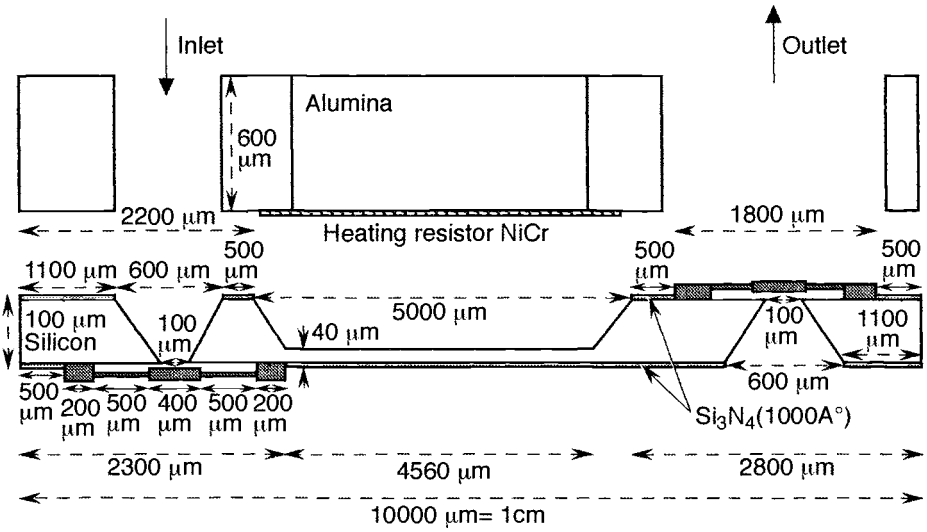


Fig. 7.12. BARMINT: detail of the micropump unit.

flows through the narrow gap between the valve lid and the substrate; the valve is open. When the liquid flows in the reverse direction, the valve lid is pushed against the substrate closing the gap and stopping the flow. The following deposition and etching techniques give details on the fabrication of the micropump.

- Si_3N_4 and SiO_2 deposition on the alumina for mask-making; masks used for chemical etching and passivation
- Spin-on-glass deposition used as a sacrificial layer
- KOH anisotropic etching polysilicon deposition used to fashion the microvalve structures for the micropump
- Photolithography and dual-face alignment

7.4.2.2 Photovoltaic Converter

The micropump consumes most of the power, which is several hundred mW (ac voltage); the other elements require only 50–100 mW (dc voltage). The option of supplying energy through a photovoltaic converter has been selected in this project. Light energy supplied by a laser is guided to the photovoltaic cells through optical fibers. The converter includes six rows of eight integrated silicon cells in series that supply a total voltage of 3.3–5 V (Fig. 7.13). This design solution is convenient for "high-power" modules. The light conversion yield is estimated to reach 15%, but is currently between 5 and 10%.

7.4.2.3 Multisensor Module

This module includes chemical sensors (i.e., pH measurements) based on an ISFET, physical sensors, such as pressure and temperature, and all the associated electronic circuits for signal processing integrated in one package. The module is positioned beneath the micropump and is in contact with the fluid. These elements provide a measure of the temperature and pressure of the fluid, parameters that are used to control the operation of the micropump (Fig. 7.14).

7.4.2.4 Packaging Process

The MCM-V stack begins with the selection of the substrate material, which may be either a silicon substrate or a flexible substrate made of plastic. A silicon substrate offers the advantage of good thermal conductivity (for power dissipation) and a high Young's modulus. In addition, during the processing, etching can be fast (1.5 μm/min for a KOH wet etch). In general, reproducibility is good, and the accuracies can be done to the micrometer scale. Other advantages are compatibility with CMOS technology and the use of the cheaper polysilicon material, for example,

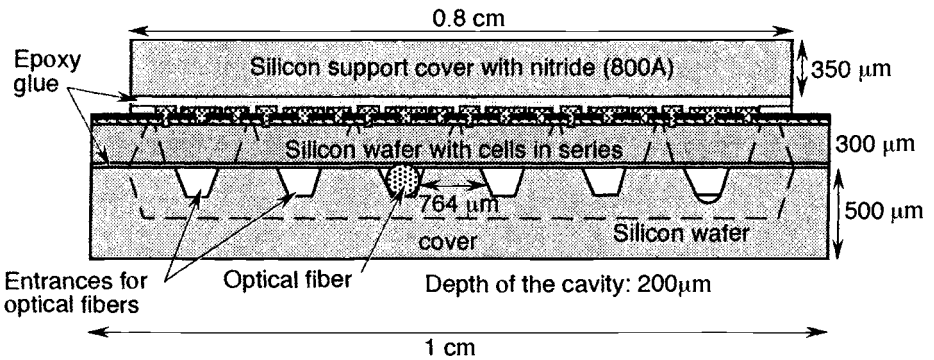


Fig. 7.13. BARMINT: side view of the power module.

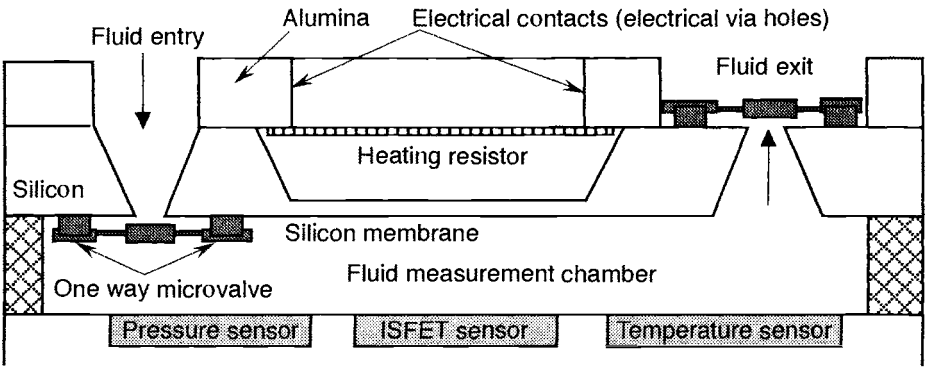


Fig. 7.14. BARMINT: schematic view of the multisensors.

porous silicon. The disadvantage in using a semiconductor is that electrical lines must be insulated by a dielectric layer. The different steps of the packaging process based on silicon substrates follow.

1. Silicon wafer double-side polished, (100) orientation
2. Deposition by LPCVD (low-pressure CVD) of Si_3N_4 (thickness 90 nm)
3. Deposition by evaporation of Cr (50 nm) and Au (500 nm)
4. Photolithography step: definition of electrical lines/bonding pads
5. Wet etching of gold and chromium
6. Photolithography step: definition of windows and holes
7. Photolithography step: opening Si_3N_4 mask
8. Anisotropic etching with KOH based solution

Flexible substrate, such as plastic material, appears to be less costly, but a poor performer in comparison to silicon. Its advantages are that it is a dielectric and its plasticity allows some flexibility in the thermal-mechanical constraint parameters. The main disadvantage of flexible film is its limited machining by stamping or milling; for example, the industrial flexible film used (FR4) does not support laser micromachining techniques. (However, other plastic films such as polyimide allows processing by eximer laser.) The BARMINT approach uses two polymer layers to achieve a standard substrate, and three layers when channels or cavities have to be fabricated. Each layer is machined, and gluing is used in the assembly. The different steps of the process are

1. Flexible epoxy or polyimide film with an average thickness of 100 mm
2. Stamping and/or milling the two films: definition of the opening design
3. Rolling a 35-μm copper leaf onto the first film
4. Photolithography of the copper: definition of electrical lines
5. Electrochemical depositing copper, nickel, and gold
6. Assembling the two layers by gluing

The next step is the chip-to-substrate assembly process. In order to keep the process uniform, the epoxy resin preform concept was selected, which provides perfect alignment between the chip/substrate and the preform. The preform can be reworked before final curing (see Fig. 7.3). The setting by stamping techniques allows shapes of 500 μm with a linear accuracy of 5 μm. The curing temperature recipe is 150°C for 30 min, which is followed by gluing the chip and coating the backside. Electrical bonding between chip and substrate is carried out by wedge bonding with gold wire, and shows a good metallization quality (50 mΩ sheet resistance).

In the final packaging operation, the chip/substrate assemblies are glued on top of one another with the epoxy preforms. The four-level stacking is an exacting process, requiring alignment of the different substrates (or foils) with a precision less than 10 μm. This is accomplished by means of alignment holes at both ends of the tape. Space between each layer is filled with a resin approximately 1 mm thick. The embedded silicon-coating epoxy takes up 80% of the volume, which is reduced after the cutting operation. Following the stacking and cutting is a molding process: the molding glue is an epoxy resin with remarkable physical properties—CTE (coefficient of thermal expansion) = 19 ppm/°C, to be improved to 9 ppm/°C. This critical step is the injection of the glue under slight vacuum in order to eliminate bubbles from the stack. After curing, the module is a solid-state block. The next step is trimming the excess resin on the sides of the block; a thickness of 1 cm is removed on the side with the holes. A few millimeters are removed on the other sides to leave a 500-μm thick resin film between the wire ends and the outer surface. The resin is removed with a diamond saw to minimize resin fracture and damage to the wires. It also gives a good surface finish. The resulting process cube shows on the lateral sides 35-μm sections of gold, which are used for interconnection.

To avoid using expensive metallization interconnect, the technique for plating PCB through-holes is used for simultaneous metallization of the six sides of the cube: a binder layer of 2-μm nickel is chemically deposited, and a conductive layer, 5–7 μm thick, is electrochemically deposited, followed by a nickel diffusion barrier and a gold protective layer.

Laser writing is the final operation for patterning the vertical connections. An Nd-YAG (neodymium yttrium aluminum garnet) laser beam is used for the patterning process. Patterns are placed on any of all sides of the block (see Fig. 7.3). To limit cross-talk, a ground plane is provided between all signal lines. A visual form-recognition device, working in conjunction with the laser, detects locations of conductors and controls positioning. Each quadrant of the cube carries lines for a specific microsystem function: two sides are dedicated to electrical signal routings, another to cooling by using conductors appearing after sawing the mini heat sinks, and the final quadrant to fluid input and output for the micropump.

Industry created this packaging technique, at first for a pure electronic application to extend the memory capacity of a module, and then with cooling lines to remove the excess heat. The space industry has also shown an interest in this packaging concept.

7.4.3 MCM-V Applications in Industry

Industrial investigations have been conducted in 2D and 3D packaging for potential use in space applications. Alcatel and IBM have developed a 320-Mbit MCM module that uses flip-chip packaging technology for all the chips; this technology has been space qualified and will be used in the SPOT V/HELIOS II satellite. The next step in this integration is a module incorporating 10-DRAM chips consisting of 16 or 64 Mbits. These are mounted with the ASICs and packaged using MCM-V technology.[27] The size is 19.5 × 16 × 19 mm, and weight is 13 g (see Figs. 7.15 and 7.16).[28]

The most advanced 3D-packaged microinstrument that will fly in space is a digital CCD microcamera[27] to go on the ROSETTA mission (WIRTANEN comet space physics and composition studies). The challenge is the travel mission, which takes 10 years to reach the comet surroundings. The microinstrument payload is to provide stereovision, using several microcameras. Four outputs from the CCD provide the video information; each is processed with four analog video circuits, followed by four sample and hold circuits, A/D converters and multiplexers, totaling about 290 components. The microcameras and associated components will be set up on the "Champollion lander"[29] for providing stereovision of the comet surface (Fig. 7.17). The mission

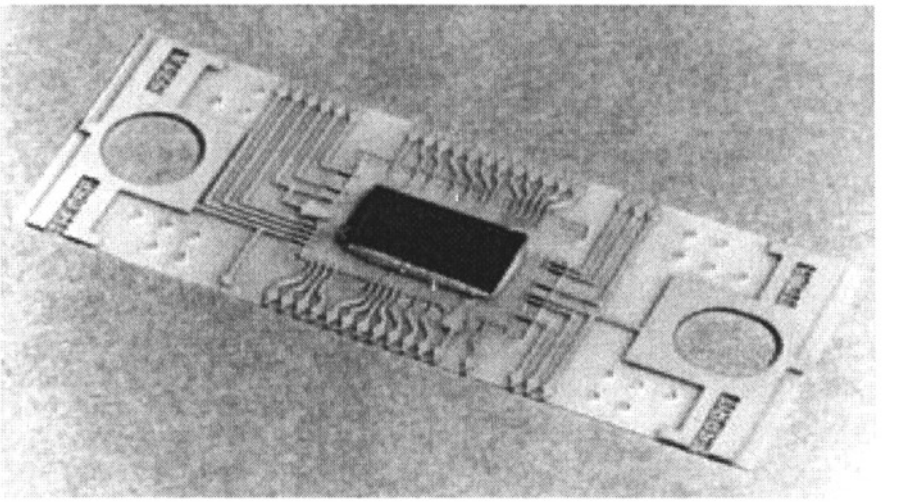


Fig. 7.15. MCM-V: wiring of chip on tape ($3D^+$).

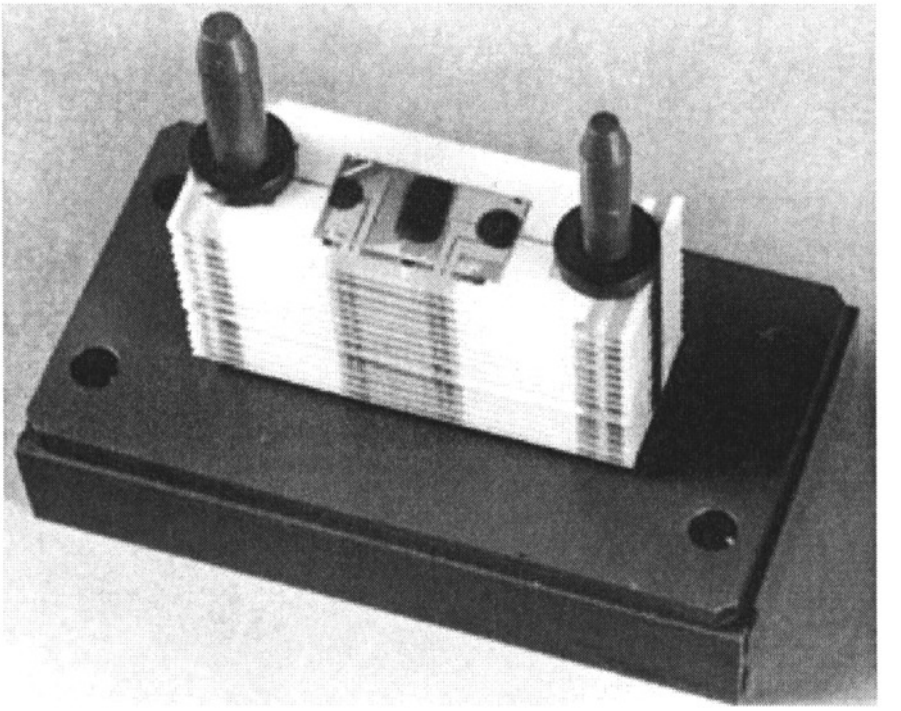


Fig. 7.16. Stack of 16 films with DRAM chips and center pad "lead on chip" type wiring ($3D^+$).

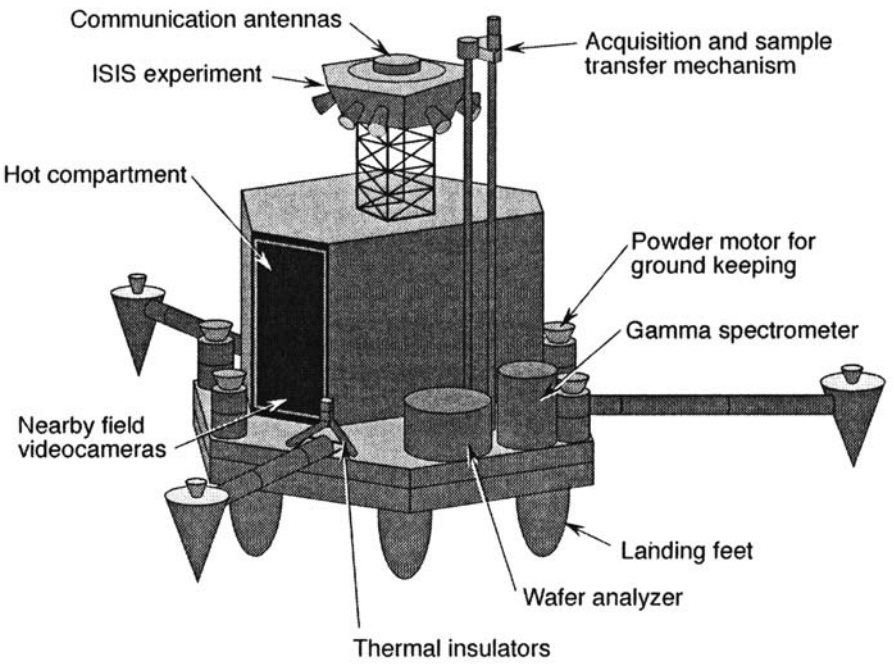


Fig. 7.17. Schematic view of the Champollion lander (CNES).

duration is a few minutes. The size of the module is 35 × 25 × 10 mm (the CCD is external to the module); the mass is 35 g (a 1024 × 1024 pixel CCD is used, optics not included in the given mass). Main specifications of the microcamera are listed in Table 7.4. The behavior of this microcamera will be tested prior to the ROSETTA mission in the experimental technological satellite STENTOR, to be launched in 2002.

Table 7.4. Main Specifications of the Microcamera

Definition	1023 × 1024 pixels
Power	Typical 1.7 W Maximum: 2.3 W <10 mW standby
Digital output	1 Mpixel/s 10 bits/pixel (1 frame/s) Double speed available
Sensitive area	3/4 in. (square)
Temperature	Storage: −150°C to +85°C
Working rate (camera output) 1 Mpixel/s	
Antiblooming included, providing an electronic shutter function	
Antireflective window in 400–700 nm bandwidth	

7.5 Reliability

7.5.1 Thermal Stresses Associated with Satellite Microsystems

In a systems-integration process, analysis of thermal properties is an important aspect of the development. Thermal control management must be incorporated into the design from the beginning. For example, controlling thermal equilibrium is much more difficult as the mass is reduced. Thermal fluctuations tend to be fast in small devices. A microinstrument or a small satellite operating in low Earth orbit (LEO) is sensitive to these effects. Real-time thermal control systems will be necessary to manage this problem, and designs for transferring heat out of micromodules will be necessary. This approach involves use of materials with large thermal conductivities, inclusion of micro-heat-exchange systems based on fluid flow or mechanical microactuation, and development of thermal micromachines. Thermal control must be managed in space applications. Although this is vexing, it is being considered for high-end microprocesssors, other wafer-scale integrated devices, and certainly MCM packaged devices.

7.5.2 Mechanical Stresses in Membranes

Crystalline silicon has mechanical properties better than stainless steel, and it is very suitable for fabrication of membranes. However, current interest is focused on polysilicon, a noncrystalline material, which is commonly used to fabricate mechanical parts like cantilevers and membranes. As a material, polysilicon has not yet been investigated extensively. One problem is that polysilicon is grown under a wide variety of forms and recipes, and grain boundary properties influence the mechanical response. Empirical methods are used to adjust the process recipe and doping levels of polysilicon to arrive at a minimum residual stress. Initial data as to the mechanical and material properties are presented in Chapter 3.

7.5.3 Thermomechanical Stresses in Packaging

Thermomechanical stresses induce failures, including breakdown of interconnection wires, changes in electrical properties, cleavage in crystal materials, destruction of membranes and micromachined cantilevers, unsticking in epoxied materials, and pressure modulation in existing microcavities following temperature variations. Thermomechanical stress can happen in the following two circumstances:

- During the assembly process, as various components are encapsulated into the resin at high temperature and are then cooled
- During operation in a temperature-cycling environment, for example, a LEO mission of a very small satellite with little heat capacity. The temperature cycling would result from the "day/night" segments of the approximately 90-min orbits.

7.5.3.1 Stress During the Assembly of the Microinstrument

For the BARMINT microinstrument, cracks in silicon substrates were detected in the initial trials after resin molding following block cutting. These defects affected mainly substrates positioned at the boundaries of the stack. This is shown in Fig. 7.18, where the reported crack on the silicon substrate is shown with a connecting wire. Mechanical relaxation of substrates during cutting caused the defects. Another type of defect is shown in Fig. 7.19, where there is a delamination between the resin and the substrate; a third is a crack that resulted from a defect made during the micromachining process. Improvements in several processing steps permitted the fabrication of five silicon substrate units without the above defects (shown in Fig. 7.20). The improvements were a reduction of the mold size to decrease the mass of resin and a related adaptation of the cutting parameters, and the optimization of the cutting parameters and the reticulation cycle[30] (defined to mean polymerization plus chain cross linking, as seen in Table 7.5).

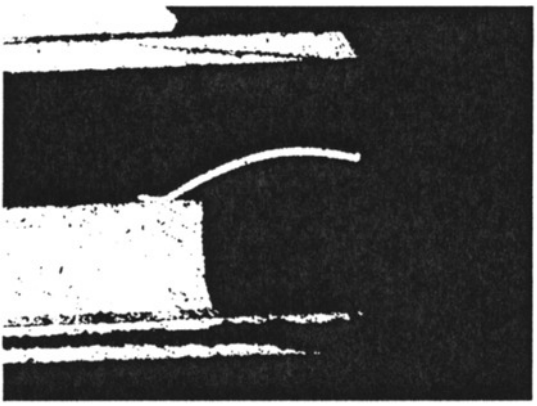


Fig. 7.18. Crack in a silicon substrate after block cutting in a 3D stack.

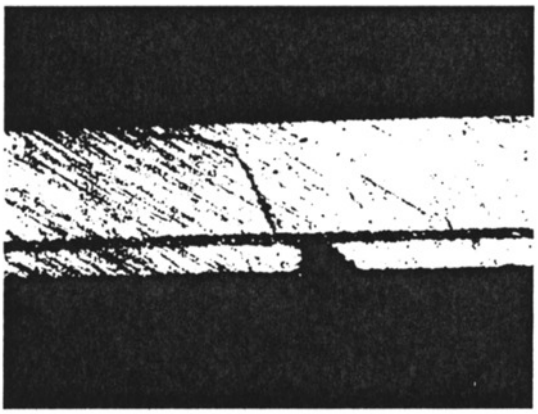


Fig. 7.19. View of delamination of a substrate and further propagation of the crack.

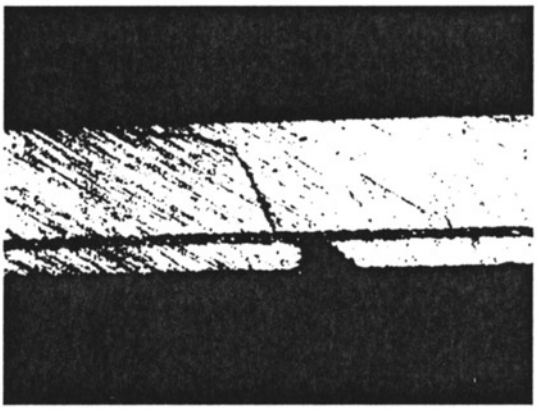


Fig. 7.20. Elimination of cracks after process improvement in a 5-silicon-substrate stack.

Table 7.5. Optimization of the Reticulation Cycle with Temperature

	Time (h)			
Reticulation Rate (%)	110°C	125°C	140°C	165°C
10	0.06	0.03	0.01	0.00
25	0.18	0.07	0.03	0.01
50	0.42	0.17	0.08	0.02
60	0.56	0.23	0.10	0.03
75	1.84	0.35	0.15	0.04
95	0.82	0.75	0.33	0.09
99	2.80	1.15	0.50	0.14

7.5.3.2 Thermal Expansion

When a structure made of two materials is in contact and submitted to a large temperature gradient, a thermomechanical stress is generated along the interface, according to Eq. (7.4):

$$\sigma = kE_A(\alpha_A - \alpha_s)\Delta T \tag{7.2}$$

where α_A and α_S are the coefficients of thermal expansion (CTE), E_A the Young's modulus, k is a geometric factor, and ΔT is the temperature difference. We consider an adhesive epoxy A, deposited on a substrate S, with a CTE α_A, which is higher than that of the substrate CTE α_S. For the chosen system, if $\Delta T>0$, the substrate experiences a compressive stress force while the epoxy undergoes a tensile stress (Fig. 7.21). This stress can be measured during the stacking process by varying the rising and falling times of the temperature. To minimize these effects, a stress model simulation was conducted to gain some information about the potential thermomechanical properties of the stack.

7.5.3.3 The ANSYS Software Simulation Applied to the Stack

A finite element method (FEM) analysis to determine the stress evolution in the resin coating was applied. The basic principle consists of approximating the continuous actual structure with a discontinuous model constituted of a finite number of elements of limited sizes, interconnected through nodes at their boundaries. By application of the virtual work theorem, the problem is restricted to the resolution of a system of linear equations of high order (often 1,000 equations up to 100,000). The unknown variables are the node displacements. The method can be applied to

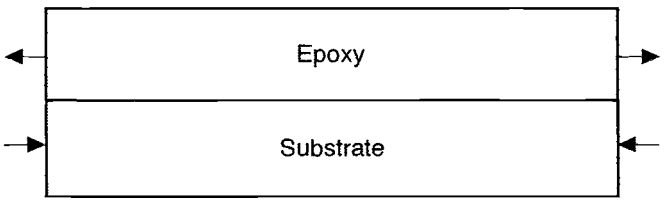


Fig. 7.21. Stresses generated at the interface of two materials.

structures made of boards, cantilevers, or bulk elements, in both static or dynamic modes. In highly stressed zones, the denser the number of nets/modes, the more accurate the solution. The software ANSYS has been used for simulating the effect of a compressive stress at high temperature close to a membrane located at the upper part of the stack and how the distribution is modified at a lower temperature. Figures 7.22 and 7.23 show these results. One typical result from the simulation is optimized geometric and assembly techniques. The straight cut windows have been modified into a T-shape to shift the high-stress zone away from the cut line (Figs. 7.24 and 7.25).

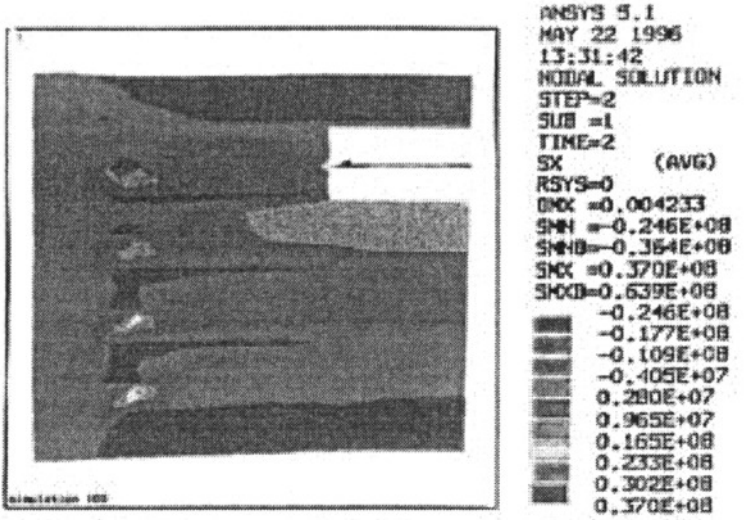


Fig. 7.22. Stress mapping inside assembly with a cavity, at +100°C (ANSYS simulation).

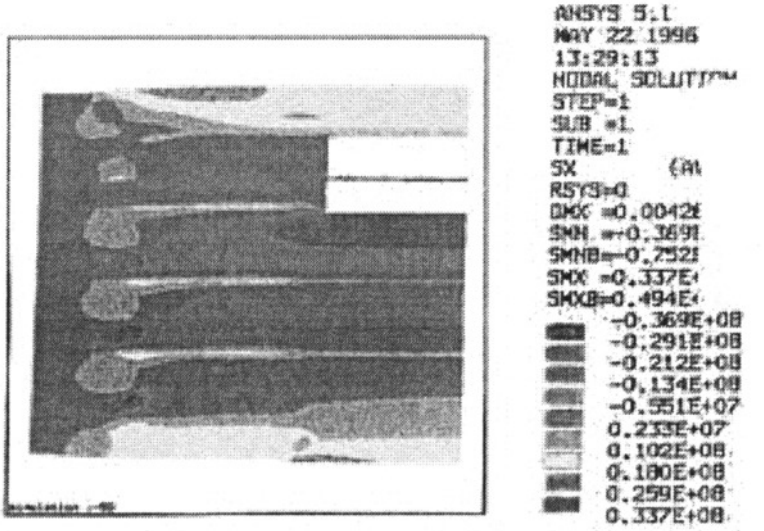


Fig. 7.23. Stress mapping inside assembly with a cavity, at −50°C (ANSYS simulation).

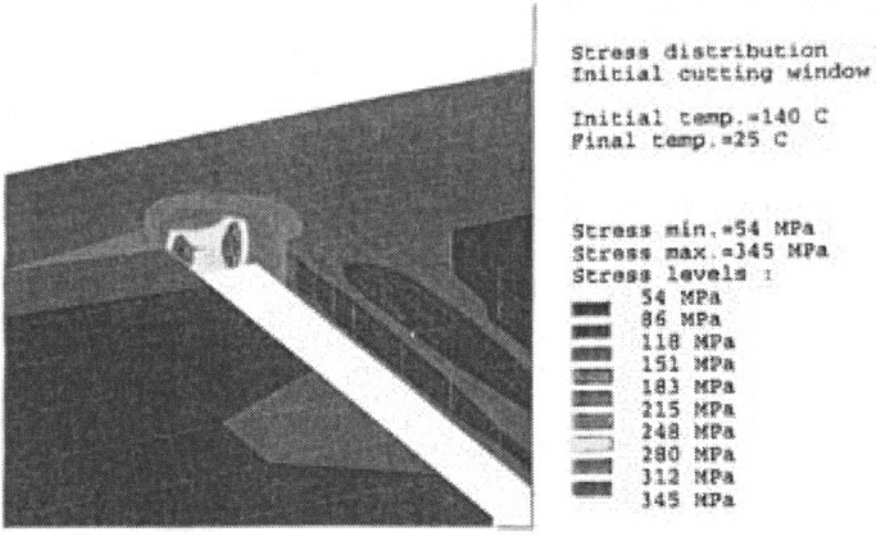


Fig. 7.24. Stress applied into silicon by a straight etched window (ANSYS simulation).

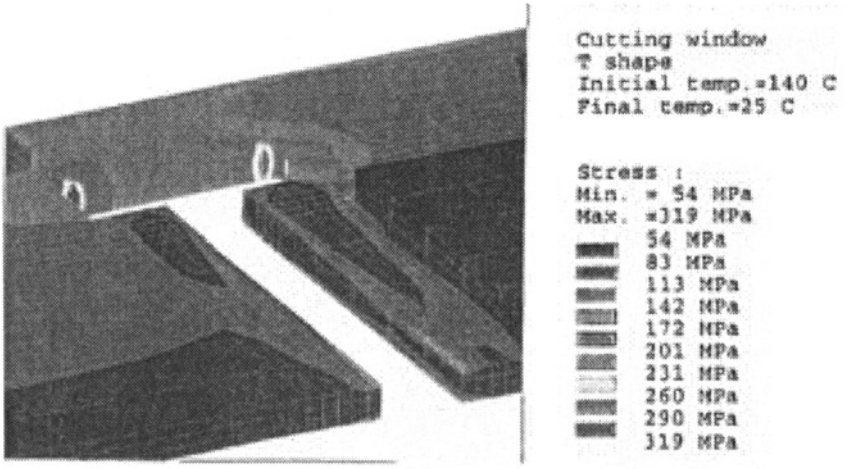


Fig. 7.25. Stress reduced at the sensitive point with a T-shape window (ANSYS simulation).

Conclusions from a number of simulations enable the formulation of general rules regarding the assembly of the microinstrument. Simulations with a pitch of 100 μm give good representation of the microinstrument properties.

- Use materials with CTE close to each other.
- Reduce the resin thickness.
- Increase the substrate thickness.
- Reduce the curvature radius by favoring a symmetrical stack and by increasing the number of layers.
- Set the thicker substrates at the bottom of the stack.
- Protect sensitive elements by sheets with CTEs lower than elements to be protected.

7.5.4 Space Reliability of Plastic Packaging

For a long time, plastic encapsulated devices were not accepted by military and space specifications. One reason for this reluctance to use the best available commercial plastic parts was a perceived risk of poor reliability. Fortunately, progress in commercial technology and data published on the relative reliability of solid-encapsulated packaging changed minds at the beginning of the 1990s. An investigation into the applications of commercial, plastic-encapsulated components for the space environment, initiated in 1994, showed they are reliable. It also called for using a deterministic methodology to assess space reliability (see Sec.7.2.9: The Deterministic Methodology). Conclusions also published by Hakim *et al.*,[31] give a positive assessment of the use of commercially-produced, component-encapsulated packaging, but caution that radiation effect is the major barrier for space application. It appears that a statistical analysis of the behavior of plastics in a space environment yields a poor reliability evaluation, which is the opposite when deterministic evaluation is used. Previously the impact of radiation on materials was evaluated per each component rather than on the whole system, which may encapsulate a radiation-intolerant component. Today radiation hardening is evaluated by considering the whole architecture of the microsystem.

7.5.5 3D Stacking Reliability Performances

The 3D stacking technology has been comprehensively evaluated through demonstrator programs for electrical, thermal, thermomechanical, and reliability performances within the ESPRIT European program.[32] The following main features tested were.

- Thermal and thermomechanical performance of the 3D-stacking package
- Structural reliability of the structure
- Reliability of the contacts between surface and internal interconnect metallization
- Long-term reliability of the overall stack

The stress test program used is given in Table 7.6. These are the main conclusions on the reliability assessment:

- A small thermal resistance variation in modules was noticed before and after the evaluation program. This indicates a minimum physical change in the structure, as such effects could be generated by delaminations and internal cracking of the resin.
- Thermal and thermomechanical performances are good enough to allow the use of the technology with no additional cooling.
- No corrosion of the surface metallization was observed. Triple-track aluminum meander structures were used on the ICs to monitor corrosion on the witness structures and the on-chip wire bonds during highly accelerated stress testing (HAST), an environment test used to check corrosion resistance to salt air.

7.6 Conclusions

This chapter has examined a global approach to microsystem development. The design approach should be an extension of the top-down methodology developed for microelectronics, including the multidisciplinary aspects required for microsystem development. We believe this is achieved by the use of a "behavioral" modeling language like VHDL-AMS with a multifunctional/multidisciplinary simulation capability.

Regarding packaging technologies, the choices are either the monolithic approach based on standard IC process plus postprocessing steps or the hybrid approach of the MCM type. Both approaches must be adapted from microelectronics to microsystem applications. The 3D-stacking approach should be an attractive way to package the most compact assemblies.

Table 7.6. Test Program Used in the ESPRIT Project for 3D-Stacking Reliability Assessment

Test	Test Conditions
Mechanical shock	5 shocks 300 g/0.5 ms
Temperature cycling	100 cycles per step to 500 cycles (−55°C, +125°C)
Thermal shock	0°C, +100°C
High-temperature storage	+125°C, 2000 h, unbiased
Humidity	2000 h, unpowered, 40°C, 95% RH
HAST	40 h, 110°C, 85% RH
Endurance	2000 h, 125°C, powered
Vibration	50 g
Power cycling	500 cycles on/off

The ultra-thin-chip-stack (UTCS) approach is also very promising but is currently in its infancy. More research effort is needed before UTCS technology challenges the other approaches. Both monolithic and hybrid microsystem packaging technologies successfully tested on demonstrators that can be transitioned to space applications.

Regarding reliability, it is preferred that thermomechanical stress-analysis tools be applied for the assembly process and for the specific application as well (i.e., a space mission). To better assess microsystem reliability, deterministic rather than statistical methods should be used, even if initial results reveal no cost savings. For space applications, radiation effects are still considered to be the major issue for microsystem technology insertion.

7.7 Acknowledgments

I wish to thank Dr. D. Esteve for reviewing this chapter and providing helpful discussions to add value to this work; C. Alonzo, B. Jammes and M. Couzineau for discussions dealing with microsystem design approaches; and A. Coello-Vera from Alcatel Espace, T. Duhamel from Matra Marconi Space, J. P. Fortea from CNES (Centre National d'Etudes Spatiales), C. Val from $3D^+$, and A. Val, Z. Sbiaa, and the BARMINT team for their contributions.

7.8 References

1. S. Janson, "Spacecraft as an Assembly of ASIMS," *Microengineering Technology for Space Systems*, monograph 97-02, edited by H. Helvajian (The Aerospace Press, El Segundo, CA, 1997) pp. 143–201. First published as Aerospace Corp. report ATR 95 (8168)-2 (1995). See also: A. Heuberger, "Silicon Microsystems," Microelectronic Engineering, 21, 445–458 (1993).
2. G. Massobrio and P. Antognetti, *Semiconductors Device Modeling with SPICE*, 2nd ed. (MacGraw-Hill, New York, 1988).
3. J. A. Connelly and P. Chol, *Macromodeling with SPICE* (Prentice Hall, Englewood Cliffs, N.J., 1992).
4. D. Pellerin and D. Taylor, *VHDL Made Easy* (Prentice Hall, 1996).
5. H. E. Tahawy, D. Rodriguez, S. Garcio Sabiro, and J. J. Mayol, "VHDELDO: a New Mixed Mode Simulation, *Proceedings IEEE International Conference on C.A.D.* (1993), pp. 546–551.
6. *SABER Reference Manual*, Release 3 1a (Analogy Inc., February 1987).
7. R. A. Saleh, D. Rhodes, E. Christen, and B. A. A. Antao, "Analog Hardware Description Languages," *Proceedings IEEE Custom Integrated Circuits Conference* (1994), pp. 15.1.1., 12.1.8.

8. MAST Modeling Class, Modeling Language for the SABER Simulator (Analogy Inc., October 1994).
9. H. C-C. Chang, A Top-Down, Constraint Driven, Design Methodology for Analog Integrated Circuits, Memorandum UCB/ERL, M95/21, University of California, Berkeley, April 1995.
10. E. W. Becker, W. Ehrfeld, P. Hagmann, A. Manerand and D. Munchmeyer, "Fabrication of Microstructures with High Ratios and Great Structural Heights by Synchrotron Radiation Lithography, Galvanoforming, and Plastic Molding (LIGA Process)," *Microelectronic Engineering* **4**, 35–56 (1986).
11. H. Lehr, "New extension of LIGA technology," *Micromachine Devices*, 3–13 (November 1996).
12. M. J. Madou and S. R. Morrison, *Chemical Sensing with Solid-State Devices* (Academic Press Inc., 1989).
13. D. A. Doane and P. D. Franzon, *Multichip Module Technologies and Alternatives—The Basics* (Van Nostrand Reinhold, 1993), p. 160.
14. P. Lall and S. Bhagath, "An overview of Multichip Modules," *Solid-State Tech.*, 65–76 (September 1993).
15. "Unique Epoxy Resin and Printing Encapsulation System for Advanced Multichip Modules, PLCC, BGA, PGA, TAB, COB, and Flip-Chip," *Int. J. of Microcircuits and Electronics Packaging* **17** (2), (1994).
16. J. C. Lyke, "Packaging Technologies for Space-Based Microsystems and their Elements," *Microengineering Technology for Space Systems*, monograph 97-02, edited by H. Helvajian (The Aerospace Press, 1997), pp. 103–141.
17. R. R. Tummala, "Multichip Packaging—A Tutorial," *Proceedings IEEE* **80** (12), (1992), pp. 1924–1941.
18. R. R. Tummala, "Multichip Packaging in IBM—Past, Present and Future," *Proceedings Int. Conf. on Multichip Modules* (Denver, CO, 14–16 April 1993), pp. 1–11.
19. C. Val and T. Lemoine, "3D Interconnection for Ultra Dense Multichip Modules," *Proceedings IEEE 40th Electronic Components and Technology Conf.*, Las Vegas (May 1990), pp. 540-547.
20. C. Val, "Challenge to Densify Packaging in Europe," *Proceedings IMC* (Tokyo, 30 May–1 June 1990), pp. 29–37.
21. Thomson CSF and Alcatel Espace (private communication).
22. BARMINT European ESPRIT Project no. 8173.
23. EUREKA/PROMETHEUS/PROCHIP, Project no. 2211. See also: D. Esteve, P. A. Rolland, J. J. Simonne, and G. Vialaret, "Prometheus-Prochip: Status of Sensor Technology Applied to Automotive Collision Avoidance," *Proceedings SPIE: Infrared Imaging Systems Design, Analysis, Modeling, and Testing VI*, Vol. 2470 (1995), pp. 386–395.
24. J. J. Simonne, P. Bauer, L. Audaire, F. Bauer, Workshop on the Technology of Ferroelectric Polymers (Albuquerque, NM, 17–19 October 1995).
25. R. A. Wood, "Uncooled Thermal Imaging with Monolithic Silicon Focal Planes," Proceedings *SPIE Infrared Technology XIX*, Vol. 2020 (1993), pp. 322–329.
26. F. Bauer, French patent 8221025, U.S. patent 4611260.
27. $3D^+$ (private communication), and C. Val, "Challenge to Densify Packaging in Europe," *Proceedings IMC* (Tokyo, 30 May–1 June 1990), pp. 29–37.
28. A. Coello-Vera and M. Masgrangeas, "256 Mbits MCM-V Memory Stack," *Proceedings Int. Conf. on Multichip Modules* (Denver, CO, 19–21 April 1995), pp. 24–29.
29. Centre National d'Etudes Spatiales (CNES) (private communication).
30. A. Val, Thesis no. 2529 (Univ. Paul Sabatier, Toulouse, France, 9 December 1996).
31. E. B. Hakim, R. K. Agarwal, and M. Pecht, "The Demise of Plastic Encapsulated Microcircuit Myths," *Agard Conf. Proceedings562: Advanced Packaging Concepts for Digital Avionics* (San Diego, CA, 6–9 June 1994), pp. 9/1–7.
32. A. Coello-Vera and M. Hayard, *Proceedings 47th Astronautical Congress* (Beijing, China, 7–77 October 1996).

8

Space Electronics Packaging Research and Engineering

J. Lyke* and G. Forman†

8.1 Introduction

Packaging provides an essential bridge between the materials and structures that represent electronics components to an end user or another bridging assembly. Examples include computer chips, printed wiring boards (PWB), digital watches, and satellites. Packaging is recursive or hierarchical in nature. Usually an item that can be handled is a package for things within, and those things within are packages for other things, and so on. Packaging can be identified as having several fundamental roles or functions, such as getting electrical power in and waste heat out. It is possible to show that those roles change or expand as more is expected from packaging. Tiny structures that interact with the environment, sense light, and channel fluid are examples of things that define new roles for packaging.

While packaging represents many key scientific disciplines, including electrical engineering, mechanical engineering, polymer chemistry, material science, and mathematics, the practice of packaging is both *ad hoc* and driven by *de facto* standards, the latter of which define various facets of an infrastructure. This infrastructure represents a barrier to evolution, regardless of whether that evolution represents beneficial improvements to the technological status quo or, in the case of new technologies, if that evolution represents an expansion in the traditional roles.

This chapter will introduce basic concepts in packaging technology, not only to provide a better understanding of related technologies and concepts, but to provoke the extension of packaging beyond traditional roles to include microelectromechanical systems (MEMS). The confines of the chapter prohibit prescribing a discipline for inventing new forms of packaging, but it should motivate a better understanding of some basic principles of engineering electronics packaging solutions for systems. We hope also that it will make clear why packaging is needed and has evolved into its present form and why there may be considerable room for improvement.

The following section presents some basic concepts in packaging technology. Section 8.3 addresses the principle of engineering packaging solutions and introduces new, nontraditional technologies, such as MEMS. Finally, Sec. 8.4 examines a few advanced packaging concepts and case studies and discusses a new three-dimensional (3D) heterogeneous packaging framework under development for compact packaging of high-performance systems.

8.2 Basic Concepts

Almost everything looks like a package! Packaging is an arrangement of separately fabricated structures that achieves a desired function, which is different from monolithic systems, where all structures are formed in a single overall fabrication process. In fact, it is almost always necessary to package even monolithic structures.

Many systems, even entire platforms such as satellites, may be viewed topologically as electronics packages. This view is useful, even necessary, when optimizations of size, weight, power,

*Air Force Research Laboratory, Kirtland Air Force Base, New Mexico

†General Electric Corporate R & D Center, Schenectady, New York

and performance reduction are considered. Not doing so sometimes results in suboptimal treatments of a packaging problem, which in some cases is a desired result. It is often necessary to retrofit a complex function within an existing nonoptimal system, where very tight size, weight, and power budgets exist for that function. In these cases, little opportunity exists for shrinking the particular circuit board involved, much less an entire satellite. It is tempting for a technologist to address wider optimizations than may be practical. In fact, the ripple effect of packaging optimization can be so dramatic as to provoke a large-scale system redesign. In satellite systems, which do not have to deal with the issue of accommodating the handling by "parasitic humans," the possibilities can be staggering and have led to serious consideration of micro- and nano-satellites. In some cases, size and weight minimization possibilities are limited only by the size of apertures needed for communications and energy gathering from solar panels, which must have surface areas consistent with physical reality. Further considerations of even these problems have led to discussions of deployable, even inflatable structures. In the vacuum of space, bereft from most of the deleterious effects of gravity and atmosphere, many creative possibilities exist for minimizing the ratio of stowed volume to deployed surface area, which might be viewed as a metric of packaging optimization in these platforms.

Packaging has structure and function, but it is predominately viewed as overhead, a price to be paid for accessing the inherent power of an integrated circuit or for protecting delicate sensors. Technologically, it has not yet been demonstrated that the one-piece or monolithic fabrication of an entire system on chip is possible. The earliest attempts to build wafer-sized chips began in the 1960s and probably led to the whole notion of cutting apart wafers into dice. There existed at that time no integrated circuit (IC) industry to prejudice researchers; it was economic reality that dictated the severing of silicon planes into yieldable pieces. Similarly, as more functions are expected from a monolithic process, complexity grows and yields shrink. This basic realization motivates packaging and its hierarchy.

Packaging (or a hybrid arrangement of components) is not that bad (as an alternative to monolithic construction). By combining the best-yielding pieces of things built in the same or different processes, systems can be formed with some practicality. The idea has a lot of merit, since it is possible to put within a common package components fabricated from processes that would be incompatible in a monolithic process. Furthermore, as most IC processes are planar, packaging provides more options to be creative with geometries in the engineering of systems functions. For example, rather than attempting to build a three-axis accelerometer or gyro in a monolithic process, it is in many cases more practical to build three optimal planar single-axis units and rearrange them as an orthogonal set through packaging.

8.2.1 Roles of Packaging

Traditionally, packaging plays at least four roles in electronics: (1) power delivery to constituents, (2) signal mapping from outside a package to within and between the constituents, (3) thermal management, and (4) environmental protection. These roles, represented schematically in Fig. 8.1(a), must be minimally accommodated for every electronics element and system. In microsystems, particularly those involved with MEMS devices, the traditional roles of packaging may need to be expanded to accommodate a number of other roles. Examples of these extensions, as suggested in Fig. 8.1(b), include light (photon) transport and mapping, fluidic transport and mapping, mechanical (linear and rotary coupling), and environmental access for chemical sensing.

Power delivery refers to the mechanism by which electronics within a package assembly receive electrical energy to effect operation, usually by electrical conductors, in some cases similar to those used for signal transport but sometimes more robust to ensure low resistive loss. Sometimes the system can receive energy in a nonelectrical form and convert this energy into electrical

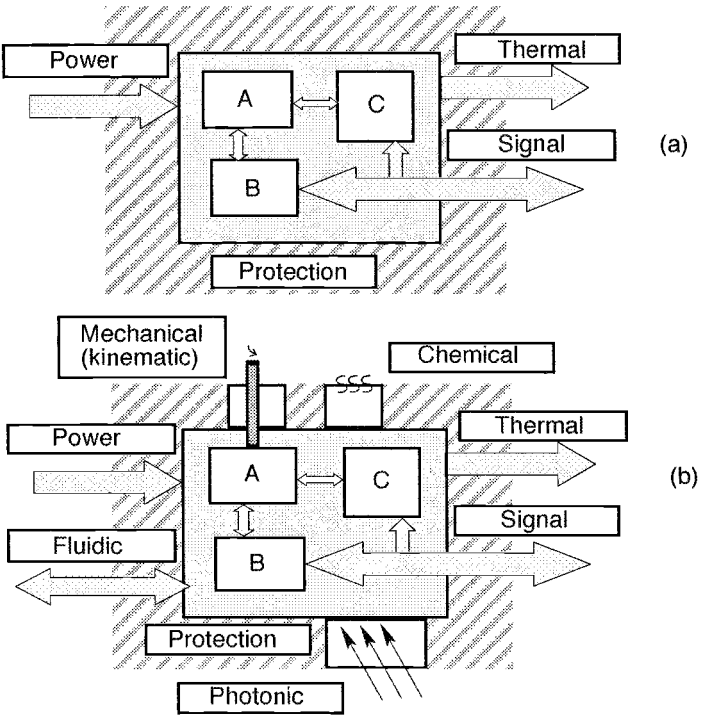


Fig. 8.1. Functions of packaging: (a) traditional (b) expanded.

form. Solar cells in a self-contained system are one example. A fluidic coupling could be envisioned in which a tiny turbine might generate energy. In very low-power systems, it is possible to exploit power conversion from radio-frequency (RF) electromagnetic waves, as is done in some commercial RF identification systems (although it must be indicated that the RF power can be distinct from RF communication). In some cases, a system or element has no external power source, usually only in the cases where temporary operation is required and/or the element is serviceable for replacement of the internal power source. Consumer electronics are common examples. It is possible to mix modes of power delivery, most notably where an internal source is regenerated from an externally coupled electrical or nonelectrical source.

Signal mapping is the most often regarded function of packaging. In the case of ICs, this function is accommodated internally through multilevel interconnections and externally through terminals referred to as bond pads. In traditional packages, an internal shelf usually is the attachment point for electrical component terminals within the package, and a wiring system is mapped to the outside of the package to facilitate, for example, attachment of a package to a PWB. When multiple nonmonolithic components are placed within a single package, then the resulting assembly is referred to as a multichip module (MCM). Connectors are the most common high-level signal access method for complex electronics assemblies, such as PWBs and boxes.

The notion of "plug and play" deserves some particular mention here. Packaging is the "plug" part of plug and play, usually in the form of connectors. As the role of packaging is to provide not only the capability to access functionality of components within, plug and play is suggestive of the idea of obtaining this access automatically or autonomously. Are there attributes that packaging can provide to extend the concept of plug and play beyond the relatively constrained form in which it is practiced today? Such attributes might include:

- The ability to physically reroute signal and even power interconnections
- The more intimate connection of mating plug and socket "surfaces" (for example, by implementing a MEMS version of a zero-insertion force socket whereby many thousands of signals could be mapped, aligned, and connected in a seamless, automatic way).

Clearly, the plug-and play concept is more than physical packaging structures, but perhaps the coordinated physical (plug) and logical (play) aspects of interfacing could themselves be recognized as another role that packaging supports, one pertaining to system configuration.

8.2.2 Packaging Hierarchy

A painful lesson learned by many over the last decade is that MCMs alone provide only a partial solution to packaging advancement. Often, MCMs are invoked as cure-alls in a technology development program, yet in many cases the MCMs do not achieve the dramatic improvements suggested by simply comparing the MCM substrate to the bulk of the packaged components replaced. The key to achieving system-level improvements in electronics density clearly lies at the higher levels of what is referred to as the packaging hierarchy. PWBs, boxes, and even the system platform itself interact with, and affect the density of, possible packaging solutions.

In the traditional packaging hierarchy, level 0 (L0) refers to interconnections between individual transistors on an IC. Level 1 (L1) interconnect refers to the transition from diced wafers to a package. Level 2 (L2) interconnect refers to the package-to-package or PWB interconnect. Level 3 (L3) interconnect commonly refers to the interchassis board-to-board interconnect. Finally, level 4 (L4) refers to the system platform when viewed as a collection of electronics chassis components linked together by cables and connectors.

The traditional packaging hierarchy (Fig. 8.2) has occurred as a happenstance over the last 4 decades. Its present form is convenient and represents the embodiment of most modern electronics. The traditional hierarchy, however, is limiting at higher levels (above level 2), representing the substantial reliance on monolithic ICs for rendering the vast majority of today's electronics.

8.2.2.1 On-chip Interconnections (L0)

Formed in some cases billions at a time, the IC interconnect has enjoyed a substantial investment since the monolithic IC was invented in the 1960s and has been responsible for explosive performance advances. The improvements in IC performance and density have been so powerful that the IC has dwarfed and superseded many other packaging technologies, beginning with the abortive attempts of wafer scale integration (WSI) in the 1960s. Many attempts have been made since then to augment and accelerate the gains made possible through monolithic IC improvements alone. Since monolithic IC integration has been so successful, little incentive has existed in the electronics industry to seek substantial improvements in higher levels of the packaging hierarchy, a consequence being poor efficiency in most of the packaging hierarchy above L0.

8.2.2.2 Chip-to-Packaging (L1), Hybrid MCM/Substrate (L1.5)

Chips are typically interfaced to anything else through packages, which are often regarded as the first true level of the hierarchy. MCMs are sometimes referred to as level 1.5, since they are often treated as the more traditional packages that predominately contain only one component. Packages to IC designers represent a severe "speed bump" caused by the sometimes significant parasitic impedances of associated packaging structures. Herculean efforts are made to keep things within a single die.

8.2.2.3 Package-to-Package (L2)

Level 2 is usually synonymous with "boards." The first significant forward step and "kink in the armor" in the current L2 packaging is the Pentium II processor, which employs a cartridge for

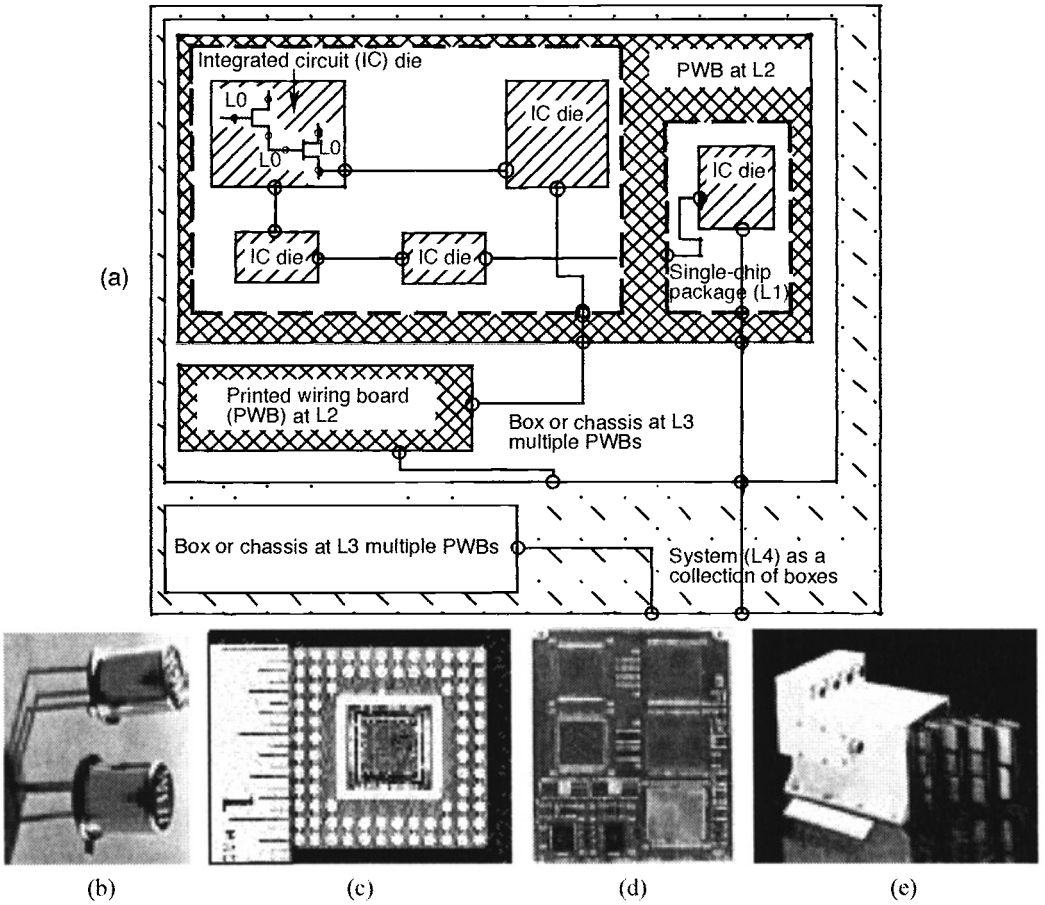


Fig. 8.2. Traditional packaging hierarchy. (a) schematic representation. (b) example L0 (transistor), (c) example L1 (package) (inch scale), (d) example L2 (board, 8 × 12 in.), (e) example L3 (box, 15 × 9 in.).

electrical and thermal management, rather than the previous Pentium (single-chip packages) and the Pentium Pro (a fairly crude hybrid). L2 tends to be an inefficient integrating medium for MCMs, sometimes more dramatically so, as a result of package bulk and fanout requirements.

8.2.2.4 Board-to-Board

Level 3 integration is typically for complex systems, and includes the integration of multiple boards into a chassis. As such, a personal computer is a typical L3 system. The L3 packaging inefficiency is embodied in traditional bus specifications such as VME (virtual-memory environment). Because the speed bump is so severe (in the sense of the interconnections between transistors of nonmonolithically integrated components), bus-specifications affect the clock skew much worse orders of magnitude than those possible within L0. System efficiencies measured as volume utilization show that conventional packages have efficiencies of less than 1% and as low as 5% even when MCMs are used in a system.

8.2.2.5 System

Level 4 is box-to-box interconnect, usually the highest level present in a system, such as an aircraft or satellite. The penalties of cable and connector in weight, propagation delay, and power

consumption are so severe that when these barriers can be overcome, it is not uncommon to consider as possible as two orders of magnitude improvement in size, weight, and power.

8.2.3 Packaging Metrics

Metrics in packaging attempt to quantify some benefit from a change of a technology (improved wire-bonding) or an approach (3D vs 2D). The desire to obtain improvements to size, weight, and power promote the development of metrics that reflect these characteristics. Improvements in performance could suggest other metrics. Then there are technology-specific metrics, such as wiring and contact density, which pertain to attributes less meaningful to end users, but can be applied to assess the adequacy of a technology to achieve a certain required performance level. As such, we regard two classes of metrics: reduction metrics and performance metrics, some of which remain undefined.

8.2.3.1 Reduction Metrics

Reduction metrics denote improvements in density of packaging. The most meaningful measure of planar size reduction is substrate efficiency, defined as

$$\eta = \frac{A_{comp}}{A_{sub}} \tag{8.1}$$

where η is substrate efficiency, A_{comp} is the area consumed by components, and A_{sub} is the area consumed by the substrate. The metric could be extended to effective substrate efficiency, in which the areas of only usable components are computed. Substrate efficiency addresses one degree of how aggressively a particular MCM is filling a particular area with silicon or other components.

As a 3D metric, the substrate efficiency metric is largely meaningless. Attempts to apply it to 3D systems lead to impressive but somewhat nonsensical numbers that are greater than 1 in value, which is difficult to interpret. A 3D equivalent metric, the volumetric efficiency, is more sensible, but also difficult to apply since a standard is needed for comparison. One approach involves using the volume displaced by silicon fabricated onto a typical wafer (defined as 0.5 mm thick) as the unit reference:

$$\eta_{vol} = \frac{V_{comp}}{V_{si}} \tag{8.2}$$

where η_{vol} is volumetric efficiency, V_{comp} is the volume consumed by components plus the volumes inscribed by mounting overhead in cubic millimeters, and V_{si} is the surface area of the equivalent amount of silicon multiplied by 0.5 cubic mm. This metric reveals staggering inefficiencies in most packaging systems. Conventional PWBs have η_{vol} values of less than 1%,[1] and simply replacing small groups of components with MCMs will probably not result in η_{vol} values greater than 4–5%.

The present definition of volumetric efficiency is somewhat suspect, since silicon die are not always 0.5 mm thick, and it may be possible to distort the metric when thinned silicon is introduced.

Weight and power reduction metrics have not been reported in any consistent manner, and most improvement claims in these areas are based on heuristic arguments. Weight reduction seems to grossly track with volumetric efficiency. Power reduction on the other hand is generally slight (e.g., 8–15%), since most ICs are not designed to exploit the reduced capacitance and trace length of MCM substrates compared with any other wiring medium.

8.2.3.2 Performance Metrics

The only performance metrics reported pertain to attributes of packaging, such as wiring density and contact density. Wiring density is a product of the number of wiring channels in a layer of interconnects multiplied by the number of layers. The linear form is concerned with the number of channels per linear dimension (e.g., 30 lines/cm); whereas the areal form is based on the amount of interconnect in an areal dimension (e.g., 125 cm/cm^2). Contact density refers to the number of I/Os (input/output) between the boundary defined at one level of packaging (e.g., IC chip) and the next level (e.g., the package). Contact densities are defined normally by wiring pitch in the case of bond pads and packages and by the number of signals per inch in connectors. As in the case of weight and power metrics, no metrics have been previously described for performance improvements because of reduced electrical latency in MCM substrates.

8.2.4 Packaging Taxonomy

A taxonomy refers to an organizational scheme, and the many concepts in packaging can be confusing at first. The taxonomy summarized here deals with MCMs, packages, and 3D assemblies.

8.2.4.1 MCM Technologies

MCMs combine more than one component onto a unifying substrate or package that usually has some nontrivial interconnection within the substrate between components. A generic depiction of an MCM is given in Fig. 8.3. An MCM, besides components, consists of an interconnection substrate, a mechanical substrate, and a package. The package contains a body and I/O terminals for contacting to the MCM substrate on the inside, and the terminals extend to form leads on the exterior of the package to form the next-level interconnect. For the most general representation, the structures are distinguished. Sometimes the mechanical and interconnection substrates coincide, and sometimes one or both of these substrates coincide with the package structure.

A general taxonomy or organization system for MCMs must consider several material and structural configuration concepts that relate the MCM to both the components it contains and the next-level package. Figure 8.4 presents an attempt to capture the attributes that generally define a type of MCM. A type of MCM can be defined generally by starting at the top of the diagram and working down, choosing one of the several arrows at each level. In this diagram, we chose to consider first the interconnect arrangement relative to components in an MCM. Two possibilities exist: patterned substrate and patterned overlay. For patterned substrate approaches, components are mounted onto a substrate. We next consider the way components are integrated into the MCM. The means by which components are mounted to a substrate is referred to as the element-attach

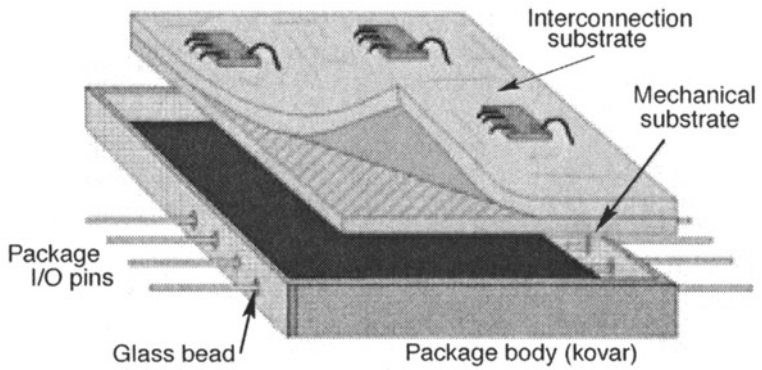


Fig. 8.3. Anatomy of MCM.

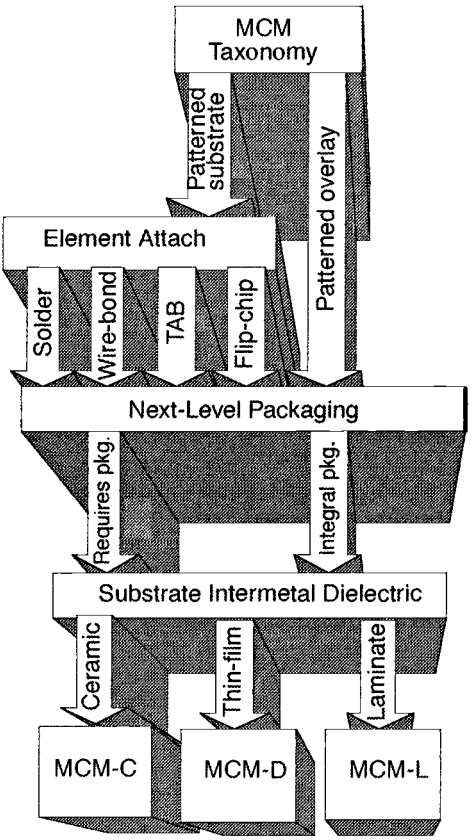


Fig. 8.4. MCM taxonomy (2D).

approach. Though a variety of approaches exist, only the dominant ones (solder, wire-bond, tape-automated bond, and flip-chip) are considered. For patterned overlay MCMs, the substrate is formed directly over components, metallurgically interconnecting to component bond pads, eliminating the need for another element-attach approach. Another differentiating characteristic of an MCM is its relation to an enclosure. There are two basic enclosure or package configuration cases. In the first case, a separate package structure with a body is required. In the second case, the substrate of the MCM *is* the package, which is also referred to as an integral package case. Finally, we consider the substrate material itself, more specifically the intermetal dielectric of an MCM, around which the common industry definitions of MCMs are based. There are three basic kinds of substrates, which are given the names MCM-C, MCM-D, and MCM-L.

It must be indicated here that the taxonomy is nonrestrictive. While Fig. 8.4 suggests that an MCM is categorized by selecting a particular path from top to bottom, it is in fact possible to have several coexisting possibilities within a single system. For example, if components are mounted onto the surface of a patterned overlay MCM, then both options of interconnect arrangement exist. It is common to have a single MCM with wire-bonded, flip-chip, and soldered assemblies on the same substrate. Figure 8.4 serves to illuminate the facets that define MCMs, but it is not an orthogonal classification scheme.

8.2.4.1.1 Interconnect Arrangement

The interconnect arrangement refers to an assembly ordering, but more importantly defines the relation between MCM components and the substrates (interconnection and mechanical) of an MCM. As an assembly ordering, the industry has coined terms such as "chips first" and "chips last" to designate patterned overlay and patterned substrate configurations. Figure 8.5 illustrates the two approaches. As shown in the patterned substrate case, the interconnection substrate and mechanical substrates are both "below" the component. In the patterned overlay case, however, the components are between the mechanical and interconnecting substrates. The patterned overlay is referred to as "chips first," since the components must usually be housed within a planar mechanical substrate before the interconnection substrate is formed. As it turns out, the mechanical substrate is not always required for either configuration, a case that is referred to as chip-on-flex. As remarked before, the patterned substrate and overlay approaches are not mutually exclusive, and their combination provides a powerful means for increasing density, for example, by permitting the addition of surface-mount sensors.

8.2.4.1.2 Element Attach

Since patterned overlay approaches form interconnections over components, the approach creates substrate and component attachment in the same process. For patterned substrate approaches, components must be electrically and mechanically attached to the interconnecting substrate medium. Even in the case of patterned overlay, surface component attachment is frequently important. A number of well-known techniques exist for surface attachment of bare components. The most common include solder, wire-bonding, tape automated bonding (TAB), and flip-chip or controlled collapse chip connection (C4), as it is sometimes called.

Soldering in the ordinary sense (flip-chip approaches are also solder-based) can provide electrical and mechanical attachment of components to a substrate, although as in the case of wire-bonding, mechanical attachment of the components through an adhesive (die attach) is often prescribed. This is desirable, especially in aerospace applications, to secure components independently of the solder mechanism.

Since its inception, wire-bonding has been by far the dominant form of electrically interconnecting ICs (the wire bonder was invented in 1960, preceding the invention of the monolithic IC in 1961[2]). Recent estimates indicate that 4 trillion wire bonds are formed annually worldwide.[3] In wire-bonding, gold or aluminum wires, spooled through the capillary of a wire-bonding apparatus, establish a flexible system in which a series of bridging conductors are formed between integrated circuit bond pad terminals and corresponding terminals on another component or substrate. Interconnections are formed ultrasonically (cold-welded at ambient temperatures) or thermosonically (at or below 150° C) to achieve a good metallurgical connection to bond padding

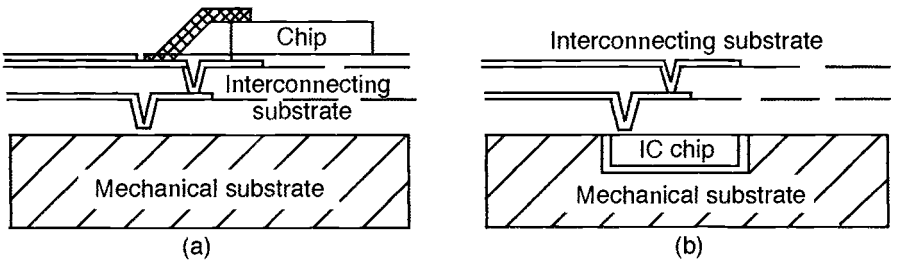


Fig. 8.5. Patterned-substrate (a) and patterned-overlay (b) substrate configurations.

surfaces. Wire-bonding continues to defy predictions of its demise. In the 1980s, predictions were made that wire-bonding would be replaced by TAB at I/O counts above 100.[4] Instead, wire-bonding has continued to find use at I/O counts above 500, and the pitch in wire-bonding machines has decreased from 150 μm to less than 80 μm (Fig. 8.6).

TAB technology involves the application of preformed frames to first the IC and then the substrate. They are observed to have the advantages of allowing pretesting, since the ICs can be transported and exercised after mounting onto TAB frames. Furthermore, TAB devices can be gang-bonded, which in theory allows a higher volume throughput to be achieved. However, the predictions of TAB sweeping the entire IC field as the choice element attach approach have not reached fruition. TAB's disadvantages include expense and lack of flexibility. Part of the expense associated with TAB is in the nonrecurring cost of TAB frames, which are unique to a particular IC bond-pad arrangement.

Flip-chip technology, established by IBM in the 1960s, involves forming a pattern of solder balls over the surface of an IC die, then inverting the die and aligning it with similar patterns on a substrate, followed by reflow. The process has a simple elegance in that electrical and mechanical attachments are performed simultaneously. Flip-chip has the ability (as do patterned overlay technologies) of placing bond pads over the entire die surface, which greatly increases the total number of contacts possible on a given IC. Flip-chip approaches continue to gain momentum, particularly with more complex ICs. One disadvantage cited for flip-chip technology is the need to introduce additional process steps to form a solder bump-compatible metallurgy over the IC's surface, which makes the technology difficult to apply to preexisting die, especially if they have already been removed from a wafer. Post-facto interconnection redistribution can result in suboptimal electrical performance. As such, flip-chip approaches are more easily mechanized in cases where new IC designs are involved.

8.2.4.1.3 Next-level Package

MCM substrates are sometimes considered incomplete assemblies and in those cases require mounting within a package. In some cases, however, the essential packaging structures are embodied directly into the substrate. Such constructions are referred to as integral packages. Usually, in the case of hermetic assemblies, a lid must be introduced, and sometimes a leadframe must also be soldered or attached in some other matter.

8.2.4.1.4 Material Classification

The industry usually classifies MCMs by only the type of substrate used. MCMs are then categorized as MCM-C, MCM-D, or MCM-L. More specifically, the types of dielectric used within the wiring layers of an MCM are actually addressed in this classification scheme.

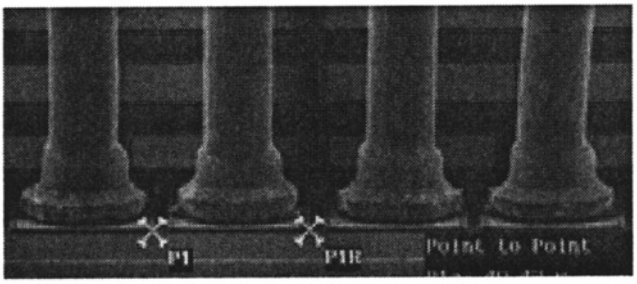


Fig. 8.6. State-of-the-art gold ball wire-bonding (50-μm pitch). (Courtesy ASM Pacific Assembly Products, Inc.)[5]

Ceramic-based MCMs, for example, are referred to as MCM-C technologies. Almost any aerospace MCM built before 1990, and almost any assembly bearing the name hybrid, is probably built with MCM-C technology. Several variations exist within the class of MCM-C, but most commonly, the designation describes a multilayer system based on alumina (Al_2O_3) or aluminum nitride (AlN), in which multiple metal-dielectric layers are cofired to form a substrate. Since these materials must be fired at very high temperatures (>700° C), they usually employ refractory metals for conductors and are referred to as high-temperature cofired ceramic technologies. More recent work in low-temperature cofired ceramics has enabled introducing metals with better electrical properties. Thick-film approaches, used in the original types of hybrid microcircuits (the "original" MCM), employ prefired substrates onto which a metal (e.g., gold) layer is formed.

The deposited thin-film or MCM-D approaches employ dielectric layers of polymer (e.g., polyimide, benzocyclobutene) and high-performance conductors usually of copper (with barrier metals such as titanium or nickel) that can be patterned with high resolution (e.g., 15–50-μm line widths). Also in the MCM-D class are approaches that employ a Si-SiO_2-aluminum like that used in VLSI (very-large-scale integration) technology. This system is capable of higher resolution, but the associated processes require modification for good electrical performance. For example, typical VLSI interconnections are built with thermally grown oxides only 0.5 μm thick; whereas much thicker oxides are required to achieve sufficiently low capacitance for high performance.

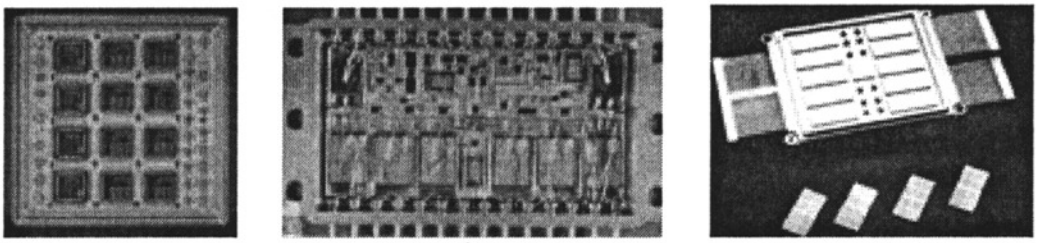
The third industry material class, MCM-L, employs laminate materials (e.g., FR-4, polyimide, and BT [bismaleimide triazine] resin), usually the same employed in the fabrication of PWBs. The most common name for MCM-L assemblies is chip-on-board (COB). The MCM-L is a grey area sometimes, since on the lower end of technology (line widths above 0.004 in/100 μm) it is possible to directly introduce bare die into ordinary PWBs. The higher end of the MCM-L technology employs line widths approaching 25 μm, which when combined with microvias, is very competitive in performance range and cost with other classes of MCM technologies.

The C-D-L scheme is not sufficiently rigorous to describe all the approaches in materials for MCMs. For example, some supercomputer designs employ a mixture of ceramic (MCM-C) and polyimide (MCM-D) in the same substrate. Consequently, a number of perfunctory attempts to establish hybrid designations, such as MCM-C/D, MCM-E/F,[6] and MCM-L/O, have been attempted. None of these schemes have received any real acceptance from the industry. In aerospace work, the MCM-D approaches have been most common, dating from the 1960s. General confusion and reluctance have slowed the acceptance of MCM-D and MCM-L approaches, but these material systems are making gradual inroads into a variety of space applications.

8.2.4.1.5 MCM Technology Examples

This section outlines a few planar MCM technology examples, not a complete cross-section of processes, but a representative sampling. Outside the normal taxonomy is a class of rapid-prototyping technologies for MCMs or MCM-like assemblies, which are briefly discussed.

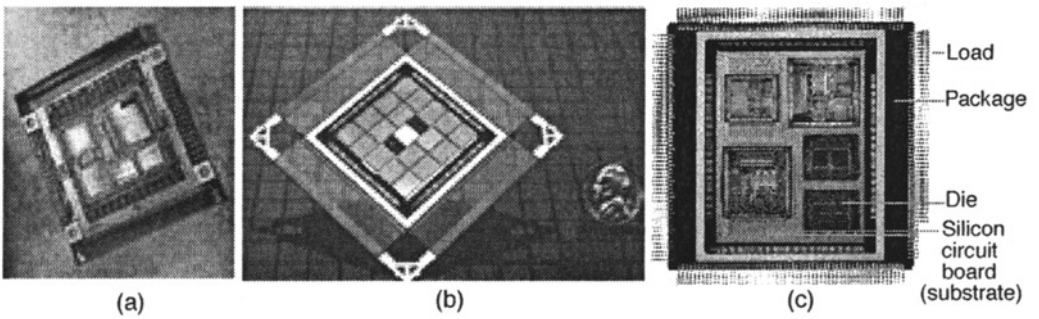
Patterned Substrate Technologies. Several representative ceramic-based (MCM-C) patterned substrate examples are shown in Fig. 8.7, each demonstrating a different form of the commonly available element attach technologies. TAB is shown to be useful for high-density connections in complex digital die [Fig. 8.7(a)], where multiple occurrences of several die exist, but for power hybrid assemblies [as shown in Fig. 8.7(b)], special forms of wire-bonding are required. The application of flip-chip technology to memory module [Fig. 8.7(c)] is straightforward, as demonstrated in a space-qualified application for a French satellite.[7] One of the more interesting "closet" examples of an MCM-C design is the Pentium Pro, which is based on a two-chip MCM (processor and cache).[8]



(a) (b) (c)

Fig. 8.7. MCM-C examples. (a) communications module, employing TAB bonding[9] (courtesy Harris Semiconductor, Harris Corp.), (b) power converter hybrid, using (in some cases very large) wire-bonds[10] (courtesy Teledyne Microelectronics), (c) French SATURNE memory module (IBM), employing flip-chip bonding (courtesy TechSearch International).

MCM-D examples are shown in Fig. 8.8. In the early 1990s, the most common form of MCM-D was a copper polyimide (Cu/PI) process, and suppliers of this technology were usually captive (e.g., Hughes Aircraft Co., Boeing, Control Data Corporation, IBM, Digital Equipment Corp. [DEC]). DEC's spin-off of MicroModule Systems established a merchant MCM-D capability, one of the few remaining commercial sources of Cu/PI based MCMs today. Its Gemini module is shown in Fig. 8.8(a), a complex processor core containing a complete Pentium-based computer.[11,12] Other examples of MCM-D technology exploit existing VLSI-like processing capabilities. Figure 8.8(b) illustrates a module built in the VHSIC (very high-speed IC) chip-on silicon (VCOS) technology, fabricated on the same line as the components that populate it. The advantages of this polyimide-aluminum form of MCM-D technology include a very high wiring density, comparable to that of the integrated circuit itself, and in this case, a vertical integration of a chip and MCM. However, the use of thin polyimides (about 3–4 μm) results in a lossy interconnect, restricting the performance range to lower clock frequencies. The example module shown, a memory module, was built at Lockheed-Martin (Manassas, VA) on the world's first qualified manufacturer's list certified MCM-D process. A slightly more progressive MCM-D technology is represented by nChip's Si-on-Si process [Fig. 8.8(c)], in which an aluminum-SiO_2 process, even closer to the native technologies of VLSI, is used to form sophisticated MCMs. Unlike traditional VLSI, which suffers from the limits of thermally grown oxides (about 0.5 μm thickness), nChip's special low-stress oxides allow for much thicker dielectrics (4–7 μm), permitting much greater electrical performance as a result of reduction in capacitance.



(a) (b) (c)

Fig. 8.8. MCM-D technology examples. (a) copper polyimide substrate, with wire-bonded parts[12] (courtesy MicroModule Systems), (b) aluminum polyimide process (lossy RC), with flip-chip attachment, (c) Si/SiO_2 substrate with aluminum conductors (courtesy Flextronics International Ltd.).[13]

Patterned Overlay Technologies. The most well-known patterned overlay MCM approach is the high-density interconnect (HDI) process, developed by General Electric under government and corporate funding. The standard HDI process is illustrated in Fig. 8.9. In addition, newer forms of the HDI process have been developed. The microwave HDI process,[14,15] for example, can accommodate air bridges and monolithic microwave integrated circuit components. Power HDI has more robust copper (thicker) structures for improved power delivery. Each of these variants of HDI is based on premilled substrates, usually ceramic. Newer versions of the HDI process include the plastic HDI and single-sided chip-on-flex (SSCOF) processes.[16] In these processes,

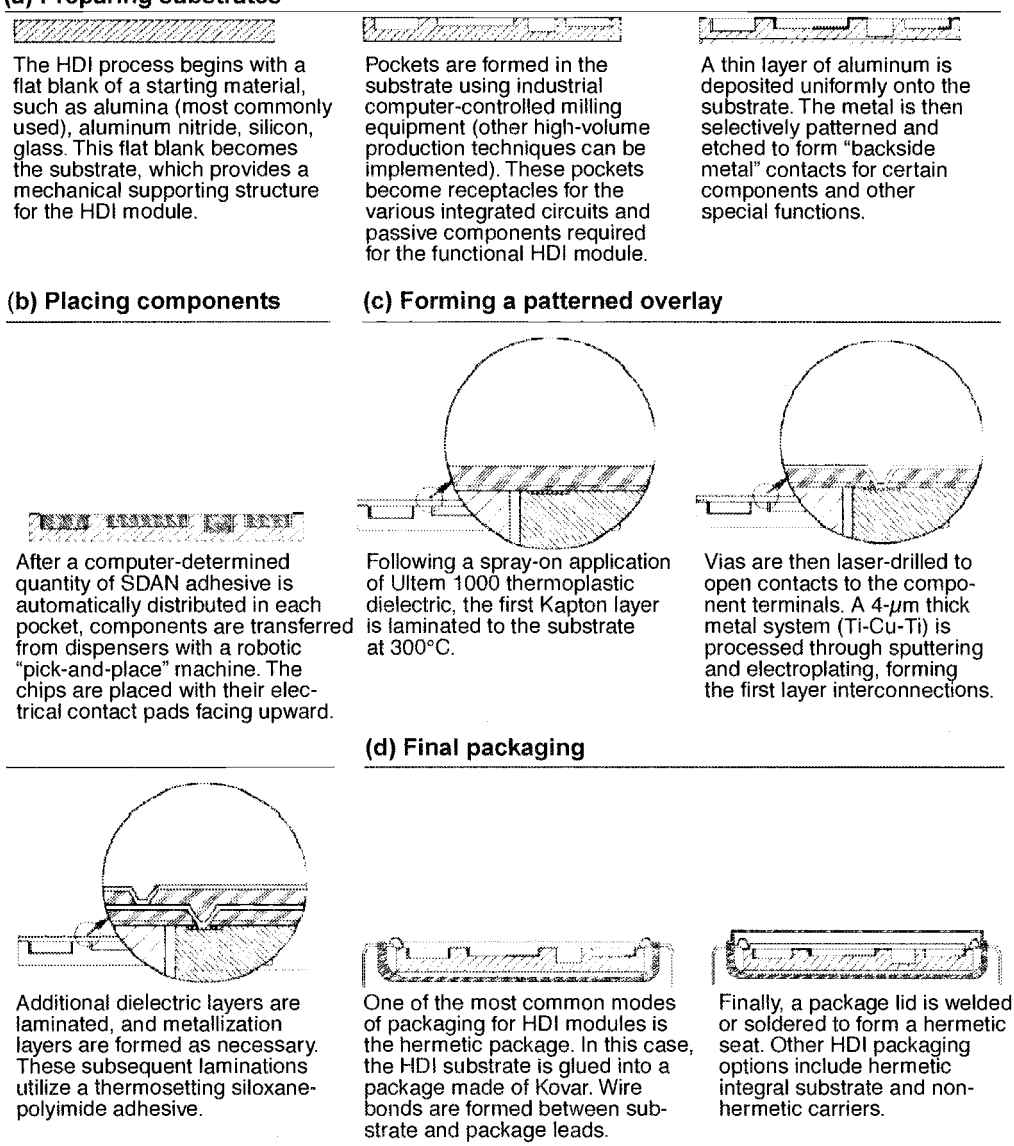


Fig. 8.9. Standard HDI process.[17]

components are placed upside down (bond pads facing down) onto a drumlike film of stretched Kapton, coated with adhesive. For plastic HDI, the ensemble of components is potted (for example, with Plaskon polymer), while for SSCOF no mechanical substrate is formed. The Kapton sheet is then inverted, and traditional HDI process steps [Fig. 8.9 (c)] are employed for creating a patterned overlay. In these processes, leadframes can be added to create integral packages for low-cost assemblies, or the entire assembly can be enclosed within a hermetic package. To achieve even further reductions in process cost, it is possible to use copper-clad Kapton "drums" that are prepatterned with interconnect. Such films are commonly available in high-density flexible circuitry work, and reduce the HDI-specific processing to via connections between components and flex-based conductors. One or more additional layers of HDI can be added to achieve higher wiring capability.

Rapid Prototype Technologies. Several options for "quick turn-around" and rapid prototyping technologies exist that promise to achieve faster fabrication cycles than those normally associated with MCMs (8–26 weeks). MCC, for example, offers a quick turn-around interconnect (QTAI) technology, based on a copper-polyimide process that can be formed within 2 weeks. The QTAI process involves a series of cuts and links formed on a nominally prepatterned substrate. Since most of the substrate interconnect is already formed, the QTAI process averts a lengthier fabrication cycle.[18,19]

Another rapid-prototyping technology is based on an antifuse substrate technology. Developed originally by Mosaic Computers[20] in the 1980s, this rapidly configurable substrate technology is employed in a variety of aerospace applications by Pico Technology, Inc. The substrates are formed *a priori* and are programmed as required to realize an application. These substrates are densely patterned with multiple interconnect layers and thousands of unprogrammed antifuses [Fig. 8.10(a)]. The programming process forms near short circuits at selected grid points, thereby creating connections between conductors in the multilayer system. Bond pads for signal, power, and ground are strategically positioned about the substrate such that almost any conceivable die placement arrangement will have access to one or more as required to connect components to the substrate. Such an approach can create complex MCMs rapidly, as shown in Fig. 8.10(b). A recent Air Force Research Laboratory (AFRL) contract with Pico Technology is exploring a system for portable MCM creation in which "blank" antifuse substrates can be programmed from a personal computer attached to a special apparatus. This software/hardware combination will establish the advent of field-programmable MCM substrates, giving organizations the capability of autonomously creating substrates as rapidly as designs can be formed.

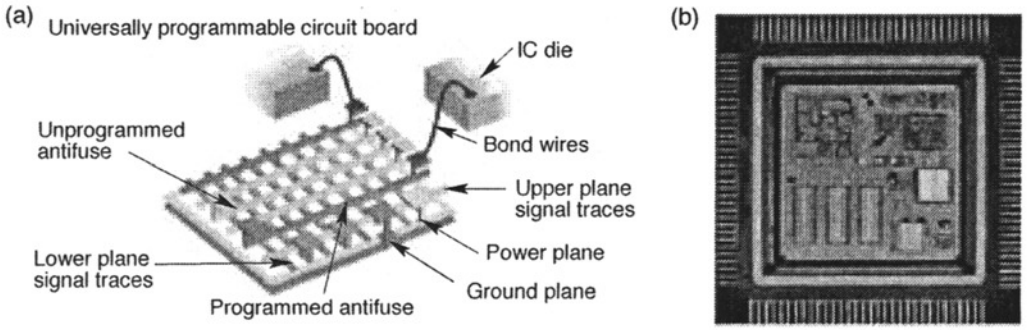


Fig. 8.10. Pico Technology's rapidly programmable substrate technology. (a) antifuse concept, (b) encoder MCM example (courtesy Pico Technology, Inc., Toledo, Ohio).[21]

8.2.4.2 Package Technologies

The package (L2 of the packaging hierarchy) contains one or more components. Packages, in a traditional sense, are the containers of ICs or MCMs that might be soldered onto a PWB . To most people, the package is the chip. Packages are limited in physical size, as they must be large enough to be handled by humans, but small enough to mount within other assemblies.

Simultaneous growth of signal count and increase in operating frequency of contemporary single-chip packages (SCP) and MCMs demand spatially efficient, adequate, electrically effective packages for test and operation. Many military still demand hermetic solutions. Despite the great variety of package formats, some designs still have unique requirements.

Packages are enabling or inhibiting, depending on one's perspective. Efficient packages provide an ideal protective environment for the components within, while providing a convenient handle and delivery mechanism for electrical and thermal performance. For MEMS devices, the package is a limiting factor. Accelerometers can be made smaller than bolts, but poor packaging creates an embodiment only marginally better than a non-MEMS alternative. Monolithic ICs usually require packages, but with some chip-scale packaging approaches, it may be possible to convert raw silicon into a form ready for assembly. Sometimes an MCM is its own package, but in other instances the MCM must also be protected with an external package.

Here we consider the types of packages used in electronics. The breakout of package usage by type is shown in Fig. 8.11. Packages can be broken into the dichotomies of through-hole/surface-mount, plastic/hermetic, and peripheral/areal.

8.2.4.2.1 Through-Hole Approaches

Through-hole approaches are traditionally represented by transistor "cans" [see Fig. 8.2 (b)] and dual in-line packages (DIP) (Fig. 8.12). Through-hole packages have existed as long as have PWBs. Their conductive leads penetrate a PWB, and the combination of mechanically crimping leads against the board and forming soldered connections provided a sort of "belt and suspenders" approach for securing electronics electrically and mechanically to a board. For ICs the dual in-line package has been a dominant form of package until the pin grid array (PGA) was developed for devices with higher I/O count.

Despite a significant loss in market share to surface-mount technology, through-hole technology remains in abundant use in systems. In 1995, 34% of all IC packages were through-hole, a fraction that is expected to decline to 10% by 2001.[22] It is unlikely that through-hole technology will completely disappear, however. For example, penetration of a circuit board is required for other reasons, such as providing access portals for sensors, actuators, heat sinks, and coolant flow.

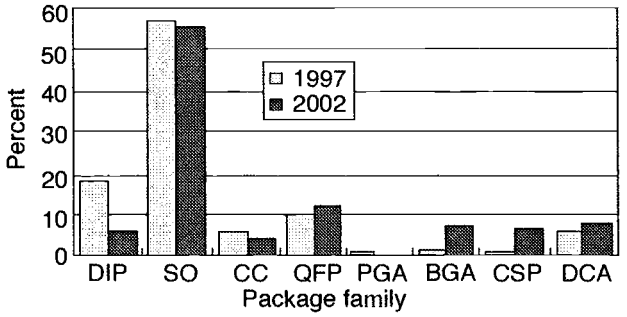


Fig. 8.11. Histogram of package usage.[23] DIP—dual in-line package, SO: small outline, CC: ceramic carrier, QFP—quad flatpack, PGA—pin grid array, BGA—ball grid array, CSP—chip scale package, DCA—direct chip attach (bare die).

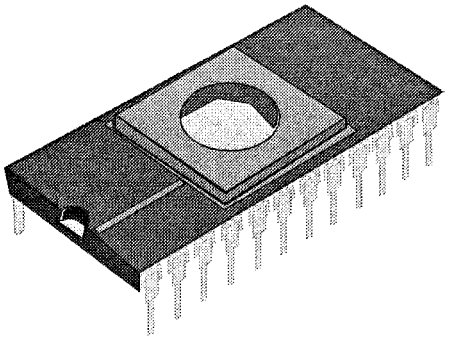


Fig. 8.12. Ceramic version of dual in-line package (CERDIP).

8.2.4.2.2 Surface-mount Approaches

Recently, the quantity of surface-mount (SMT) packages used in assemblies worldwide exceeded through-hole packages. A summary of SMT packages follows:

- Metric and thin quad flat package (MQFP and TQFP). QFPs are square, flat packages with a single row of leads emanating from each edge at pitch ranges from below 20 mils to 50 mils. They constitute the most commonly used package style today for lead counts above 20. MQFP sizes range from 10 mm to 40 mm square, and I/Os from 44 to 304. TQFP sizes range from 7 mm to 28 mm, with I/Os from 32 to 256. The size and lead configurations normally conform to JEDEC (Joint Electron Device Engineering Council) standards.[24]
- Thin small outline packages (TSOP). A very thin package, predominately for memory applications, which calls for backgrinding a normally 20-mil-thick silicon wafer to as thin as 7 mils.
- Thin very small outline package (TVSOP). More aggressive form of TSOP, with 16-mil lead pitch.[25]
- Very small package array (VSPA). An advanced package style developed by the Panda Project (Boca Raton, Florida) for high-density applications.[26]

8.2.4.2.3 Ball and Other Grid Arrays P

The ball grid array (BGA) package may be viewed as the single most important development in packaging this decade. Though BGAs represent a small percentage of the total package usage today[27] (Fig. 8.13), BGA technology is one of the fastest rising in the advanced packaging field. They have been applied to workstation and computer processors, complex gate arrays, field

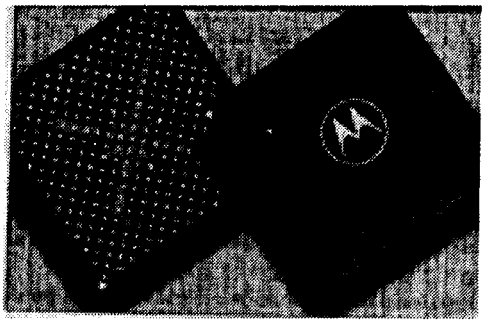


Fig. 8.13. BGA package.[28] (Courtesy ASAT Inc.)

programmable gate arrays, and consumer and automotive applications.[29] Few technology forecasters in the early 1990s viewed BGAs as having the dramatic potential they currently display.

BGA Advantages. BGAs can support higher numbers of I/Os much more efficiently than can perimeter-based packages.[30] In Fig. 8.14, the same I/O pattern is represented twice on the same IC, first as a pattern occupying three rows about the perimeter of the IC and second as an areal pattern resembling a stop sign. I/O capacity in the areal pattern can grow as N^2; whereas a perimeter layout approach has I/O growth at most as 4 N (where N is I/O spacing).

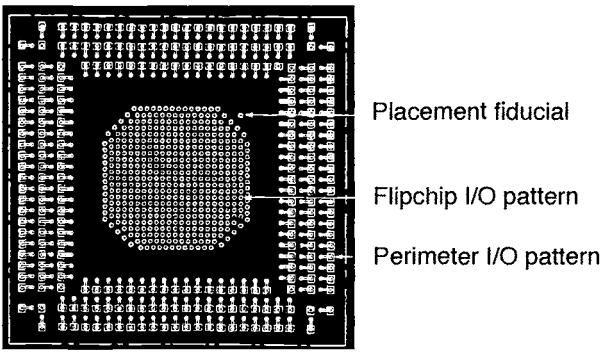


Fig. 8.14. Dual I/O distribution of the same pattern on the same IC (source: IBM).

BGAs have many other important advantages. As surface-mount approaches, they have improved routability over through-hole approaches of similar complexity, since pins do not penetrate the mounting board, thereby occluding valuable routing real estate. Their electrical performance is significantly better than other traditional package types.[31] BGA packages are smaller and lighter than other package types. They are less fragile than QFPs, which have delicate leads that if bent even slightly require special handling for assembly.

One of the most striking advantages of BGAs over other package styles is ease of assembly (notwithstanding the application of underfill) because of the self-alignment property of solder ball arrays. The surface tension of solder acts to bring the array of solder connections into alignment with the mating pattern of contacts on the board onto which a BGA is assembled. As shown in Fig. 8.15, misalignments (which can be as severe as half the ball diameter) can be corrected as the assembly is heated to the solder reflow temperature point. Though it had been a criticism of BGAs that they would not be permissible in assemblies because of their inability to inspect solder joints, it has been the experience of some assemblers that BGA failure rates are extremely low, in fact lower than that experience in QFP designs. X-ray inspection processes[32] were developed for BGAs, but in some cases manufacturers have dropped the x-ray inspection regimen after finding that with proper design, failures were so few as to warrant such procedures unnecessary. In the cases where rework is required (assuming no underfill), effective procedures have been developed for BGA repair using hot-air convective tools.[33]

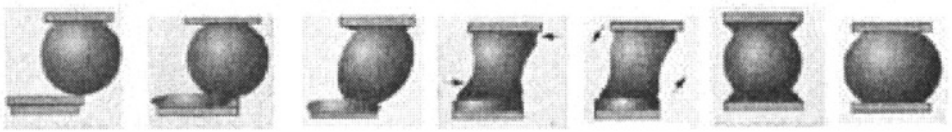


Fig. 8.15. Self-aligning property of solder ball arrays.

BGA Types. Several common BGA versions are shown in Fig. 8.16. By far the most common type of BGA in current use is the plastic ball grid array (PBGA) [Fig. 8.16(a)]. The PBGA involves forming a BGA package from circuit board material, usually BT resin. One side of this board contains a BGA I/O pattern, and the other side accommodates an IC direct chip attach, usually through wire-bonding. In the formation of the BGA, the component is mounted and encapsulated, and the solder balls are then attached in one of several ways, resulting in a finished assembly. Enhanced forms of PBGA, usually with improved thermal management features, have also been developed.[34,35] In one type, a combination of flip-chip attach and metal backing plate results in a BGA system with high thermal and electrical performance [Fig. 8.16 (b)].[36] Tape or flex BGAs employ thin-film interconnect (e.g., polyimide) instead of board materials (e.g., BT resin).[37] Ceramic versions of ball grid arrays (CBGA) have also been developed.[38]

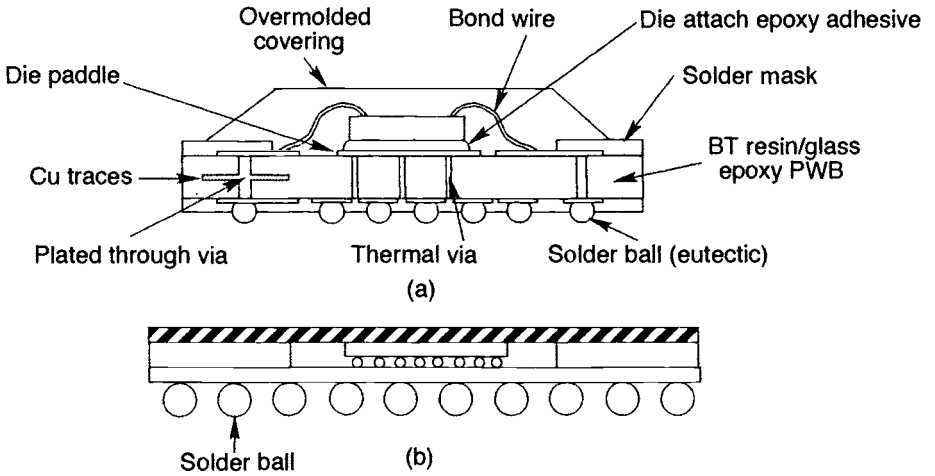


Fig. 8.16. BGA packages: (a) Plastic (PBGA), (b) Flip-chip based plastic with metal substrate.

BGA Issues. The most significant problem in BGA application is thermal expansion mismatch. In the case of PBGAs, the problem occurs within the package itself. In the region of the package where silicon is mounted, the expansion of the board material is constrained relative to its expansion on the silicon-free regions. The differential expansion sometimes creates cracking in solder balls in the boundary between the two regions. Overall package expansion can occur, particularly in metal and ceramic BGAs, in which cases the stress is an increasing function of size. Fractures often occur at solder ball interfaces.

Several concepts have emerged to deal with mismatch problems. In some cases, BGA designs avoid placing balls in constrained regions. To improve robustness of the solder ball/package interface on CBGAs, a "dimpled" BGA (DBGA) has been proposed, in which green state ceramic and paste are contoured as receptacles for each solder ball. To increase the range of compliance, a number of approaches have been employed to increase the height of the solder balls.[39] These approaches involve stacking several balls in each position or forming solder columns, and such assemblies are referred to as column grid arrays (CGA). A more common practice, however, involves the introduction of an underfill, a polymer injected underneath the mounted package, which surrounds the solder balls and creates a much more robust mechanical attachment. It is only with reluctance that assemblers use underfill, though, as most underfill compounds render the associated BGAs nonrepairable, and the underfill application process is time-consuming[40] and

problematic. Problems are not altogether confined to the solder; some of the more traditional plastic package reliability problems can also exist, such as popcorning.[41] Package coplanarity problems have also been reported.[42]

Other Grid Array Technologies. The other most commonly used grid array package is the PGA, which is usually a ceramic assembly in which pins protrude downward. PGA packages have reasonable electrical performance and thermal performance, but are bulky, require significant insertion force into assemblies (unless they are equipped with zero- or low-insertion force sockets), are difficult to rework, and create significant routing problems because of the significant amount of board real estate that is removed to make room for pins. The land grid array (LGA) is similar to a BGA without solder balls, and is used in some workstation cores.[43] LGAs require a compressive force and a compliant contact system, such as a socket, as there are no pins or solder bumps to secure them to an assembly.

8.2.4.2.4 Chip Scale Packaging

A chip scale package (CSP) is a package that is the same size or slightly larger than the component that it contains (up to about 1.5 × in area).[44] CSPs represent the state of the art in minimally packaged ICs, allowing board assemblies to approach the density of MCMs while preserving the ability to pretest components. Most CSPs employ a compliant contact redistribution system in a peripheral grid or a distributed array. Several general CSP types are shown in Fig. 8.17:

- Micro-BGA. Pioneered by Tessera (San Jose, California) [Fig. 8.17(a)], it relies on a polymeric pad formed over the die surface for compliance, and the BGA lead system wraps from the die bond pads on the perimeter to a surface formed over this pad.[45]
- Rigid-substrate mini-BGA [Fig. 8.17(b)]. Designed using a PWB substrate, with die attached by wire bond, TAB, or flip-chip.
- Lead-on-frame [Fig. 8.17(c)]. Relies on bond pads located away from the perimeter to achieve a smaller package area.
- Molded [Fig. 8.17(d)]. Miniature molded versions of standard packages.
- Wafer-level [Fig. 8.17(e)]. Connections formed on wafer through postprocessing (no separate package).
- Floating-pad [Fig. 8.17(f)]. Based on the HDI process, employs a novel floating pad structure for mechanical compliance.[46]

CSPs obviously employ denser I/O pitches than standard perimeter packages, which quickly become pad-limited at higher I/O counts. Standard BGA pitches are 1 mm and 1.2 mm; whereas CSP mini- and micro-BGA pitches are at 0.5 mm. The BGA forms of CSP resemble flip-chip mounted die, but strictly differ in the coarser size and pitch of the solder balls and the presence of a compliance system. Flip-chip devices, on the other hand, are typically underfilled as some BGA systems. CSPs have been shown to have good reliability without underfill.[47]

Do CSPs replace MCMs? For many of the less complex MCM designs, it may be possible to implement a CSP-based alternate design. Part of the limitation of CSPs lies in the substrate's ability to support dense wiring. MCMs represent a sort of limit argument in substrate density, and in fact, the CSP creates a blurring between the boundaries of MCM and traditional packaging. The MCM, besides providing for a more compact design, offers "relief" to the wiring requirements of a substrate, permitting simpler boards in an overall system design, while the CSP, in some cases, forces more complexity at the board and/or next level of packaging. Application constraints provide insight to the proper use of CSPs and MCMs. An obvious guideline is to examine the tradeoffs in cost and complexity as the wiring "burden" is shifted from the board level to the MCM level while keeping an eye on the associated impacts on size of each prospective implementation.

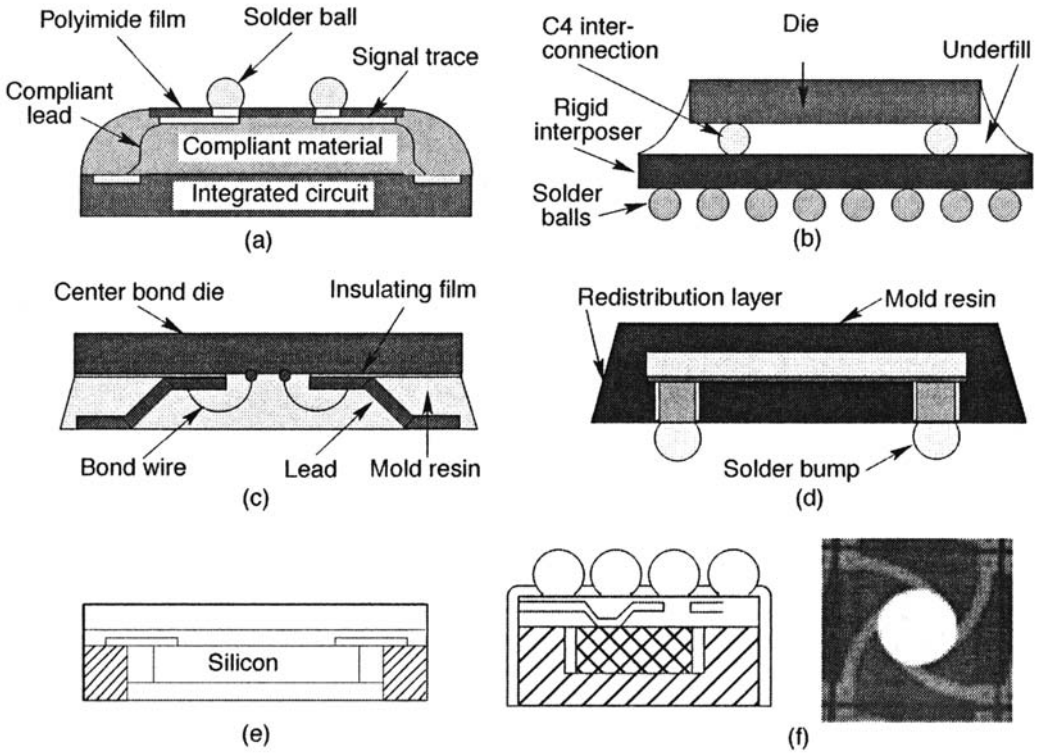


Fig. 8.17. Various CSP types. (a) interposer with flex circuit,[48] (b) interposer with rigid package,[48] (c) lead-on-chip,[48] (d) miniature mold,[48] (e) wafer-level package,[49] (f) HDI-based, with close-up of floating-pad structure used in pad array.

8.2.4.3 3D Packaging

Three-dimensional packaging is suggestive of approaches that can overcome the barriers of the next-level packaging problem. The approaches are almost always formed from a collection of ICs, MCMs, or packages that are stacked in a very compact configuration. The desired characteristics of a 3D approach include:

- High volumetric efficiency
- High-capacity layer-to-layer signal transport
- Genericness (the ability to accommodate random circuitry and the widest possible variations in layer types)
- Heterogeneity (the ability to accommodate multiple "modalities," e.g., analog, digital, power, RF)
- Serviceability (the ability to isolate and access a fundamental stackable element for repair, maintenance, or upgrade)
- Ease of interface and integration to a variety of existing and emerging next-level packaging technologies
- Adequate heat removal to the next level of packaging
- High pincount delivery to the next level of packaging with high electrical efficiency
- Adequate structural support
- A practical scheme for extending to an arbitrary number of layers

Establishing a categorization system or taxonomy for 3D packaging is difficult, given the diversity of packaging concepts involving the stacking of die, MCMs, and packages. A taxonomy is proposed in Fig. 8.18. We consider first that two regimes exist relative to layer count. In few-layer 3D, the assemblies are generally mounted onto a next-level assembly flat, like a deck of playing cards. In many-layer 3D, the number of "playing cards" is high enough to topple over. In packaging, such stacks are edge-mounted instead. The distinction is important for thermal management purposes. The term "layer fixity" simply refers to the ease of disassembly. The possibilities of nonseparable and demountable are fairly self-explanatory. The element interconnect scheme describes the mode through which the primary stacked elements "communicate." We identify three possibilities: plane-edge (similar to backplanes), plane-plane (such as mounting a BGA to a circuit board), and edge-edge (a twist literally on the edge-edge scheme). There are

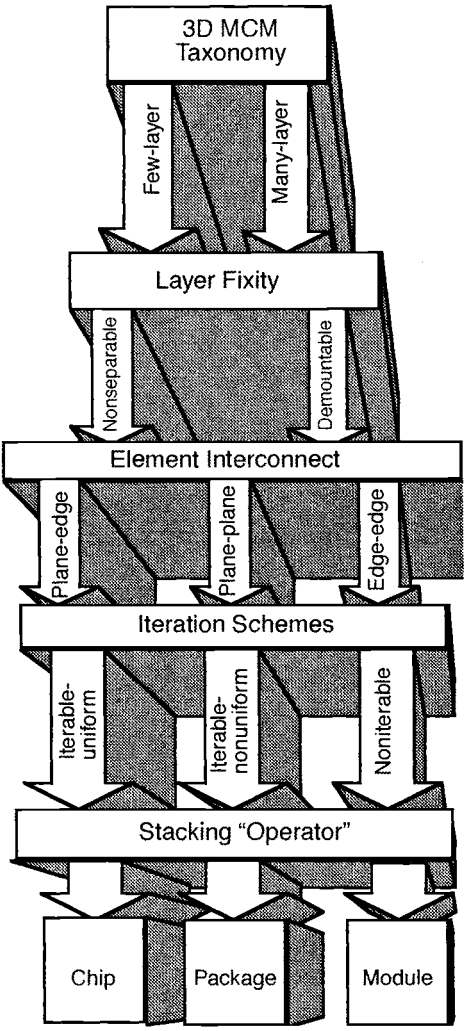


Fig. 8.18. 3D-MCM taxonomy.

three general conceptual possibilities that describe how stacking occurs in an assembly. The iterable-uniform case is like stacked playing cards, namely identically sized assemblies stacked in an identical manner. The iterable-nonuniform case covers the possibilities of telescopically stacked or stagger-stacked assemblies. In the former case, the assemblies stack directly upon each other, but each assembly is progressively smaller. In the latter case, identically sized layers are staggered as they are stacked. The noniterable designation is the "catch all" default, covering other approaches, such as folded MCMs, double-sided MCMs, and compound MCMs (MCMs inside other MCMs). Finally, the stacking "operator" is considered, that is, which type of element is being stacked (IC die, MCMs, or packages).

The remainder of this section discusses examples of chip, package, and module stacks to illustrate some of the interesting possibilities that have been attempted in the construction of microminiature systems.

8.2.4.3.1 Chip Stacking

Chip stacking appears to be the dominant form of 3D packaging, most commonly applied to increase the density of memory components in a system. Most chip-stacking approaches would be classified as few-layer, iterated, nonseparable, plane-edge approaches. Examples of memory stacking approaches are shown in Fig. 8.19. These approaches are popular as "baby-steps" in 3D packaging, as they allow stacked components to replace normally planar components in a design. For memories in particular, increased density is a fairly simple architectural modification in most cases. The first approach, the short-form memory stack technology developed by Irvine Sensors Corp. under USAF funding ([Fig. 8.19(a)], employs a few-layer concept employing wafer-level processing, derived from the many-layer "sugar-cube" approach [Fig. 8.19(c)]. The StackTek concept (Fig. 8.19(b)] applies a special frame to each die and stacks them, bussing all interconnects at the side edges. Many other examples of short-form stacks are found in the literature.[50]

8.2.4.3.2 Module Stacking

Module stacking approaches involve stacking elements or modules that contain one or more components. Unlike die stacking approaches, the components of each layer need not be identical in function, shape, and size, but the module boundaries usually have identical shape and size. A summary of some of the more eclectic approaches is provided in Table 8.1. Illustrations of some of these approaches are provided in Fig. 8.20.

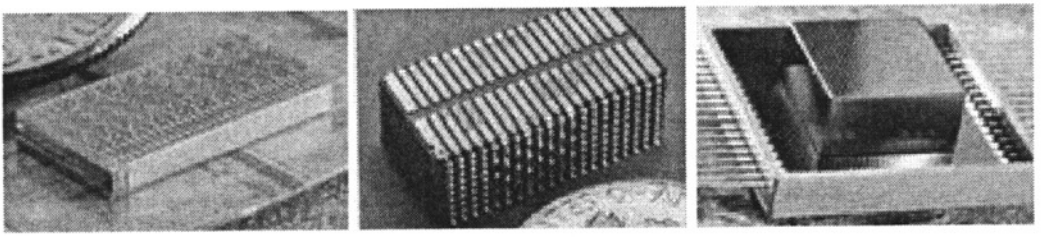


Fig. 8.19. Memory chip-stacking approaches. (a) short-form die stack based on wafer-level processing (courtesy Irvine Sensors),[51] (b) short-form leadframe stack (courtesy Staktek Corp.),[52] (c) many-layer version of (a) (courtesy Irvine Sensors).[53]

Table 8.1. Summary of 3D-MCM Stacking Approaches.

Technology	Organization	Domain	Fixity	Interface	Iterability
3D high-density interconnect	GE	both	nonseparable	edge-plane	iterable
HDI-folded flex	GE	both	nonseparable	various	noniterable
MCM-V	3D-Plus (France)	many-layer	nonseparable	edge-plane	iterable
Double-sided ceramic	Honeywell	few-layer	nonseparable	plane-plane	noniterable
Mezzanine-stacked substrates	ATT, Space Computer Corp.	few-layer	separable	plane-plane	iterable
3D silicon MCMs	LETI (France)	few-layer	nonseparable	plane-plane	iterable nonuniform
Mosaically arranged data compression and processing (MADCAP)	Boeing (formerly Rockwell)	many-layer	nonseparable	edge-plane	iterable

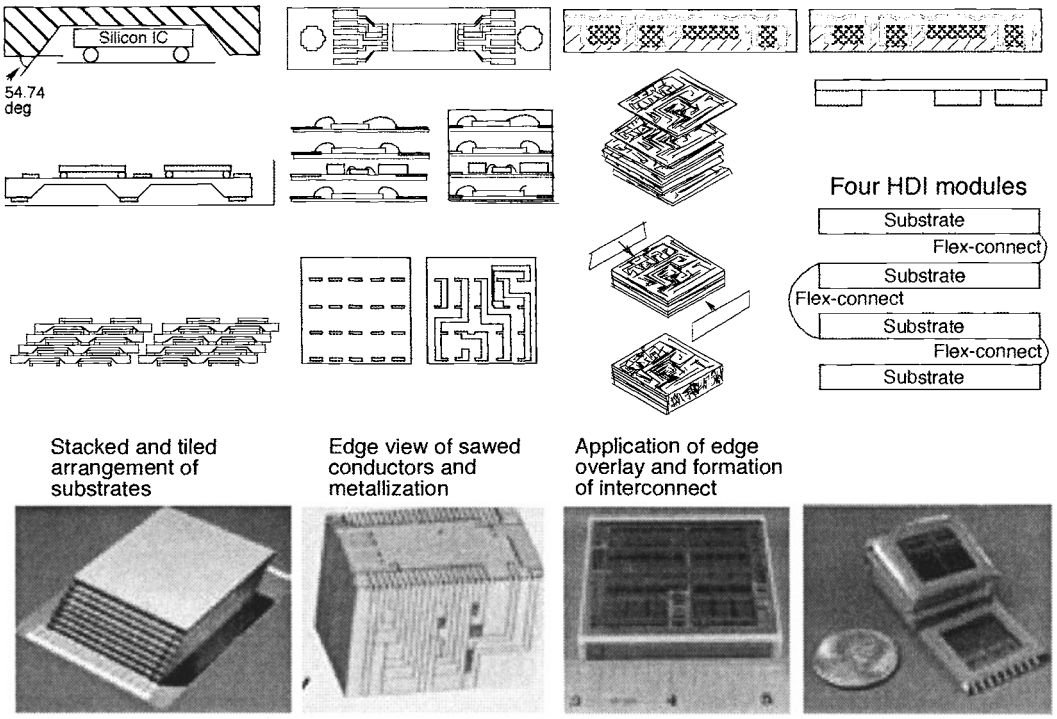


Fig. 8.20. Representative 3D module approaches. (a) 3D Silicon MCM (photo courtesy *Laboratoire d'electronique de technologie et d'instrumentation*—LETI, Grenoble, France; photo by Artechnique), (b) MCM-V (photo courtesy 3D Plus Electronics, France), (c) 3-D HDI four-layer stack, (d) folded HDI MCM.

8.2.4.3.3 Package Stacking

Package-stacking approaches offer a simpler, low-technology method of achieving higher density than possible with 2D approaches. The most common implementation involves stacked-memory components. One package-stacking supplier, Dense Pac Microsystems, Inc., employs a variety of concepts, one of which is shown in Fig. 8.21, to achieve a compact stack that is not as dense as the approaches shown in Fig. 8.19, but can be made at a lower cost.

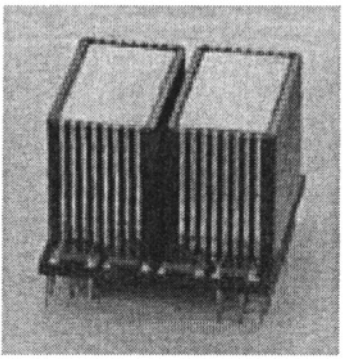


Fig. 8.21. Dense-Pac Microsystems stacked-package memory system.[54] (Courtesy Dense-Pac Microsystems.)

8.2.4.3.4 Boards, Boxes, and Cables

The ability to mount electrical components onto a planar structure and interconnect their terminals without a tangled nest of point-to-point wiring connections was made possible with the advent of printed circuit boards. Board technologies can be crudely divided into: (1) single-sided (copper clad), (2) double-sided (copper clad), (3) multilayer with drilled vias, and (4) microvia technologies. Boards are most commonly thought of as rigid, planar structures onto which components can be mechanically affixed and electrically interconnected. Another important class of printed wiring technologies, referred to as flexible circuit technologies, is capable of conformal application.

Often an aerospace platform's electrical system is partitioned into units, and boxes contain the electronics of each unit. The boxes are serviceable enclosures containing a collection of one or more circuit boards, which are secured by screws, wedge-locks, and connectors. Significant wasted space usually results, as board-to-board pitches must provide sufficient clearance for all components. Board-to-board connections are formed using two primary methods. The first method involves a backplane (edge-plane) interconnection approach. The second method involves a mezzanine (plane-plane) interconnection approach. Sometimes boxes can embody essential performance enablers, such as the resonant cavities in microwave designs. As always in the hierarchy of packaging, the box must address all necessary roles of packaging, including thermal management, power delivery, signal transport, optoelectronic, fluidic, and other couplings. Except for the structure and thermal management, these functions are usually provided through connectors.

It has been remarked that connectors should actually be called "disconnectors," since that is what they permit. Connectors are born of the need to rapidly plug or assemble systems, and they are especially important when those systems have to be serviced. Connectors define a $23 billion[55] industry, which attests to their pervasiveness in any electronics assembly. Cables usually

bridge connectors between enclosures and provide interconnections between noncollocated elements at different levels of an assembly. In space systems, connectors and cables consume a surprising amount of bulk and mass. Connectors are used most often in one of the following ways.[56]

- Board to board
- Component to component
- Board to cable
- Cable to cable

Connectors and cables are generally thought of as routing electrical power and signals, but they can also route photonics and fluids.

8.3 Engineering Considerations for Packaging

Packaging is the "wrapper and mapper" of systems to their components at various levels of regard. For example, the package of a Pentium or PowerPC microprocessor is a protective wrapper of the delicate IC within the package. The wrapper should be as small as possible, since structure adds size and hinders performance. Conflicting with the goals of miniaturization are the mapping requirements, such as power distribution, heat removal, and electrical access to terminals. The package provides a heat sink to keep the IC operating within reasonable temperature bounds and maps that heat sink through the package structure to the IC. Power and signal delivery are mapped through the pins or the I/O of the package.

The goal of good packaging design is to make packaging "invisible" to the circuits that are affected and to the system that uses the associated assemblies. A good packaging system adds the minimum amount of size and weight to a collection of components and enables essentially the full performance potential of components to be delivered to the system. This section addresses the balance of considerations one must deal with to engineer assemblies of monolithic components.

8.3.1 Engineering to Constraints

The development of packaging involves the coordinated consideration of many constraints in materials, geometries, and architectures.

8.3.1.1 The Integrated Circuit Die

The size of an integrated circuit die is determined by yield, the amount and complexity of circuitry within the die, and the number of I/O terminals (bond pads) and how they are accessed.

8.3.1.1.1 Yielded Size

A rough estimate for the number of yieldable die is given (modified from Moresco[57]) as

$$N_D = \pi\left(\frac{(D-a)^2}{4(X+d_{saw})}\right) \tag{8.3}$$

where N_D is the number of die, D is the wafer diameter, a is the scrap border region at the perimeter of the wafer, X is the length dimension of a square die, and d_{saw} is the dimension of the saw street. The actual size of the die is determined by the quantity of functionality required, the density of which is obviously increased as feature sizes of the semiconductor technology are decreased. In the system-on-a-chip philosophy, where vast amounts of functionality are required, the size of the die is limited by the inherent yield of the semiconductor process. For processes with good yields (>30%), Murphy's model applies:

$$Y = \left(\frac{1 - e^{-AD}}{AD}\right)^2 \tag{8.4}$$

where Y is yield, A is chip area, and D is defect density ("killer defects"/cm^2).[58] From a yield standpoint, smaller die are obviously better, but packaging bottlenecks and additional assembly complications seem to drive a high degree of "monolithicness" whenever possible.

Gate-Limited vs Interconnect-Limited ICs. Two general cases exist for resource utilization within a complex digital microcircuit. The first case, interconnect density-limited, involves circuit elements (digital gates) built at densities far greater than the interconnect wiring can support. In the second case, gate density-limited, the IC density is limited by the size of those circuit elements. For complex digital processes, interconnect density limitations result in wasted silicon, and the addition of more wiring layers in silicon has been a popular trend in the 1990s to improve utilization. Of course, adding metal layers has its own form of penalty, usually in the form of yield loss because of severe topographical warpage of interconnections. Technologies such as chemical mechanical polishing (CMP),[59] which serve to minimize the accumulation of topographic misalignments, are now common practice in many processes.

8.3.1.1.2 I/O Terminal Count and Placement

The number and arrangement of I/Os have the most profound impact on the complexity of the next level of packaging. The vast majority of components have fewer than 200 I/Os, as shown in the chart in Fig. 8.22, based on 1996 market research.[22,60] Even though high-pincount components only constitute 1% of the total market, 500 million units still represent a significant quantity (48 billion packages were assembled in 1995[60]).

I/O Count. One of the most important models for the terminal density of electronics systems in general comes from an empirical relationship known as Rent's rule. The relationship is named after E.F. Rent, who developed a functional description for the number of pins required by a digital design based upon the number of logic gates required.[61,62] Some packaging experts have stated that Rent's rule can with reasonable accuracy predict the unconstrained terminal count of assemblies such as super-computers.[63] Furthermore, Rent's rule has been used to analyze the architecture of field-programmable gate arrays (FPGA). It has been suggested that it is possible to predict which applications optimally map to FPGAs by fitting parameters of Rent's rule to FPGAs based on their internal mix of logic and interconnect routing resources. As such, it is possible to predict which types of functions are poorly mapped onto particular FPGAs, or conversely how to establish more optimal FPGAs for particular applications.[64]

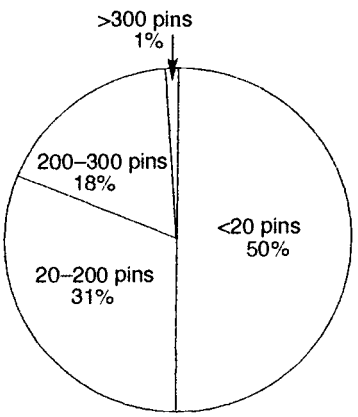


Fig. 8.22. Breakdown of packages by pincount (%).

The Rent's rule relationship is of the form

$$I = bC^p \tag{8.5}$$

where I is the number of I/Os, b is the average number of connections per circuit, C is the number of circuits (gates), and p is a positive component.[65] Rent's rule is domain-specific; that is, the coefficients b and p differ based on the class of function. The pin count of simple, regular structures, such as memories, follow a relatively slow growth compared to that manifested by high-performance computer systems and random logic (as commonly implemented in gate arrays). Moresco[57] uses $b = 3.2$ and $p = 0.434$ for Rent's parameters in a group of logic cells, while Bakoglu[61] uses $b = 1.9$ and $p = 0.5$. Bakoglu also suggests that the Rent's parameters for static memory are $b = 5$, $p = 0.12$, and for microprocessors, $b = 0.82$, $p = 0.45$. I/O growth for several different circuit types is shown in Fig. 8.23.

Rent's rule, however, requires careful interpretation. As it is heuristically accepted that the terminal count in an electronics system is generally a growing function of the number of subelements contained, at the highest level of a system the terminal count decreases. A personal computer, for example, which contains many millions of gates, exhibits a very low terminal count when viewed at a system level. This "necking down" phenomena is sometimes referred to as "breaking Rent's rule," a strong motivation for systems on a chip. However, the same personal computer system, when randomly bisected (say by chopping it with an axe) exhibits many thousands of "terminals" represented by the newly cleaved conductors. Rent's rule it seems may depend on reasonable or optimal partitions of systems and subsystems, and finding these optimal system partitions is computationally complex.[66] Systems that are not partitioned in concert with Rent's projections may suffer in performance. Even systems with optimal partitioning may be bandwidth-limited if the serviceable terminal count cannot meet the Rent's projection for that particular system class. This phenomena is observed in some multiprocessing systems, especially those that allow restrictions in internode terminal count due to packaging technology limitations. Rent's rule also only provides a projection of signal terminal count. The total system terminal count must include the terminals dedicated to power distribution, which can approach a significant fraction of the total signal count.

I/O Placement. Whatever the actual terminal count of a system becomes, the number of terminals can create a packaging problem. Since most connections are done through wire-bonding,

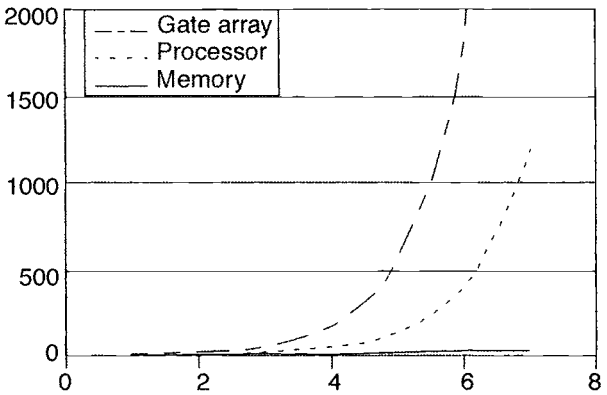


Fig. 8.23. I/O predictions based on Rent's rule for different types of circuits as a function of the number of circuits (gates).

perimeter location of bond pads are often required in practical configurations. Given the current practical limits of wire-bonding (60–75μm),[2] it is quite possible for high I/O devices to waste lots of silicon simply from the need to provide the minimum fanout requirements for bonding. In such cases, the chip is referred to as pad-limited, meaning that were it not for perimeter bond-out requirements, the chip could be made smaller. In these cases, switching to an areal I/O distribution can be advantageous. In a recent research project, dramatic die size reductions (75%) were achieved by utilizing HDI technology to access bond pads distributed about the surface of a die, rather than construct a much larger, pad-limited die (Fig. 8.24).

8.3.1.2 The MCM Substrate

For MCM designs, the substrate is a critical integrating medium of a number of components. The components are usually assumed monolithic, but this restriction is not essential. Provisions in an MCM assembly must be made consistent with the potentially wide variety of components it may contain. Other considerations must enter at this level because of the larger physical scale of the assembly.

8.3.1.2.1 Size of an MCM Substrate

The factors affecting MCM size are similar in principle to those in the monolithic case: required functionality and yield. Additional consideration is given to MCM size and aspect because of carrier utilization and structural considerations.

Yield. In the case of an MCM, a larger number of yield mechanisms must be accounted for, and one expression for the total yield Y presuming independent mechanisms is given by:

$$Y = Y_{comp} Y_{int} Y_{assy} \tag{8.6}$$

where y_{comp} is yield due to components, y_{int} is yield due to interconnect, and y_{assy} is yield due to assembly. Of these, the component-based yield is by far the dominant loss mechanism, and has led to what is commonly referred to as the "known good die" problem (see Subsec. 8.3.4.1.1). Interconnect yield loss, process-related, is usually of more concern in new processes and is manifested in conductor trace and via opens, shorts, and leakage paths, along with other process-related problems of all varieties. These are usually more rare in well-established processes, where facility cleanliness, stable process parameter windows, and design rules work in concert to minimize the probability of such defects. Assembly yield mechanisms refer to the variety of problems that occur as components are united to the substrate, substrates are shipped, and to some degree,

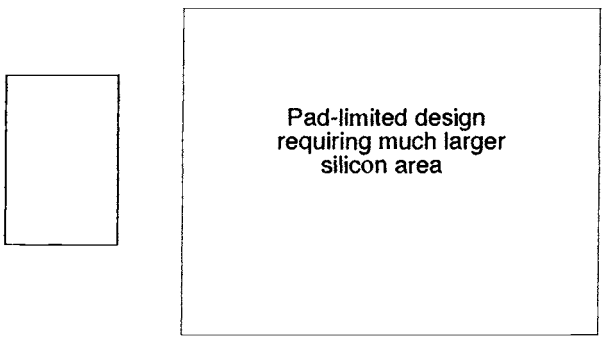


Fig. 8.24. Example of die size economies in basic instrument controller gate array when built as an "HDI-ruled" pad array vs a perimeter design.

as they are integrated into a higher-level assembly. Process controls are often more difficult to exercise in assembly as material surface and component preparation varies across a wide base of manufacturers. MCM developers attempt to combat these problems creatively by specifying careful component inspection procedures (often with more stringent acceptance criteria than specified in military standards) and sometimes adding extra component treatments to improve adhesion characteristics.

The size of substrates affects yield in an intuitively obvious way. Larger substrates can contain more components, which aggravates component and assembly yield loss, while interconnect yield loss degrades more or less continuously with increased size. A simple case of a module containing devices of area a and yield y reveals that total yield can be specified as:

$$Y = y^{(l/a)^2 \eta} \tag{8.7}$$

where l is length (square substrate is assumed here) and η is substrate efficiency. Since y is by definition less than 1, total module yield gets worse with increasing size. *The sum of these arguments suggests that smaller MCMs fare better overall.*[67]

MCM carrier utilization is an often-neglected motivation for carefully controlling MCM size and aspect, especially for some of the MCM-D processes that utilize carriers below 6 in. diam. Carriers for MCMs are analogous to wafers in ICs, and a grouping of carriers form a lot. The difference can be dramatic, as shown in the simple example in Fig. 8.25. In Fig. 8.25(a), a 6-in. carrier (5.5-in. usable diam) is shown with a single 2-× 4-in.-MCM substrate. Obviously, only one MCM can be arranged onto a complete carrier. In Fig. 8.25(b), 17 small 1-in.-sq MCMs can be arranged onto the same carrier, improving carrier utilization from 34% to 72%. If the two designs represented, for example, memory modules, the Fig. 8.25(b) design would often be more cost-effective, since the yield of each smaller substrate would be higher and amortized lot-processing costs lower than the Fig. 8.25(a) design.

As a final consideration to the selection of MCM size, and certainly not the least important, is the structural implications of MCM size selection. MCMs are typically much larger than any individual IC component; large physical dimensions can complicate the design of a higher-level package, board, or direct system mounting structure. Most MCMs are planar and not intended to experience flexure in their operation. Larger MCMs require additional stiffening structures, which in turn add size and weight to the system embodiment. Any material incompatibilities in next-level packaging can further exacerbate mounting because of mismatches in thermal expansion coefficients. The additional complications serve to decrease reliability, and the advantages of MCM implementation become questionable.

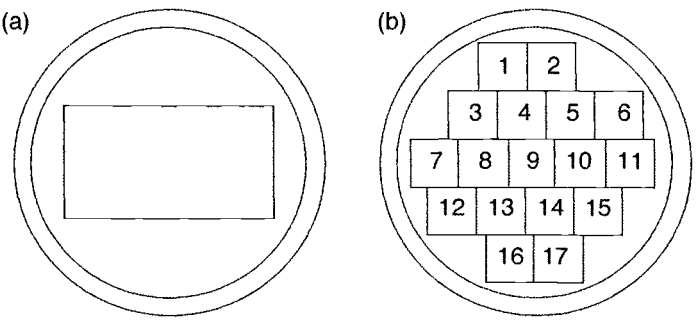


Fig. 8.25. Carrier utilization (6-in. diam) in MCMs. (a) 2 × 4 in., (b) 1 in. sq.

8.3.1.2.2 MCM Component Selection and Compatibility

One of the greatest strengths in an MCM embodiment of electronics is that a number of choice components from individually optimized fabrication processes can be united onto common substrates. For this reason, it is often better to use MCMs to implement designs involving MEMS, photonics, and other "mixed-mode" designs (designs that involve mixture of power, analog, digital, and microwave components). In many cases, the resulting MCM will achieve performance comparable to that of a truly monolithic implementation, while offering a substantial flexibility not possible in any monolithic fabrication process. Yet to access the features of a particular component in an MCM, for example to release gears and cantilevers in a MEMS device, requires careful consideration of the MCM substrate fabrication and assembly processes. In most cases, merging diverse components will have impacts far subtler than would be experienced in, for example, the monolithic integration of a MEMs device with a fiber-optic transceiver and digital processor. The most notable examples in current research are the merging of simple MEMS devices within an MCM, where the impacts of release chemistry must be handled without damaging other components within the MCM.[68] Changes in processes that are needed to accommodate traditional nonphotonic silicon and GaAs devices are referred to as intrinsic accommodations. Dealing with features beyond those needed in these traditional devices is referred to as extrinsic accommodations. Intrinsic accommodations include material compatibility (e.g., thermal expansion mismatch) or environmentally influenced (e.g., extremely high G-force) factors. Extrinsic accommodations cover special cases imposed by application. Examples include:

- Alignment of a MEMS accelerometer-gyro mounted onto a substrate
- Single-mode optical fibers, which have in most cases submicron alignment tolerances
- Provision of optical windows or environment access portals, a possible requirement at a substrate level
- The need to "map" aspects of the outside world into the package

Some of the newer self-assembly techniques being explored for MEMS devices offer potential solutions in some cases to alignment problems,[69] and micromachining techniques allow for the integration of fluidic channels[70] for improved thermal management.

8.3.1.2.3 Wiring and Contact Density Implications

Wiring in any assembly, whether an IC, MCM substrate, PWB, or harness, is a "supply and demand" game. Point-to-point connections must be made within and between subassemblies at a particular level in the packaging hierarchy, and substrate technologies must match closely the ability to deliver an appropriate level of interconnection. Too little wiring capacity creates intractability, while too much wiring capacity is costly and a poor design trade. In the former three cases, strict limitations are imposed on planar multilevel wiring media, while in the latter case, point-to-point wiring is channeled through cables or individual jumper wires within an assembly.

Planar multilevel wiring is made possible by the concept of a via, which is simply a connection between wires formed on one plane to another. Vias in VLSI and MCM processing are normally formed from one level of metal to the next level only. Some MCM processes, such as HDI, can form "spanning" vias that cross several levels of interconnect (say from level 3 to level 1). These concepts are illustrated in Fig. 8.26. These vias are sometimes called staircase vias, which cannot be stacked directly. Forming a connection from level 1 to level 4, for example, requires a staggering of the vias from one level to another, creating the staircase effect. More aggressive VLSI processes have begun to employ filled vias, which can be stacked upon each other, providing a

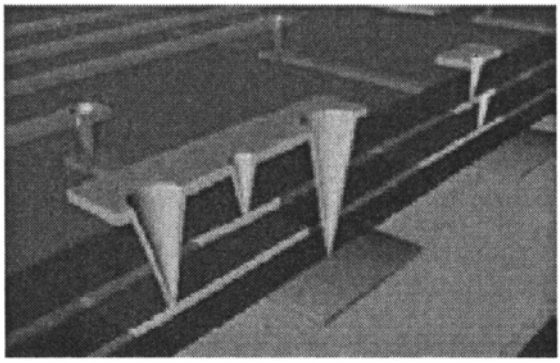


Fig. 8.26. Depiction of several forms of vias. Normal and "spanning" vias shown.

greater wiring density. Multilayer PWB processes have traditionally employed "barrel vias," in which a hole piercing the entire substrate is plated, creating a potentially severe routing problem in dense designs. The newer microvia processes[71] employ more of the VLSI-like structures, dramatically extending the capabilities of PWBs to handle wiring-intensive designs.

Wiring capacity can be related to the theoretical capability of the wiring medium to be filled with conductors and the efficiency with which this density can actually be used in circuitry of interest. Efficiency in a particular medium is affected by the amount of real estate consumed by vias. The number of vias in a design depends on the number of breaks in a wiring plane required in a design, which is driven by the routing complexity. This efficiency actually varies across a substrate, as local congestive regions decrease the efficiency. One of the most significant wiring challenges is referred to as the "escape problem." The escape problem is caused when packages (in a PWB) or ICs (in an MCM) have so dense an arrangement of I/Os that they create a situation where it is difficult or impossible to place a wire on every contact.[56,61,65] Solving the escape problem requires one or more of the following solutions: (1) the introduction of a wiring medium with greater density and efficiency (more compact via structures), (2) better I/O distribution, and (3) designs that lower wiring demand through alternate partitions and technologies (e.g., multiplexed higher bandwidth connections replacing large numbers of slower, parallel connections).

8.3.1.2.4 MEMS Copackaging

In MEMS device packaging, either the MEMS element is used as a stand-alone device (dropped into another assembly) or the MEMS device must be copackaged with other electronics (i.e., hybrid packaging). In the vast majority of cases, the latter option is the only practical one, given the difficulty of obtaining adequate integration of processes (i.e., finding a process that is good for both MEMS and microelectronics). Furthermore, to more appropriately reap the benefits of MEMS devices, copackaging is a sensible means for obtaining adequate efficiency in size and weight. In very simple cases, MEMS can be combined with electronics in a common package involving no MCM substrate by wire-bonding directly between the MEMS and microelectronics component, as shown in Fig. 8.27.[72]

In other cases, when MCM substrates are involved, MEMS devices may be integrated within one or more locations on an MCM substrate. Given the range of packaging options described previously, MEMS devices can exploit a number of planar and 3D techniques to achieve dramatic system reductions.

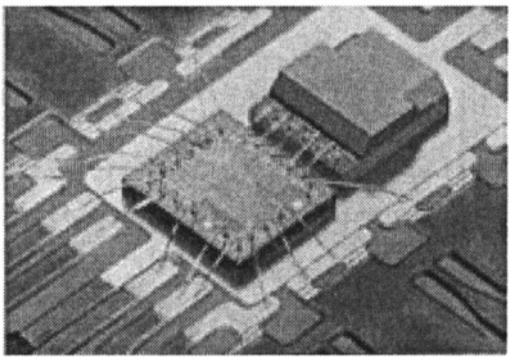


Fig. 8.27. Simple packaging of MEMS accelerometer with microelectonics IC die.[72] (Courtesy *Electronic Packaging and Production.*)

8.3.2 Basic Packaging Engineering Concepts

8.3.2.1 Electrical Interconnect Performance

An interconnections network handles the signal and power distribution functions in packaging. Desired electrical performance characteristics of these interconnections are domain-dependent, and the performance that can be realized from these interconnection structures depends on material and geometric parameters. Generally, ideal interconnections would have perfect electrical conductivity, zero loss, and infinite isolation. For cases in which the time-of-flight of the electrons is appreciable, the ability to deliver signals effectively from one point to another also depends on the impedance of the source, load, and interconnections. In the lumped-element regime, the electron time-of-flight does not influence performance. The distance over which the lumped-element approximation is valid is referred to as a lumped-element distance, a distance that decreases at increasing frequencies. When interconnect lengths exceed the definition of the lumped-element distance for a particular application, impedance control at the interconnections is required. This is one definition where the microwave frequency domain begins.

Signal integrity refers to the ability of the packaging to maintain fidelity of the electrical waveforms impressed. Distortions can come from many sources. The general categories of distortion are loss, delay, and noise. Loss results from signal attenuation and dispersion (a frequency-dependent form of loss). Delay is the result of a finite time-of-flight and in the case of digital systems by interconnection loss that induces delay transitioning from one state to another. The primary noise contributions in interconnections come from noise directly injected on the power-supply conductors (such as simultaneous switching noise), transmission line reflections, and undesired, coupled energy from neighboring interconnections and electromagnetic interference sources. We briefly review the sources that compromise signal integrity and examine the more realistic interconnection requirements driven by application.

Resistive Loss. Assuming interconnections as ideal is often adequate only in low-performance electrical systems or application domains where imperfections in interconnections are tolerable. For example, zero-resistance interconnections, a desirable attribute, is rarely a problem in low-performance digital electronics. The series loss due to the voltage divider across the virtually infinite input impedance of a digital gate input is inconsequential. As shown in Fig. 8.28, when a low-frequency signal drives a MOSFET (metal-oxide-silicon field-effect transistor) input (typical in VLSI) where the input impedance is virtually infinite, very little attenuation is experienced, but when a matched receiver is employed, or more generally a finite input impedance is encountered,

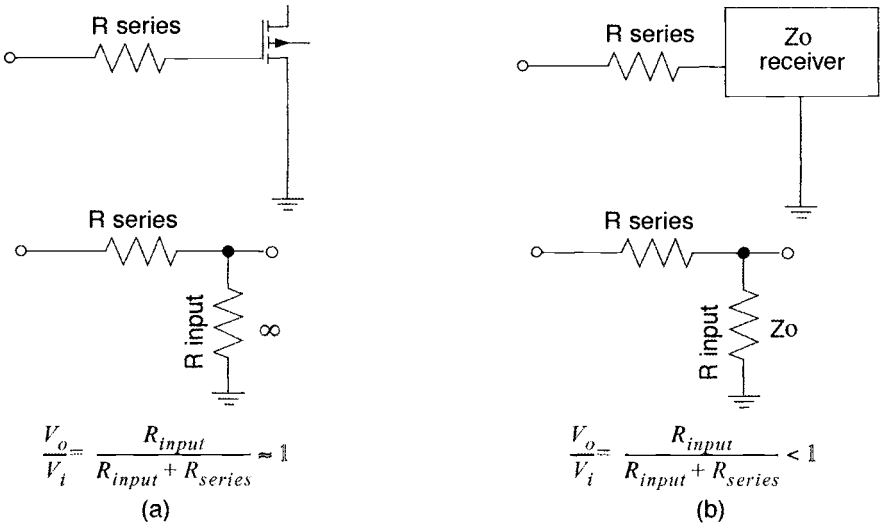


Fig. 8.28. Effects of input impedance on resistive loss. (a) infinite input impedance, (b) matched impedance. (∞ represents the infinite input impedance of a MOSFET gate.)

then voltage divider loss occurs. If for example, a very long interconnection has a 50-Ω series loss, a signal experiences a 25% attenuation when driving a 150-Ω receiver load.

Dispersive Loss. Dispersion refers to a frequency-dependent loss mechanism. Whereas a purely resistive loss mechanism does not affect the shape of signal, dispersion mechanisms attenuate some components of a waveform more than others, and the resulting composition is both attenuated and distorted in shape. The dominant sources of dispersion are material related and geometry related. An example of material property related dispersion is a dielectric permittivity that changes with frequency (usually decreases). Geometry-caused-dispersion sources include, for example, transmission lines that have an inhomogeneous material composition (permits propagation of undesired electromagnetic modes). Another dispersion mechanism that is affected both by geometry and material parameters is the loss in conductors at high frequencies due to the skin effect.

Signal Delay. The delay arises because the electron group velocity has a finite time-of-flight, which is given simply as

$$v_p = \frac{c}{\sqrt{\varepsilon_R}} \tag{8.8}$$

where c is the velocity of light and ε_R is the relative permittivity of the interconnection medium. Using Eq. (8.8), the time-of-flight delay for a distance l is given by:

$$\tau_{TOF} = \frac{l}{v_P} \tag{8.9}$$

In digital systems, an often more significant delay results from the distributive resistive-capacitive parasitics in the interconnection network, which distorts the pulsed waveforms in logic circuits. This dispersive distortion results in a delay in switching events marked by a change in the waveform crossing a certain fixed threshold. For those systems, the total delay is given by the sum of time-of-flight and resistance-capacitance loss-induced delays.

Delay is combatted in systems by using short interconnections of lossless (inductance-capacitance dominant) transmission line structures. Advanced packaging can more directly tackle the interconnection length, but loss is controlled by interconnection geometry and material parameters that are mostly process influenced.

Simultaneous Switching Noise. Another facet of interconnect performance is the quality of power delivery to components within packages and MCMs. Particularly at lower voltages (e.g., < 3.3V), the resistive loss in power delivery has a more pronounced effect because of I^2R losses. More problematic in complex digital systems is simultaneous switching noise (SSN). SSN is manifested in the form of a voltage spike, caused by circuit drivers in ICs rapidly switching states. The induced noise voltage as a result of SSN is approximated as:

$$\Delta V = NL_{eff} - \frac{di}{dt} \tag{8.10}$$

where N is the number of drivers switching at the same instant, L_{eff} is the inductance of the power delivery interconnections, and di/dt is the time rate change of current.[73]

Further exacerbating the effects of noise injected into power and ground conductors is the imperfect nature of conductors, which puts into question the assumption that a true ground reference or a power plane is exactly the same voltage at any point. Since it is not possible to eliminate inductance altogether, localized energy storage in the form of decoupling capacitors (between the power and ground terminals of particular circuits) may be essential, especially in high-performance systems, where many millions of gates can be switching within a system at the same time. Capacitance between power and ground serves a second important role: reducing the impedance of the power and ground planes. Often the two functions inflict competing requirements on capacitors, so it is often necessary to incorporate two capacitors at each decoupling point, one for localized energy storage and one for reducing the power-ground impedance. The number of drivers switching in one instant N can not be reduced substantially without compromising functionality, but staggering the switching intervals of some portions of a circuit or system element can provide benefit, if system synchronization is not compromised.

Asynchronous system designs that do not use a global clock may suffer less from problems in simultaneous switch noise, since in principle, many switching activities might be spread out in time, providing a much lower effective value for N. On the other hand, other design schemes, such as wave pipelining, which operate on multiple global clocks, could suffer potentially more from SSN, since more of the functional logic within a given monolithic component (or perhaps MCM) is operating in an instant.[74]

Transmission Line and Full-Wave Effects. When a significant portion of a signal's frequency content is above the lumped element distance, then reflections caused by impedance mismatches can have a pronounced distortion effect on the waveform. This can be particularly damaging in pulsed digital waveforms, as such effects can create distortions severe enough to cause multiple triggerings of logic functions. This distortion is caused by the superposition of time- and amplitude-shifted reflected versions of the original signal. When a wave is launched in the form of a signal from a logic driver, it traverses an interconnection indefinitely until a physical discontinuity, such as a circuit load, is encountered. A reflection occurs as a necessary consequence of the wave equation's boundary conditions when that discontinuity is not identical to the characteristic impedance associated with that interconnection. Termination of the interconnections in the characteristic impedance can eliminate this problem (assuming no other discontinuities in the interconnection create a reflection problem), but matched impedance receivers are uncommon in complex digital IC designers. Hence, by the previous definition, digital designers will always strive

to keep the *Manhattan distance* of circuits (the true length of the boxlike interconnect network vs the Euclidean distance, the true point-to-point distance) less than a lumped element distance. Failing to do so requires in some cases the use of special circuitry for matched drivers and loads, as well as impedance control in interconnections. Many claims in the literature of controlled impedance have little meaning when only the interconnection manifold is considered, given that the circuit elements attached to the manifold create most of the transmission line effects problems.

At increasing frequencies, each change or transition in the interconnection structure, including corner turns and vias, can contribute to undesired electromagnetic effects. In these cases, the system is considered to be operating in a *full wave regime*. Here, the 1D transmission line approximations become less approximate, as they are based on a transverse electromagnetic (TEM) assumption. At sufficiently high frequencies, longitudinal components are manifested in interconnections, making the associated waves non-TEM, and resonant modes can be established by any conducting material of the appropriate length. The entire interconnection manifold then can interact with the ordinarily dominant propagation modes, a situation requiring the special disciplines of 3D electromagnetics modeling in full-wave regions.

Electromagnetic Coupling. Crosstalk is a transmission line phenomenon involving the coupling of energy on a conductor from neighboring conductors. While crosstalk is considered a near-field phenomenon, some electromagnetic interference sources can be external to a substrate, circuit, or assembly. For planar interconnection media, the analysis of crosstalk is straightforward, assuming that the interconnection manifold can be viewed as a very sparse matrix. In other words, if coupling effects can be approximately confined to a few nearest neighbors, then crosstalk can be determined by using a coupled transmission line formulation for the interconnections. This formulation can be written as:

$$\frac{d[v]}{dz} = -([R] - jw[L])[i] \quad \frac{d[i]}{dz} = -([G] - jw[C][v] \tag{8.11}$$

where the $[R]$,$[L]$,$[G]$, and $[C]$ are the $N \times N$ per unit resistance, inductance, conductance, and capacitance matrices, respectively, and $[i]$, $[v]$ are N-element column vector representations of current and voltage. The assumption of a sinusoidal waveform reduces the basic coupled partial differential equation into this ordinary differential equation system. The system can be solved by establishing the per-unit parameter matrices through an extraction procedure, in which transverse cross sections of the associated interconnections are analyzed and solved with correct current and voltage boundary conditions. When the boundary conditions impress one source voltage only in a group of conductors, the crosstalk on neighboring conductors is solved. If the $[R]$ and $[G]$ off-diagonals are nonzero, then conductive leakage mechanisms exist; whereas if the $[L]$ and $[C]$ off-diagonals are nonzero, then electromagnetic coupling (EMC) mechanisms exist.[75]

Consideration of the coupling problems gives insight on their mitigation. By keeping interelement capacitance and inductance small, or the length of the interconnections short, coupling effects are reduced. All other things being equal, higher permittivity materials (e.g, ceramic vs polyimide) create higher amounts of crosstalk. In "unguarded" structures (no intervening grounding conductors or planes), the greater the separation, the less the crosstalk. These considerations each motivate particular approaches in materials, processing, and layout.

The more general problem of nonlocalized energy coupling or electromagnetic interference (EMI) is difficult to treat analytically, although some software tools exist that examine the manifestation and more importantly the effects to a necessarily limited degree. As a general guideline, sensible grounding and shielding practices are necessary, motivating the many standard *ad hoc* practices (e.g., EMI gaskets, conductive screens). Some of the packaging approaches can create

excellent approximations of Faraday cages. The more *ad hoc* practices of EMI mitigation become more exacting forms of design discipline in microwave systems, particularly at frequencies greater than a few GHz. Chapter 13 in this book explores some of these details.

Interconnection Performance Based on Application Regime. In the several application regimes, divided loosely into digital, analog, power, and microwave, different electrical parameters are emphasized depending on that domain. For example, digital signals are less sensitive to noise and can accept some degradation to gain wiring density. Series loss becomes important only in the sense that it affects delay. For analog signals, however, isolation is very important, as a result of the continuous dependence in analog functions on signal values. Interconnection resistance becomes more problematic. Controlling resistance values in these cases is often more important than minimizing them. For power waveforms, on the other hand, any resistance is undesirable. Finally, at microwave frequencies, many problems can occur from variations in impedance in structures because of dispersive interconnection structures (non-TEM structures, frequency-dependent permittivity, and skin effect) and discontinuities.

Material Considerations in Electrical Performance of Packaging. One of the top-10 trends identified in a recent technology forecast is for improved dielectrics and metals in integrated circuits. For dielectrics, the lower the permittivity, the better, at least for signal-bearing intermetal dielectrics. For metallization, especially in ICs, improved conductivity is seen as the most important enhancement for 0.25-μm ICs.[76] See Table 8.2.

8.3.2.2 Thermal Performance

Thermal management addresses the viability of a packaging system at its various levels to (1) maintain a specific thermal environment for a subset of the components it contains, and/or (2) keep component temperatures below a specified design point. The former concern is associated with the need to control the temperature of, for example, an oscillator circuit for stability or perhaps an imaging focal plane array to reduce Johnson noise. In one case, the temperature is close to ambient or slightly elevated, while in the latter case, it may be necessary to cryogenically cool particular components. The latter concern, keeping component temperatures below a design target, is a common requirement of electronic systems. The basis of most concerns in thermal management is the view that reliability follows an Arrhenius relationship and that it plummets as temperatures are high. Such a view drives a design requirement to keep, for example, the junction temperatures of transistors below a certain value (e.g., 150°C). While the Arrhenius assumption

Table 8.2. Electrical Characteristic of Representative Packaging Materials[77,78,79]

Material	Alumina	Polyimide	Silicon	FR-4
Representative class	MCM-C	MCM-D	MCM-D	MCM-L
Permittivity	9.0	3.4	3.9 (SiO_2)	4.5
Conductor material	W or Mo	Cu	Al	Cu
Sheet resistance	10.0 mΩ/sq	0.97 (1/2 oz) to 0.16 (3 oz) mΩ/sq	NA	0.97 (1/2 oz) to 0.16 (3 oz) mΩ/sq
Line widths/spaces	2–5 mils	10–42 μm	1 μm min	1–10 mils
Height between metal layers	125 μ	4–37μ	0.5– >4μ	40–60 μ or more

is sometimes taken for granted but should always be viewed in the context of reliability physics when analyzing failure modes.[80] Thermal management is also important to prevent failure caused by mechanical sources, such as fatiguing.

Heat transfer usually involves accessing surfaces of structures, which can be complicated in some packaging approaches. The mechanisms for heat transfer include conduction, convection, and radiation. Conduction refers to the intermolecular transfer of kinetic energy, which requires continuous interfaces. Thermal conductivity is a material property that conveys the effectiveness of a material to transfer heat by conduction. Convection cooling refers to heat transfer through liquid or gases. The effectiveness of convective heat transport is determined by fluid velocity and certain intrinsic fluid properties. Finally, radiative heat transport involves electromagnetic waves, and the rate of flow is determined by the fourth power of temperature (T^4).[81]

In space systems, only conduction transfer is useful within the interior of the spacecraft, and special panels are sometimes used to radiatively transfer heat by rejecting it into deep space. Convective approaches involving air are obviously impractical, while some concepts for self-contained fluid systems have some application.

8.3.2.2.1 Analysis

Most thermal analysis is based on the heat conduction equation,[82] and for many packaging problems the 3D time-dependent form is reduced to a simple one-dimensional, steady-state equation, which leads to a cookbook approach involving thermal resistances. Here, each interface between the junction of a transistor and the ambient is modeled as a resistor, which provides a loose, intuitive framework for considering the impacts of interfaces on thermal performance (Fig. 8.29).

In Fig. 8.29, each thermal resistance is given by $R = X/(kA)$, where X is layer thickness, A is the area through which heat is flowing, and k is thermal conductivity. In this simple model, the total thermal resistance is the sum of all intervening resistance components. An example expression for a plastic package might be given by:

$$R_{ja} = R_{device} + R_{mold} + R_{DieAttach} + R_{Leadframe} + R_{Convection-to-environment} \tag{8.12}$$

The difference in ambient and junction temperature with the dissipated power are related by:[83]

$$R_{ja} = \frac{(T_{junction} - T_{ambient})}{Power} = R_{jc} + R_{ca} \tag{8.13}$$

The total resistance, R_{ja}, can be conveniently divided into the sum of R_{jc} (thermal resistance of junction to package case) and R_{ca} (thermal resistance of package case to ambient). In general, lower levels of packaging hierarchy (at and below L2) deal with minimizing R_{jc}, while higher levels are concerned with minimizing R_{ca}.

As one might envision, these expressions are of limited applicability in most real cases. They approximate infinite planar slabs, a situation unrepresentative of most real packaging situations. Consequently, many thermal management problems must be delineated in formats that enable numerical analysis solutions such as the finite element method, where a problem domain is reduced into many elements, each with boundary conditions that can be solved to determine the overall

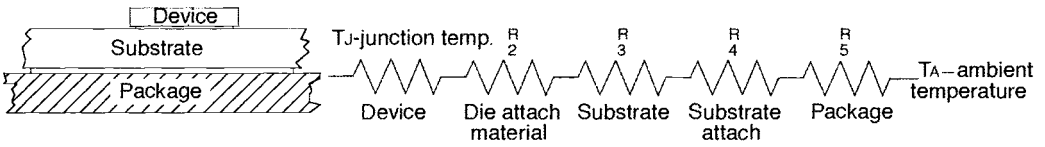


Fig. 8.29. Simple, 1D thermal model showing packaging structure as a series of resistances (after Ref. 81).

temperature distributions. While 2D analyses can provide some insight, 3D analyses are usually required to obtain reasonable estimates. Computer-aided analysis is an art form, since it requires meaningful data on material properties, knowledge of how problems might be partitioned to exploit symmetry or reduce dimensionality, and an understanding of the methods and idiosyncrasies of the particular tools to achieve convergence.

Hierarchical View of Thermal Management. Thermal management must be addressed in a hierarchical sense to be meaningful. It may make little sense, for example, to put thermally efficient MCM substrates onto PWBs with low thermal conductivity. Here, efficiency refers to the ability of a packaging entity (chip, module) to efficiently transport undesired heat generated within the assembly to the physical boundaries of that assembly. In aerospace systems, there are sometimes requirements to heat assemblies as well, and sometimes this can be done by intentionally adjusting thermal conductivities. As such, each cascading level of the packaging hierarchy can be thought of as a thermal "shuttle," and in order for thermal management to work as a whole, each segment must perform this function effectively. The following paragraphs present some thermal management concepts and issues at several levels of the packaging hierarchy.

Chip level. The assumption that power dissipation occurs uniformly across an integrated circuit die is viewed as reasonable for typical silicon-based digital microcircuits,[84] as silicon has good thermal conductivity. In the cases where a diversity of structures is present within a monolithic die, especially MEMS devices, the assumption must be revisited. It is conceivable that some classes of MEMS structures, because of the evacuation of large amounts of silicon, manifest an effective thermal conductivity lower than bulk silicon, creating a greater isolation of electronic functions within the same die. It is also conceivable that as feature sizes continue to shrink, the heat flux densities of silicon can reach a point where good conductivity of the bulk material is inadequate for beneficially spreading heat over the die (laterally and through the thickness of the die). Consequently, mounting the die directly to even a good heat sink will not adequately handle generated heat within the die. This problem exists in some nonsilicon semiconductors. In technologies such as gallium arsenide, more localization of heat occurs, as the bulk material has a much lower conductivity. The problem is combated by thinning the entire die to about 50 μm, coating the back with gold, and securing the component to a thermally efficient substrate or package.

Chip-to-Substrate. Monolithic components are secured to a substrate or package typically through a die-attach material. This material must have reasonable thermal conductivity and be uniformly distributed across the bottom surface of the die, as voiding in the die-attach material effectively creates an infinite thermal resistance in the voided area. The die-attach material can sometimes buffer the differences in thermal expansion coefficient mismatches between the die and substrate. This practice is problematic as the thickness of die-attach material would need to increase to be more effective at buffering, which would also increase its thermal resistance.

In the case of flip-chip devices, the mechanical and interconnecting substrates coincide, and a separate backside contact system is sometimes used to effect efficient heat removal. The bottom (unprocessed) side of the silicon is processed for the purpose of heat removal. For example, the back surface of the die can be textured or micromachined to enhance thermal transport. For supercomputer applications, a variety of approaches have been attempted[65] each of which add structures to the packaging system. In other cases, thermal bumps (vias) have been introduced to the chip (substrate), which are electrically unnecessary but serve to lower overall thermal resistance by increasing the number of thermally conductive channels between the die and the substrate. Underfilling the die provides some additional improvement by replacing air/vacuum regions between the die and substrate with a material having nonzero thermal conductivity.

Substrate-to-Package. In some MCM concepts, the MCM is the package. In others, the MCM substrate must be placed within a package, using a substrate attach material, usually an adhesive.

Substrate. In MCMs, thermal conductivity is the most important factor in facilitating heat transport from components to the next level of packaging. In the MCM-D and MCM-L cases, the interconnecting substrate is usually a poor thermal conductor, while MCM-C substrates offer significantly better thermal transport. Further enhancements to these patterned substrate technologies are possible. For MCM-D and MCM-L technologies, thermal vias can be distributed throughout the substrate underneath the die mounting locations, so as to increase the effective thermal conductivity locally. A variant of the patterned substrate approach discussed earlier, referred to as recessed patterned overlay, improves thermal management by "excavating" a region of the interconnect-bearing substrate area and mounting die there (Fig. 8.30). These approaches reduce thermal resistance, but do so at the expense of routing capability.

In patterned overlay approaches, which separate electrical and thermal paths, it is possible to achieve simultaneously high thermal[85] and electrical performance. In other cases, where the interconnecting and mechanical substrates coincide (refer to the previous taxonomy discussion), some compromise is necessary to either thermal or electrical performance.

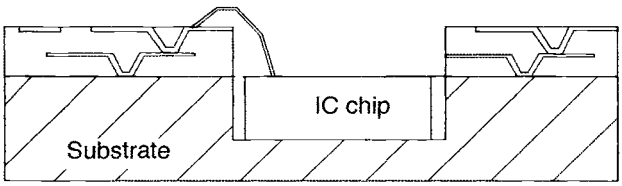


Fig. 8.30. Recessed patterned substrate.

3D MCMs. Modules stacked in an arrangement must effect special measures to deal with thermal management. Two important regimes exist in 3D assemblies from a thermal management standpoint: few-layer and many-layer MCM stacks. The few-layer regime [Fig. 8.31(a)] refers to a situation in which the number of layers in a 3D stack are few enough to permit adequate heat transfer through the stack vertically. The many-layer regime [Fig. 8.31(b)] cannot achieve adequate thermal management from vertical transport alone, but must rely on lateral thermal transport. This result follows intuitively even from a simple thermal resistance model. The success of thermal management in either case depends on designing transport mechanisms consistent with the regime. In both cases, the need for adequate thermal transport may compete with efficiency, and the associated structures should be included in any "fair metric" of 3D system density.

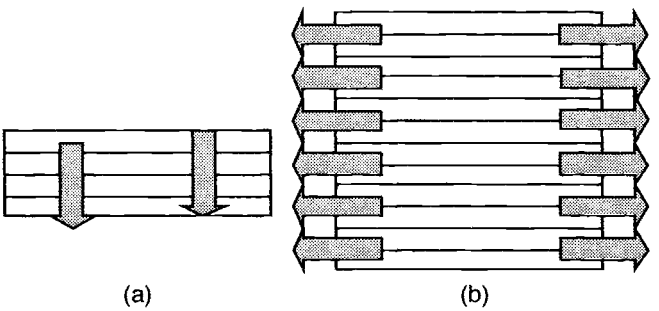


Fig. 8.31. Thermal paths of two 3D thermal management regimes. (a) few-layer (b) many-layer.

Package. The package for a single-chip or MCM assembly can be viewed as yet another set of thermal resistances, although packages can improve overall thermal management by spreading heat more uniformly. Additionally, the package can support more advanced thermal management approaches (e.g., embedded fluid flow channels) because of the extra physical structure. Most packages are designed to be mounted directly to boards, and in many cases the board is a vital link in the thermal management system. This is not always the case, as demonstrated in some supercomputer designs.

Boards and Boxes. The PWB plays a lesser role in thermal management, for good reason: boards are typically constructed from FR-4 or polymide laminates, which are very poor thermal conductors. In fielded military systems, ruggedized boards often rely on forced-air convection cooling, which is not an option in space systems. Space systems based on the standard board-box model rely often on conduction through the board. Two approaches are often used. In the first, boards are affixed to an aluminum cold plate, and conductors that must pass from one board to another do so through either the backplane or holes formed into the cold plate. In the second approach, a single multilayer PWB is used with one or more thick copper planes to provide grounding and lateral heat spreading/conduction. In both cases, the board-to-box conducts heat through wedgelock structures, which can be a critical path.

Box-to-Platform. Boxes in space systems are mounted to structural panels, and the nature of the usual interface between the box and panel is a plane-to-plane contact with good thermal conductors over a large surface area. When two presumably flat surfaces are placed together, the microinterface is typically rough (Fig. 8.32), and heat transport is a hybrid between conductive and radiative heat transport. To improve thermal transport, the effective contact area must be increased through the use of a softer material (e.g., thermal grease) between the two surfaces. Solutions include greases, oils, foils of soft metal (lead, indium), composite material and adhesives, and surface treatments (e.g., deposited films).[86]

Trends in Thermal Performance. It is generally projected that the average heat flux of microelectronics assemblies has increased during the last several decades. In microelectronics, the reduction in feature size has permitted increased frequency and a corresponding increase in power dissipation all in the same unit area (less power per device, but more devices). Advanced packaging approaches exacerbate the situation, especially 3D approaches, as they make it possible to accumulate large numbers of such components in the smallest possible size. Of course the need to reduce voltage in microcircuits because of the potential of gate oxide breakdown has favorably affected power and heat flux per unit area, but this mitigating effect does not fully compensate

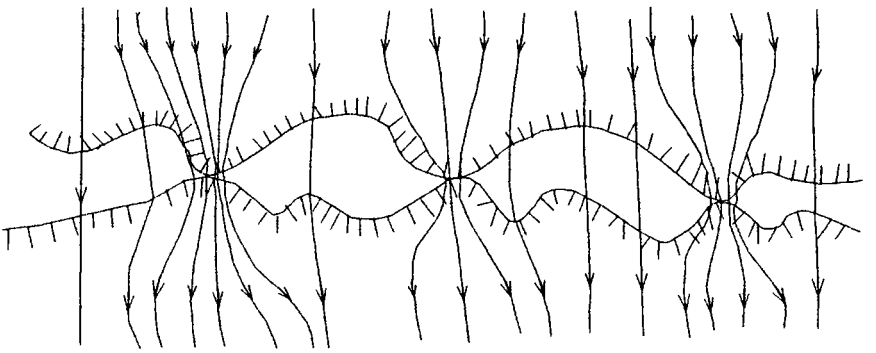


Fig. 8.32. Heat flow between surfaces (after Fletcher).[86]

for the increase in circuit and packaging density. As such, complex 3D packaging arrangements must be "punctuated" by adequate thermal management approaches. In terms of sheer density, these accommodations, though a necessity, serve to erode any fair metric that measures the quantity of components, say per cubic centimeter. Claiming higher densities based on incomplete packaging structures, at least in real systems, is illusory. These considerations lead to the suggestion of limit arguments in packaging density as a function of "thermal content."

8.3.2.3 Mechanical Reliability Considerations

Many problems in mechanical integrity within packaging systems occur when joined materials experience differential expansion. Usually this occurs as a result of mismatches in thermal expansion coefficients between material interfaces. The magnitude of stress per unit area in a film-to-substrate interface can be expressed as:[87]

$$\sigma = \frac{E_f}{1-\nu}\Delta\alpha\Delta T \tag{8.14}$$

where E_f is the elastic modulus of the film, ν is Poisson's ratio for the film, $\Delta\alpha$ is the magnitude of difference in the thermal expansion coefficients, and ΔT is the magnitude of temperature change. Other mechanisms for manifesting stress include moisture absorption, material phase transformation, and (in metallic materials) grain growth.[87] In components on substrates that experience severe thermal cycling as a result of (for example) rapid power cycling, severe differential stress can occur even when $\Delta\alpha$ is zero, as is the case for a silicon-on-silicon substrate assembly. Differential stress has many manifestations in packaging, including die and package debonding, film delamination, conductor lead and solder ball fractures, and popcorning.

Other mechanical problems in packaging come about because of deflections in large structures created by resonances at low frequency. Lid deflection, for example, in large packages has been studied extensively.[88]

8.3.2.4 Materials and Compatibility

Constructing a viable packaging system requires development of a "wrapper" that does not impede the efficient delivery of "services" (e.g., signals in-and-out, power in, heat out), and sometimes the materials in a packaging system compete with this desire. Specific examples illuminating this issue are provided:

- Signal interconnects benefit from low capacitance, but power interconnects benefit from high capacitance.
- High electrical conductivity is desired in sensor interconnections, but the high thermal conductivity usually associated with this property creates thermal losses, making sensor packaging more problematic.
- Many dielectric materials with good conductivity have high permittivity, suggesting the need to trade between high thermal conductivity and high electrical performance.
- Die attach often must mate semiconductors and substrates with widely different thermal coefficients of expansion. Since most die-attach materials have poor thermal conductivity, this condition suggests the need to trade between high thermal conductivity and reliable die attachment.
- High operating temperatures (>150°C) create many obvious compatibility problems.

Other issues of compatibility exist in fabrication processes. In patterned-overlay MCM processes, for example, components embedded within substrates must withstand the processing temperatures. When MEMS devices are introduced into MCM processes, copackaging with other

electronics components requires a manufacturable solution that includes a MEMS-device release process and does not damage the MCM and components. Recent investigations have illuminated aspects of this problem and have highlighted solutions.[68,89]

As such, the balance of considerations in material and process compatibility against performance limitations in IC and MCM processes set boundaries on application domains and use environments. Changes in the IC component size or process, for example, may further open or constrain these boundaries. For example, a larger IC builds more differential stress in thermal cycling when on a dissimilar substrate. While a 1-cm die size might work over a 100°C excursion in a particular case, it may be that a 1.25-cm die size will fail as a result of shearing from its mount. Furthermore, a 1-cm die with a higher power cycling performance may also fall outside a permissible use boundary.

8.3.3 Approaches to Design

The proper design of packaging is intertwined with the design of a system. A breakdown of a structured engineering process that integrally addresses packaging is shown in Fig. 8.33. Though automation is desirable, design reflects the ability to map aspects of problems into partitions that are readily solvable and may require a solution process not mapped in automation. In new technology fields or in the first-time combinations of existing technologies, boundaries are crossed that are not often captured in existing automated tools. While some may conclude that design realization is indeed a process,[90] it is important to realize that humans will continue to play a necessary role in all but the most trivial aspects of design. Several basic engineering principles are summarized in an excellent essay by MacLennan,[91] who took lessons taught by bridge makers and applied them to the design of computer languages:

- Efficiency—seeks to minimize resources used.
- Economy—seeks to maximize benefit versus cost.
- Elegance—applies a tenet that "designs that look good will also be good."

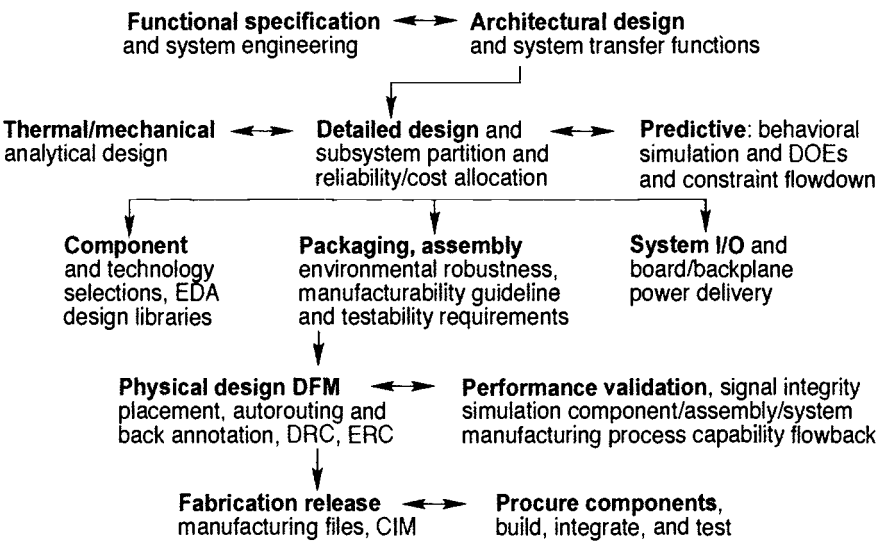


Fig. 8.33. A structured electronic design process: a design methodology that considers packaging.

Nevertheless, automation is not only a good thing, but also essential for most practical work done in advanced packaging and underlying component technologies. Designers need both power and flexibility, captured in the form of tools that free design tedium, while allowing the necessary access to certain details and aspects of design to which aesthetics can be channeled. In packaging, MCM design automation has become increasingly more sophisticated. More comprehensive assemblies (e.g., 3D packaging) are, by contrast, embryonic in their infrastructure.

8.3.3.1 Role of Automation in Design and Fabrication

A Computer Aided Design (CAD) information system generally supports all modern electronic design processes because of the complex and collaborative nature of current design processes. Automation can only be achieved by adopting a structured process. In CAD, the design process is typically structured by the Electronic Design Automation (EDA) vendor and in-house EDA management system resources that together provide tools and interfaces that may be applied in a consistent methodology. These interfaces include the graphics user involved with design (schematic/layout) capture, the library manager involved with components, the test manager involved with design-for-testability, and the factory interface involved with design-for-manufacturability (DFM) and assembly. These interfaces support an iterative design process. Because electronics payloads in space systems are made of many classical functional domains (power, RF, digital, analog), interface between many such CAD processes may exist in one system design environment. Standard interfaces have evolved to permit interchange of electronic CAD-generated data.

Electronic design libraries form the backbone of support to the two principal processes that dominate the CAD and manufacturing processes: component information systems and module or board fabrication processes. At the manufacturing site, assemblers will depend on the CAD system to provide industry standard databases that meet manufacturing requirements for quality. Automation in the design process plays the following key roles: design cycle reduction, design reuse, design sustainability, consistent design methodology, design library development and maintenance, and iterative optimization for DFM. Automation in the fabrication of electronic assemblies includes CIM (computer integrated manufacturing) as well as documentation, distribution, component and assembly inventory, process assembly and test scheduling, factory standard design file creation, version control, design rule checking, and DFM.

8.3.3.2 CAD

Design automation technology remains the most significant advantage to shrinking design development cycles. In logic design and optimization, timing analysis, hardware software codesign and verification, automation tools enable significant savings in time and reduce the potential for defects to pass into fabrication, where cost to discover errors are the highest. Many of these tools and physical layout techniques migrate from integrated-chip design tools to MCM and PC board design environments. However, the MCM design provides microsystem integration at high component densities and high speeds with several newer technologies that have begun to employ sensors and MEMS technology. The challenge of CAD tool performance for these developments has begun to be addressed in the DARPA (Defense Advanced Research Projects Agency Composite) CAD program that will begin to offer new EDA products a few years from now.

Finite-element-analysis tools can add to CAD tools the ability to explore the physical design space for mechanical and thermal simulations. EMC/EMI design tools have emerged to provide detailed signal integrity simulation for completed and predictive physical layout, including examination of crosstalk and interlayer 3D coupling on circuit performance. As increased speed and reduced geometries create concern over parasitic effects, newer CAD tools must be employed to examine performance.

8.3.3.2.1 Discussion of CAD Flow for MCMs

The CAD infrastructure for MCMs has evolved considerably over the last decade. As improved understandings of interconnect performance, component selection and availability, thermal management, and other requirements have been achieved, CAD tools have evolved to incorporate facets of this understanding. MCM CAD tools provide the design engineer with all pertinent criteria necessary for successful layout and interconnect design integration. Process technology is typically offered in the form of a design guide from the manufacturer that provides the user's EDA system with the "technology reference files" for correct construction layout. Design libraries must also support the component data management including physical data (geometry), performance data (critical pin function or power-supply connection) and schematic mapping (logical pin to physical pad) files. Often there are additional performance libraries associated with simulation, and in the case of FPGA devices, additional map files per instance of the device (a common trap for interconnect optimization routines). To permit back annotation of physical interconnect lengths or proximity effects, CAD tool interoperability can aid the information process in early identification of potential flaws being created during physical design. DFM techniques can be more effective if fabrication process guides are also available in the CAD tool database or are separately available to ensure compliance.

A sample CAD flow for the electronic-design process is given in Fig. 8.34 and shows the task areas and tool categories necessary for proper design implementation and database creation. In this case, separate front-end design capture and physical design construction may share data through interfaces provided by the EDA vendor and/or open database connectivity. While often the status of EDA standards prevents some productivity enhancement because of tool incompatibility, it is also possible to employ *de facto* standards that remain common within the data exchange community and for which translators are low-cost and universally available. Homogeneous and heterogeneous design environments are two grouping categories for user environments and CAD tools. In a homogeneous design-tool environment, all EDA CAD tools share a common user environment in an encapsulated or integrated manner (e.g., mentor, cadence). Heterogeneous CAD environments consist of several programs operating independently but linked together to create functional tool methodologies (by various vendors). Generally, heterogeneous environments can link a "shopping list" of the best software modules and can include batch-mode simulation and verification software, to aim at incorporating the best in class tools. Because EDA vendors offer configurations that may include a wide variety of software modules to enable cost/performance trade-offs, information system managers must target their requirements to design domains most likely to be encountered by their users.

8.3.3.2.2 CAD for 3D Packaging

Three-dimensional CAD remains relatively unexplored. The extension of planar MCM concepts to 3D approaches has been examined in which a 3D assembly is decomposed into a number of logically linked 2D representations.[92] In the case of stacked planar MCMs of identical size, MCMs can be viewed as components of a system whose terminals are projected onto one or more planar backplanes. The net list of such backplanes can be routed in a traditional way (as, for example, a custom VME backplane would be), defining the interconnections within the entire 3D assembly in terms of the connection patterns of contacts emanating from the planar constituent MCMs. Other research has examined the decomposition of 3D assemblies into a number of tower routing problems, more representative of a true 3D routing problem.[93]

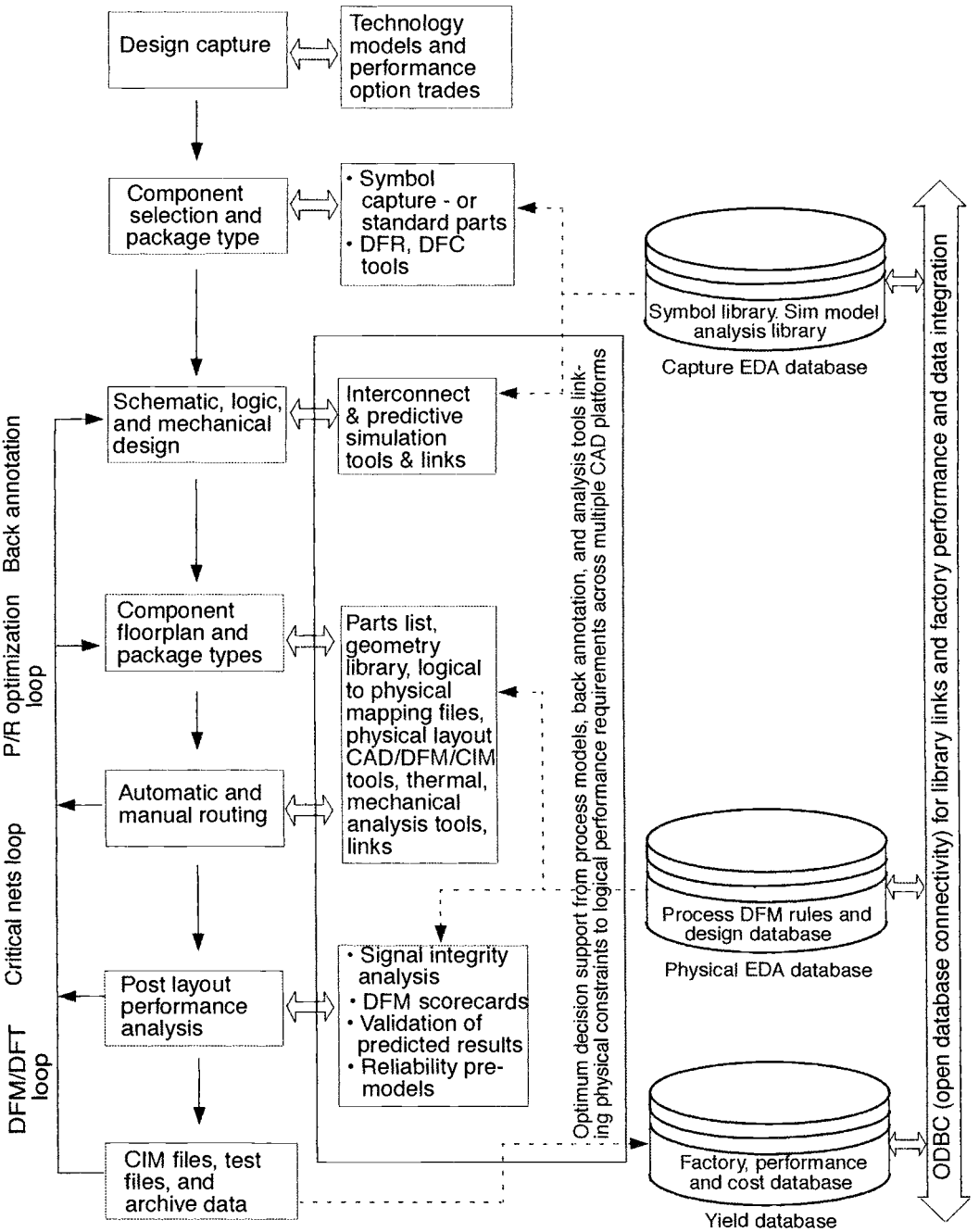


Fig. 8.34. Tasks and tools in the design process.

8.3.3.3 Virtual Process

The idea of a virtual process involves nonlocalized realization of one or more parts of an overall fabrication process, especially in a dynamic manner. Given enough diversity and sophistication, it may not be possible to establish and maintain in one facility all the techniques required to make a complex MEMS device, integrated circuit, multichip module, or system. In the packaging hierarchy, a set of relatively "clean breaks" naturally exist for transporting subelements. In a complex 3D MEMS structure, it may be necessary to design a process particular to a single device type, such as a microthruster. The features needed for good microthrusters are rarely common with the features needed for good gyros or switches. Specialty assemblies cannot command the level of investment required for a dedicated fabrication line, and by definition no existing "commodity" process will have everything needed for their construction. Unfortunately, the flexibility to freely barter microfabrication process services is much less developed than, for example, the commodity-oriented business of building PWBs. Understanding how to develop virtual processes will probably be essential for achieving maximal integration of microassemblies. In order to understand how to make virtual fabrication work, it is necessary to address a few basic questions:

- Why do high-volume processes achieve better quality control?
- How can virtual processes be brokered or arbitrated?
- How can a reasonable set of process/product qualification guidelines be set up dynamically?
- How can design rules be dynamically set up for particular processes?
- How can virtual processes be created, retired, reestablished over time, even as particular facilities and the processes that they support come and go?

8.3.4 System Issues

Systems are collections of components in one sense, so system issues span the various components at a particular level of regard. It is necessary to include level of regard, as pieces can sometimes be considered systems themselves. The boundary of what is a system is further confounded by jargon such as "system on a chip," which is sometimes referring to things that are neither systems nor chips.

It is important to realize that not all elements embody a full ensemble of the system's requirements. In other words, the pieces of a system are not necessarily a microcosm of a system. In a satellite, for example, the requirements of a satellite's bus (the vehicle) are not the same as the requirements of a payload. Hence, the system (satellite) will have subsystems (the bus and payload) that have some common requirements and some distinct to the subsystem.

8.3.4.1 Total Realization of Functions with Minimum Cost

With the variety of components they contain, MCMs represent a complex cost optimization problem. It is at least possible to obtain access to a variety of sophisticated MCM technologies, but cost is a substantial driver. The cost objective may not always be met with minimum component count or aggressive design, but may depend on the end quantities. In space systems, it may be more economical to build MCMs because of the large cost of qualification than to employ many individual discrete components. Here we consider the cost problem by reflecting upon some key factors that could affect cost.

8.3.4.1.1 Known Good Die and Known Good "X"

The so-called known good die (KGD) problem refers to the limitation of yield that occurs in MCM assemblies caused by failure of one or more untested die in an assembly. In many cases, IC die are not tested until packaged, and once packaged cannot be recovered for use as bare die

for MCM assemblies. This problem is amplified, as previously discussed, when more components are used. Large and/or complex MCMs are then driven to great expense because of yield fallout. Wafer probing can provide some insight about a die's performance, but in many cases it is not possible to exercise ICs at full speed with traditional probes. When MCMs are built, failure rates can still be high when compared to board versions of the same assembly.

KGD has been referred to as a technological nonproblem, since it is predominately caused by a lack of will on the part of die manufacturers to solve the problem. Besides, it can be argued that a number of approaches exist to solve the KGD problem physically at least. The solution of KGD involves three aspects:

- Die fixturing
- Environmental profile
- Test vectors or exercising signal patterns to which the die will be subjected

The first problem has been the center of focus in much of the KGD research; both die-level and wafer-level approaches have been defined to theoretically solve KGD. Die-level testing for KGD focuses on die that have already been sawed by a wafer; whereas wafer-level approaches are similar to more traditional wafer probes. Die-level testing approaches usually apply one of the following techniques:

- Low-impact wire-bonding (die-level)
- Wire-bonding to alternate bond pad sites on same die (die-level)
- Bladder probe (copper polyimide) (die- or wafer-level)
- Test on TAB bonding frames (die-level)
- Redistribution of chip I/Os (basis of all chip-scale packaging approaches) (die-, assembly-, or wafer-level)
- Sacrificial redistribution layer (removable pelt) (die-, assembly-, or wafer-level)
- Sacrificial substrate (e.g., saw-away flip-chip substrate) (die-level)
- Snapstrates (break-away test substrates that are dropped as an assembly onto a final substrate) (die-level)
- On-chip self-testing (self-booting or injected into assembly through simple interface) (die- or wafer-level)

The environmental profile refers to what conditions to which die or wafers are subjected for testing. In space systems, a class-S burn-in is often prescribed on ICs, which is usually a 168-h (1 week) burn-in. Yet particularly in MCM assemblies, shorter tests that identify a comparable set of failures may be more cost-effective, even though the end assembly is subjected to the longer burn-in. Examples of this accelerated test include the so-called below-a-minute burn-in (BAMBI) approach.

Even with finding solutions for the first two KGD aspects, potentially the most expensive part of establishing a KGD baseline is subjecting the devices to a set of tests adequate to identify a failure mode. The test vectors for complex ICs are often very large and very proprietary. When they are not proprietary, they are rarely in a format compatible with other test machines. Clearly, a complex MCM with ICs from several vendors faces a miserable proposition in getting every component rigorously tested in the true spirit of KGD, which is why this is considered less a technological problem than a cultural problem. Fortunately, a larger number of vendors are offering KGD products in bare die, TAB, or CSP form. MCM designs that involve engineering the ICs within the MCM are in a potentially advantageous position of engineering a KGD strategy.

Establishing known-good MCMs may not require having all KGD. In memory modules, it is common practice to include spare die in MCM designs; however, redundancy is not always

possible, especially in mixed-signal modules that cannot afford the complexity or real estate for truly redundant approaches. Other options include a strategy of repair-to-yield, in which case a substrate is reworked until fully functional. Low-cost MCMs may not have this option, in which case the MCMs are discarded, and yield statistics must be favorable for this practice to be viable. As such, simple MCMs are better as a design approach. What might have been a large 2-in.-sq MCM in 1993 might be four simple 1-in. MCMs today. The boundaries for properly sizing MCMs are based on the yield statistics of the components and the assembly process. It is soon apparent that the known-good philosophy must be extended throughout packaging designs, not just MCMs. As designs become more complex, for example a 3D-MCM stack, it is important to apply recursion in the definition of known-good assemblies, which imply known-good modules, known-good submodules, etc.

8.3.4.1.2 Process Improvements

Complex processes are costly, and reducing complexity should arguably reduce cost. Process reduction and optimization is a rather involved subject, but in principle reducing the number of steps and the amount of labor and/or time in existing process steps are worthwhile considerations. For example, photoimageable polyimides and BT[94] resins might allow reductions in processes that address the dielectrics and photoresists separately.

Process simplifications are treated as system issues when multiple component processes can be involved. For example, in a complex MEMS process with integrated electronics, it might be possible to divide components along process lines, particularly in MCM implementations. In this manner, the MEMS portion of a design might be built in a simpler MEMS-specific process, and the electronics would be built from a standard IC process. Such a concept does not impact any of the IC processes but seeks to avert the use of any processes that are more complex than necessary.

8.3.4.1.3 Substrate Design

A number of cost-mitigation practices can be exercised through substrate design. The first principle here is in deciding if an MCM is needed at all. Sometimes an MCM is an unnecessary complication, particularly when cost is a driver and performance, size, weight, and power are not. Given the choice of MCM implementation, the selection of a correct process is important. Some designs do not require the highest density process, and sometimes the highest performance design is not the highest density design. In the VCOS process, for example, or a silicon-based MCM-D process with normal (thermally grown) SiO_2, very high interconnection densities are possible at the expense of highly lossy resistance and high capacitance. Such as in the case of SSCOF-HDI, it is possible to employ prefabricated Cu-PI film to reduce cost without sacrificing performance. Sometimes, a chip-on-board process is ideal for low-cost, low-wiring-complexity applications. Using the wrong MCM process for a particular design is sometimes worse than not using MCMs at all. For example, high-density planar MCMs built in an MCM-D process are usually less cost-effective than a 3D stacked package for memory implementations, particularly when hermetic enclosures are involved.

When an appropriate MCM process is selected, a number of basic principles can be applied to reduce potential cost during design. At least two apply to the KGD problem: small substrates and redundancy. When these can be incorporated into an MCM design, they are generally beneficial. MCM I/Os should be distributed in a manner that permits a lower-complexity wiring medium at the next packaging level. When high contact density or pad-limiting is evident in a perimeter design (e.g., flat pack), a BGA or LGA approach should be considered to reduce substrate size. The carrier or panel that MCMs are built from should always be considered in the design process, and substrate sizes should be chosen to maximize utilization.

In patterned overlay design, liberal quantities of redundant via contacts should be used on bond pads whenever possible (i.e., this practice does not impact performance) to improve reliable contact and improve power delivery. This guideline may actually be necessary when probe damage is evident. In patterned overlay designs, whenever possible, double booking of the interconnect real estate should be exploited. It is often possible to put components above and below the interconnecting substrate. It is advisable to put more complex ICs underneath the overlay and surface mount passive components, particularly if tuning is required. In cases where an experimental component is involved, on the other hand, it may be advantageous to place that component on top of the overlay and employ a repair-to-yield strategy.

Any element of an MCM design that is subject to change should either be removed from the MCM design or made easily accessible. An example of this is fuse-link programmable memory components. In many cases, that component will require replacement. A general statement for any device that can be programmed but not within the system should be avoided. This would include a great variety of one-time programmable memory and FPGA devices. Fortunately, in-system programmability is an increasingly available feature in such devices, although not the case for space systems at the time of this writing.

8.3.4.1.4 Economies of Scale

Economies of scale can be exploited if they exist, and it is not always possible to determine that they do. In many aerospace applications, fewer than 10 units of a particular MCM design are required, which makes the proposition, for example, of tooling custom TAB frames particularly disconcerting. On the other hand, if an aerospace design is to be used *en masse* in several applications, some design-specific tooling can be greatly beneficial to the program in the long run in both cost and schedule. Low-quantity runs may benefit from the use of low-volume prototyping technologies described previously. Economies of scale also apply to back-end processing (assembly and test) as well as fabrication.

8.3.4.1.5 Die Optimization

Die optimization refers to the exploitation of IC design based on the unique enabling benefits of MCMs. ICs may be designed with lower power, smaller size, and higher performance when they are "keyed" to an MCM design, particularly when access to the entire die surface is possible. ICs are generally designed to drive the parasitics of conventional package leadframes and PWB trace lengths. Often it is necessary to buffer the IC signals with several driver stages, each of which consumes power, occupies space, and adds delay to the signal. In some cases, transmission line design techniques are used because the Manhattan distances associated with the interconnection manifolds exceed a lumped element distance. If those drivers are optimized for the shorter distances and lower parasitics inherent within an MCM substrate, significant economies may be realized in some cases. If, for example, a die-optimized formulation is used for an interconnection manifold and reduces the lengths below the lumped element distance, it is possible to avoid transmission line design constraints, which results in considerable design simplification. Propagation delay reduction and greater control of path-lengths can lead to improved simultaneous delivery of waveforms (reduced clock skew) and increase in the clocking frequency, which when combined with techniques such as wave pipelining,[74] can tremendously improve performance in digital designs. Recent work in the HDI process, for example, has led to the conclusion that with time-of-flight delay control with specially designed patterned overlay designs, uncertainties in the subpicosecond range may be feasible.

Die optimization reduces cost by promoting simplified MCM designs for a given performance level. These economies must be balanced against those associated with design verification. When

die optimization is thought of as promoting a "seamless" MCM design philosophy, then an MCM can be considered as a large IC device. As such, a small MCM with few components is treated as a monolithic unit. In this case, the entire MCM is tested at once, as though it were a large IC chip. The point where this assumption breaks down is a function of IC process maturity and MCM design complexity. If KGD issues surface on a large MCM that is treated as a seamless design, then the cost increases as a result of considerable complications in testing. KGD strategies are very difficult to implement on truly optimized ICs because of test fixture loading.

8.3.4.1.6 Qualifying Die vs Qualifying Assemblies

MCMs are often more expensive in space programs because of a multiplicity of test and qualification procedures that are applied at both a component and an MCM level. In more complex assemblies, a number of recursive qualifications could be envisioned, resulting in extremely expensive systems. Using a more progressive view, MCMs and more complex assemblies need to be built with known good elements, but the assemblies, not their pieces, should be qualified. If a KGD procedure can be rapidly implemented, then there seems to be little justification in performing, for example, a 168-h burn-in, particularly if the MCM will receive another 168-h burn-in. MCMs can actually reduce cost when the considerable cost of qualifying many individual components can be eliminated with the qualification of a single, integrated assembly.

8.3.4.1.7 Plastic Packaging

The controversy of plastic packaging in space is discussed in "Hermeticity and Hermetic Alternatives" under Subsec. 8.3.4.2.1. Here, we consider the benefits of plastic packaging for cost reduction in space systems. The first obvious advantage is in raw weight reduction, particularly in multichip assemblies. When combined with integral package concepts, a plastic MCM can easily save 50–90% of the mass of a competing ceramic assembly. The lighter packages tremendously simplify system assemblies. Since plastic packaging has lower mass and therefore inertia, the assemblies require less structural reinforcement, leading to less bulky and more robust assemblies.

8.3.4.2 Environment

Environment for space as it relates to package comprises two essential facets: effects of space on the packaging and its contents, and effects of packaging and its contents on the rest of the spacecraft (e.g., payloads). These facets are discussed, together with the need to control environments within a package.

8.3.4.2.1 Protection as a Role of Packaging

Packaging in a properly designed system will provide an adequate environment for operation of its contents despite external environment conditions.

The Space Environment. A summary of typical operational environmental requirements for space applications is given in Table 8.3. The table is notional, as it is impossible to prescribe a universal specification that could both cover every possible mission scenario and be met by an electronics assembly. Interplanetary mission scenarios, which can have dramatically different radiation, shock, and temperature requirements, are not addressed here.

Radiation is typically cited as a special environment for space systems. Also of concern to low Earth orbiting system is the atomic oxygen hazard. The oxygen in these orbits remaining in the atmosphere at these altitudes acts as a plasma that can attack exposed surfaces. Spacecraft charging effects are also a severe concern in some cases, and plastic materials can aggravate the problem of dielectric charging. It is a common practice in spacecraft designs to provide conductive bleed paths from every piece of metal in an electronics system to mitigate charging effects.

Table 8.3. Summary of Operational Environmental Requirements for Space Applications[95,96]

Mission Requirements	Low Earth Orbit	Geosynchronous Orbit
Temperature	−65 to + 120°C	−196 to +128°C
Thermal cycles	6000 cycles/yr	90 cycles/yr
Vibration/acoustic	Launch up to 20 G RMS, sound pressure to 145 dBs	Launch up to 20 G RMS, sound pressure to 145 dBs
Outgassing	<100 ppm	<100 ppm
Radiation	Orbit, time dependent	Orbit, time dependent
Gravity	10^{-6} to 10^{-3} G	10^{-6} to 10^{-3} G
Pressure	10^{-8} to 10^{-3} Pa	10^{-11} Pa
Plasma	0.3 to 5×10^{5} particles/cc 0.2–0.2eV	0.24–1.12 particles/cc 120–295 keV
Atomic oxygen	10^{14} atoms/(cm-sec) (1000 K exoatmospheric)	10^{7} to 10^{8} atoms/cc

Though these environments can be challenging, the harsh environment of Earth, not space, can more often provide the more difficult challenge. Many electronics in space systems are subjected to flight testing, long periods of storage, and transportation, as they eventually make their way to launch. Outside of radiation effects, the space environment can in some respects be comparatively benign in comparison to prolonged storage.

Hermeticity and Hermetic Alternatives. Plastic packaging, or nonhermetically enclosed polymeric material usage, is an area of continuing controversy in space applications. The origin of most concerns in plastic packaging took place many years ago, when significant amounts of ionic contaminants existed in plastic used for packaging. These contaminants, when combined with moisture, created a catalyst for eventual destruction of the ICs within. After nearly 20 years of continued improvement in both the materials used in plastic packaging as well as process control, plastic packages have been noted to be more reliable than hermetic packages.[96,97] These arguments are based on the fact that packaging materials have much lower ionic contamination levels. This issue still raises a significant polarization in the community.

The opposition to the use of nonhermetic packaging appears to hinge on two basic arguments: (1) something will get into the ICs and damage them, and (2) something will outgass from the packaging materials and damage sensitive parts of the spacecraft. The first argument was dominated by the issue of ionic contamination. Residual concerns also exist regarding very long-term storage, as experienced in munitions and nuclear weapons. It is important for these systems to have the ability to function reliably and immediately, even after 20 years storage. Very long-term storage raises the possibility for the manifestation of second-and third-order failure effects that would never be seen in other applications. Here, even humidity cycling over many years could create failures in nonhermetic circuits. In modern space systems, this does not appear to be a common condition, despite the cases where long storage occurs, which is usually less than 5 years.

Outgassing, the second concern of nonhermetic materials, refers to possible contamination of materials that would be emitted from surfaces and would condense onto optics, solar cells, and other sensitive payload instruments. NASA standards and databases for low outgassing materials exist, and many packaging materials meet appropriate specifications.[98] Minimizing the outgassing in spacecraft requires reducing the surface area of outgassing contributors and advanced packaging approaches to accomplish this. In some cases, the cable harnesses of a spacecraft are worse contributors to outgassing than electronics.

Precedents do exist for the use of plastic encapsulated microcircuits (PEM) and nonhermetic assemblies[99,100] in aerospace systems. It has also been shown that artificially prohibiting the use of plastic on military programs has resulted in tremendous disadvantages to include cost and lack of supply sources.[101] Of course, failure modes and some mysteries still exist regarding the use of plastics. Popcorning is a well-known problem in which absorbed moisture causes cracking in a package, usually caused by inadequate bakeout procedures. Other failure modes can occur from delamination and cracking of passivation caused by severe thermomechanical stress.[97] Finally, radiation-induced leakage currents have been reported to be higher in plastic packages than in hermetic packages, specifically after burn-in.[102]

If it is required, hermeticity can be achieved at various assembly levels. At the die level, for example, sol gels have been widely studied,[103] and more recently technologies such as ChipSeal.[104] Die-level hermeticity ideally protects die, but does not prevent package-level problems such as popcorning. At the substrate level, a number of studies at AFRL have examined the use of hermetic coatings, such as Si_2N_3[105] and diamondlike carbon (DLC)[106] at a package level. A class of polymer, liquid crystal polymer (LCP), has been shown to have hermetic performance levels comparable to the glasses used in package. LCPs can be used to create PWBs and other packaging structures, including enclosures.[107]

Shielding. Radiation effects are best combated in electronics through process hardening, which addresses the common space radiation effects (total ionizing dose, latch-up, and single-event upset) by carefully considering how semiconductors can be processed to minimize charge trapping. Even with hardened processes, careful attention to design practice is required. Some circuits and architectures, such as transmission gates and dynamic logic, for example, can be problematic even when those circuits are cast in a rad-hard process. As such, process hardening and design principles must be combined. Some improvements to existing processes can be realized in a harden-by-design methodology, which has been practiced successfully on simple gate arrays. In commercial processes, however, the hardness level can vary from fabrication run to fabrication run, so care must be exercised when hardness is not assured in the inherent semiconductor process.

Sometimes process hardening cannot be performed on components because of the expense or lack of access to the appropriate intellectual property of a design. It is possible in some cases to improve the radiation hardness of components by surrounding them with shielding materials. It is important to note several important limitations to shielding approaches. First, the degree of improvement falls rapidly as more shielding material is added. The benefit in the first millimeter of shielding material is dramatic, the benefit of the second far less, and so on. Second, shielding benefit depends on mass, and thin shields are better if made from a high-Z (atomic number) material. Finally, and most importantly, shielding only addresses total ionizing dose effects. Shielding does virtually nothing to help the critical (potentially destructive) latch-up problem, nor can it significantly mitigate single-event phenomena, well known as "soft-errors." Single-event problems can be addressed by architectures in some cases, but latch-up problems are sometimes impossible to dismiss without thorough examination of a design at a transistor and process level.

Shielding approaches are nevertheless very popular and in some cases the only way to provide some protection against radiation to many classes of components. The first non-*ad hoc* approach to shielding in packaging was established through the development of Rad-Pak, invented at the Air Force Weapons Laboratory,[108,109,111] a precursor to AFRL. Rad-Pak, now marketed by Space Electronics, Inc., (San Diego, California), and other comparable approaches have gained increasing use in space systems. They work by replacing a normal single-chip package with a package similar in appearance but with shielding materials embedded above and below to provide nearly complete coverage. They are popular, effective with due consideration of the previously mentioned concerns, but they are heavy—and much heavier still when applied to large numbers of components, one at a time in a complex design.

Fortunately, the advent of advanced packaging, along with some new technologies, can do much to reduce the mass penalty. For example, more efficient versions of Rad-Pak approaches can be formed. Research at AFRL has resulted in the development of a novel chip-scale package version of Rad Pak, shown in Fig. 8.35. As shown, this CSP Rad Pak exploits a combination of floating-pad miniball grid arrays and a high-density tungsten step lid for improved sidewall coverage [Fig. 8.35(c)]. The package achieves only hemispherical coverage by design, allowing a user more strategic control of the final shielding. Options include implementing symmetry at the substrate level [Fig. 8.35(d)] or simply applying a closing plate on the opposite side of a board assembly [Fig. 8.35(e)]. The mass penalty for the CSP Rad Pak is minimal: the package shown in Fig. 8.35 has a 4-g mass, which becomes about 8 g in a spherical coverage case.

Approaches like Rad Pak need not be restricted to single-chip packages. Even simply applying shielding to MCMs can reduce the mass penalty as viewed on a per-component basis. The greatest benefit in a shielding system is gained by shielding the smallest possible volume, implying the application of shielding to the surfaces of the most compact 3D assembly possible. This consideration is obvious, as a planar arrangement can consume much more surface area than an efficient 3D approach. The graph in Fig. 8.36 serves to illustrate the benefits of shielding a 3D arrangement of components. The shielding can be integrated within a 3D packaging system by careful structure design. Also, a new class of radiation shield coatings exist that can in principle be applied to a completed assembly.[112]

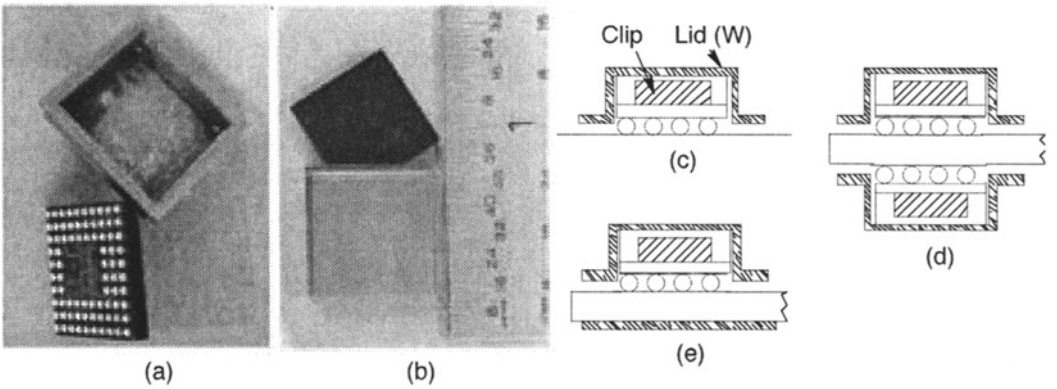


Fig. 8.35. Chip-scale form of radiation package. (a–b) photograph of separate pieces: min-HDI BGA and tungsten step lid, (c) cross-sectional view, (d) application employing symmetry, (e) application showing tungsten plate to complete shield system.

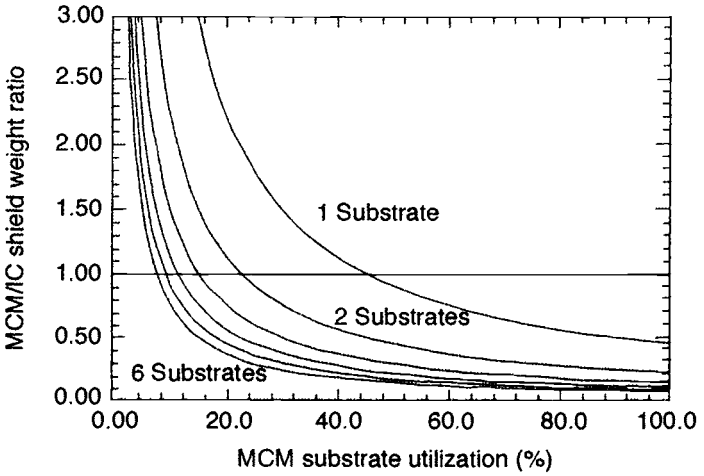


Fig. 8.36. Reduction of weight possible in a 3D packaging system normalized to a planar assembly as a function of substrate efficiency and the number of substrates in a stack.[113]

8.3.4.2.2 Control of Environment Inside a Package

While packaging is expected to provide some protection of internal components from an exterior environment, it is sometimes necessary to create special environments for particular components within a packaged assembly. Examples include microhermetic volumes, temperature control for precision operation, and isolation from vibration. In the case of microhermetic volumes, it is sometimes necessary to ensure that the ambient density around a particular component does not change and the potential of particulate contamination is minimized. This is commonly true for piezoelectric crystals, MEMS accelerometers and gyros, and internal optics perhaps for a detector. Sometimes the opposite is true; in other words, an environmental portal is required so that a MEMS sensor can sample an environment's chemical composition, temperature, and pressure. Sometimes it is necessary to maintain temperature control at a component location to guarantee precision, as is sometimes needed in precision oscillators. Many infrared systems require cooling a focal plane array to cryogenic temperatures to reduce Johnson noise. In most cases, this implies a dedicated cryocooler engine, which for space is often a closed-cycle engine. Limited research has been performed on chip-scale cryocooling techniques involving MEMS. In principle, such microcryocoolers could be embedded within an electronics package, but many challenges must be met to make these devices practical. Isolation from vibration is sometimes important in applications related to visual sensing and inertial referencing, and the engineering of the packaging system is sometimes critical to achieve reductions in certain vibration modes.

8.3.4.3 Test and Verification

Test and verification is one of the least glamorous but yet most important steps in a development process. Testing is done for verifying design, verifying product performance, assisting in the modification of hardware and software, and troubleshooting and maintenance. Design for testability is emphasized here, since without the intent of verification, it is unlikely that designs will accidentally have the necessary degrees of controllability and observability to permit this verification after the fact.

For MCMs and complex 3D systems, boundary scan approaches are a very powerful concept for test and verification.[114,115] The IEEE (Institute of Electrical and Electronics Engineers)

1149.1 standard for boundary scan has been widely accepted, not only for test, but also for device configuration and system maintenance. Boundary scan operates by inserting registers at the inputs and outputs of subcomponents, components, modules, and assemblies and linking them into several serial scan chains. Under the control of a test access port controller, it is possible to set and read patterns for the purposes of configuration and verification (Fig. 8.37). Implementing boundary scan requires intimate access of a system during the design phase, and may only be partially effective when components that do not support boundary are used. Other types of testing approaches complement boundary scan approaches, such as quiescent current monitoring (also known as IDDQ testing[116,117]) and the *ad hoc* built-in self-test approaches.

Additional complications result from not being able to physically access components within MCMs and complex packaging structures. Fixturing is, for example, a major issue with the new families of BGA and CSP components. For a complex 3D packaging system under construction, it is difficult to implement system-level testing. In-circuit emulation is also not possible for many of the system-on-a-chip concepts, particularly when specialized processing hardware is involved.

Breadboarding. In a perfect world, systems could be simulated and built to work correctly the first time. While great strides have been made in IC simulation, complex MCMs often comprise components that cannot be simulated because of lack of intellectual property access. Furthermore, many problems unforeseen in most simulation systems can afflict a system's design. In recent AFRL designs, problems were experienced in shorting together 5 V and 3.3 V supplies, one of many things that the sophisticated VLSI and MCM design tools did not check. For these reasons, breadboarding a system beforehand is not necessarily a "belt and suspenders" approach, but rather an important risk-reduction step in systems development at the current state of the art in design realization. In breadboarding, a system to be implemented in advanced packaging is brought to a

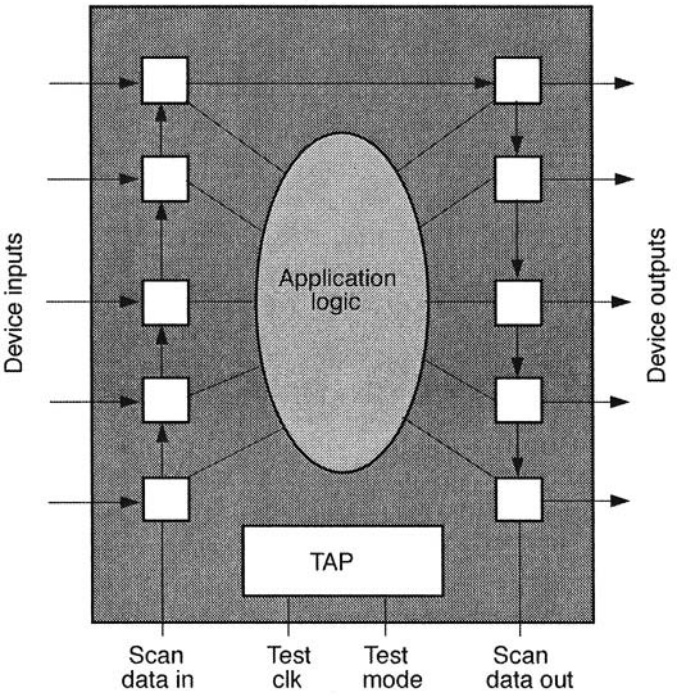


Fig. 8.37. Joint Test Action Group boundary scan concept.

form referred to as high-fidelity netlist, in which a board containing the discrete parts constituting the MCM is built with as close to identical a netlist as possible. The board is socketed for system insertion. Only after functional verification is achieved with the "high-fidelity netlist" is the design released for fabrication. After fabrication, the resulting MCM is temporarily socketed into a board identical in size and fixturing to the high-fidelity netlist format. The MCM is then inserted interchangeably and functionally verified. The procedure of high-fidelity verification is recursed as necessary in more complex approaches, such as 3D-MCM stacks. The high-fidelity netlist is not without drawbacks, particularly in die-optimized cases where fixturing access is difficult or impossible because of the severe differences in parasitic capacitance. Another pitfall that often occurs in schedule-pressured development programs is the temptation to "shot-gun," that is, to attempt to verify breadboards while the MCM is in design (or worse, fabrication).

8.3.4.4 Putting the "Play" into Plug and Play

The vision of plug and play epitomizes an effective open-systems approach, since it employs a number of open-systems standards in a synergistic manner to achieve interchangeability and interoperability. The plug-and-play concept is familiar as a result of the attention it has received from recent attempts to improve hardware interchangeability in PC hardware. The results of these attempts have been the formation of a set of industry specifications.[118] Despite these specifications, which were built upon physical (connector), electrical (e.g., 5 V / 3.3 V), logical (high-performance, low-power CMOS), software characteristics (drivers layered onto a standard operating system), many early industry attempts to achieve plug and play were problematic, leading in some cases to the less flattering descriptor, plug-and-pray.

Nevertheless, the plug-and-play concept is clearly the highest and most desirable level of open-systems implementation. In the limit argument, a true plug-and-play system would require no hardware bridges, fixes, or software patches, and it would adapt to integrate, in real time, with theoretically unknown assets if strict compliance to the appropriate plug-and-play specification were achieved. Unfortunately, by the same token, plug and play is one of the most abused descriptors used to advertise or market components and even subsystems. Often wide-ranging applicability is confused with the real ability to achieve plug and play.

8.4 Advanced Approaches

This section examines the following advanced approaches to packaging.

- Some extensions of packaging boundaries presented through case studies and techniques
- An advanced 3D packaging system that embodies a fresh look at packaging: the Highly Integrated Packaging and Processing (HIPP) program is developing this concept, one that seems to address the key issues of both micro- and high-performance systems.
- A hypothetical third-generation in MCM packaging
- An approach to remove many boundaries in packaging, multifunctional structures (MFS)

8.4.1 Extending the Boundaries that Define Packaging

Three broad approaches exist for engineering a system of packaging. The first is to simply use available technologies, such as to order from a catalog or do commodity-oriented practices, such as printed circuit board or cofired ceramic substrate design. The approach is safe albeit limited in the potential benefits accrued. The second approach involves localized optimizations, such as engineering a system element for a preexisting system in which only the new element can receive benefit of advanced packaging engineering. The final approach, reminiscent of a system-on-a-chip philosophy, encourages reviewing a large piece of a system, perhaps several subsystems, for exploitative opportunities in package engineering.

In the latter cases we have the opportunity to extend the boundaries of packaging, and not necessarily force a rote procedure. We begin considering some possibilities by first reviewing two case studies in HDI packaging, followed by a discussion on some techniques applicable to high-performance microinstruments.

8.4.1.1 Case Studies in High Density Interconnect

8.4.1.1.1 Radiation-Hardened HDI Space Computer (RHSC)

Type of Design. Stand-alone ruggedized general-purpose computer. The RHSC CPU (central processor unit) was based on the Lockheed-Martin rad-hard 1750A, and the design contained a collection of radiation components from many other vendors representing a component cost well in excess of $100,000/module, making the RHSC one of the most costly MCMs based on component cost alone. The RHSC design is unique in its capability to operate in nuclear radiation environments without upsetting real-time operation and without loss of data.

Radiation-hardened implementation contained, besides a dual lock-step 1750A microprocessor, several megabits of main memory and special-purpose integrated circuits for memory management, data protection, and system control. Though seemingly simple compared with advanced commercial microprocessors, the RHSC is designed to provide military space platforms with a survivable computer in the minimum size, weight, and power possible. In its form, the system replaces a 10.8-lb computer while providing dramatically improved robustness from the implementation of an operate-through approach that protects critical data and real-time operation from upset in a nuclear environment. Traditional practices for space systems require hermetic assemblies because of concerns of reliability degradation from long-term moisture permeation and the outgassing products of hybrids and MCMs affecting other sensitive instruments in space systems.

Advanced Packaging Technology Used

- Large format, complex digital MCM system 3.8 × 2.5 in., >500 I/Os, 12.3 W, hermetic assembly
- HDI patterned overlay MCM substrates
- HDI subtile construction (planar MCM inset within another planar MCM)
- 3D, edge-interconnected stacking methodology

In the RHSC, the HDI process was exploited in new ways, as shown in Fig. 8.38. For example, a minimodule (1.4 in. sq) containing the 1750A processor was fabricated, separately tested, and inset within a larger substrate as an HDI subtile component. As in the case of individual components, the minimodule was also placed into a larger, planarizing substrate with dozens of other integrated circuits and passive components. This compound MCM was then subsequently integrated into a stacked ensemble of two identically sized substrates using the 3D extension of the patterned overlay process. Photographs of the RHSC before and after insertion of the subtile are shown in Fig. 8.39. The complete construction of the very compact RHSC system involved no fewer than five multilayer copper-kapton interconnect systems, as shown in Fig. 8.38.

The 3D RHSC system required hermetic assembly into a package; the associated structures are depicted in Fig. 8.40. The special package developed involved a series of highly dense multilayer ceramic (MLC) inserts, placed within a kovar enclosure. MLCs were used in the RHSC because of the tight pitch (<50 mils precludes the use of the more traditional glass beads). The kovar can was designed to permit wire-bonding from lands on the top substrate to an inner bond shelf formed by the MLC inserts. After wire-bonding, the RHSC package was sealed with a lid through a standard seam-welding process, forming the desired hermetic enclosure.

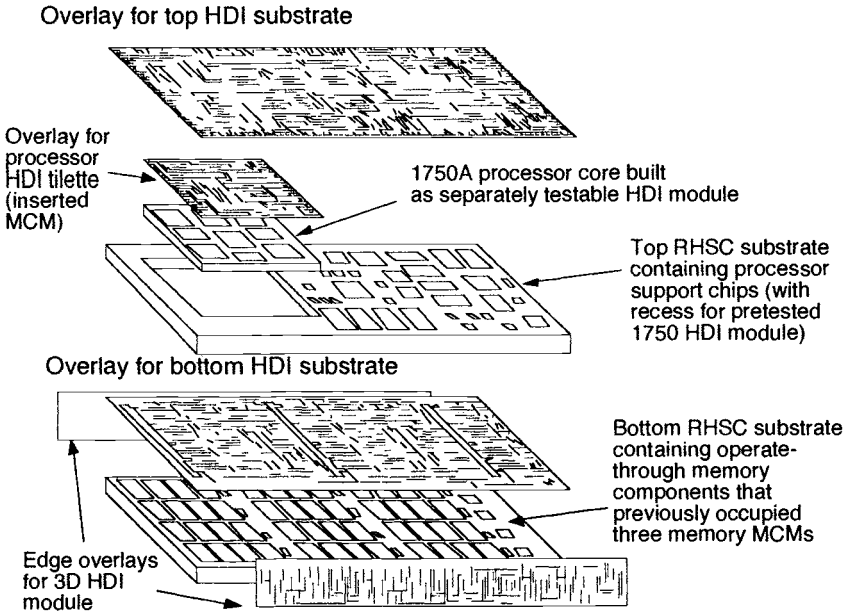


Fig. 8.38. Exploded view of an RHSC module, illustrating the use of HDI subtiles and 3D assembly.

Fig. 8.39. RHSC processor substrate, illustrating HDI subtile concept: (left) unpatterned substrate, (right) substrate with HDI subtile insert.

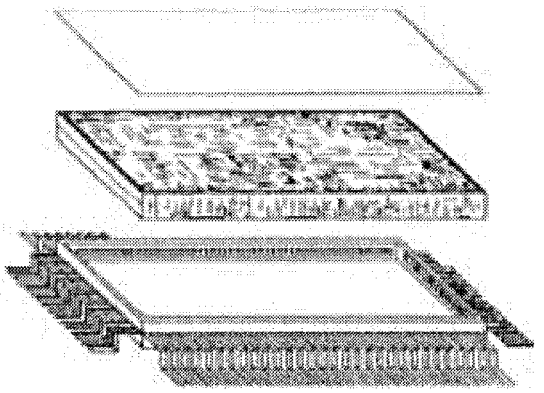


Fig. 8.40. 3D RHSC and associated hermetic package.

A Real Product or a Prototype for an MCM Technology? RHSC was built as a generic demonstration of advanced packaging and survivable computer technology for various U.S. Department of Defense applications and funded through the Defense Nuclear Agency. A limited number of prototypes were built successfully as a proof-of-principle demonstration. The design was proven in brassboard form prior to MCM design in high-radiation environment tests. The MCM itself was specially developed. The fabrication and initial demonstration were in spring 1994.

Unique Design Features. RHSC represents one of the most complex MCM systems built to date. It examined new forms of HDI, the integration of large MCM form factors in 3D form, the unique provisions/requirements of HDI MCM construction with subtiles, the trade-offs (electrical, thermal, mechanical) of a 3D MCM, the merging of IC components from a number of disparate suppliers, known-good-die, and known-good-module.

Lessons Learned

- PROM components locked into substrate were impacted by later software changes.
- Partition of the design was driven to two large substrates instead of four smaller substrates because of the I/O requirements, which drove the minimum perimeter size used in RHSC.
- Testing drove much of the I/O demand and most of the special features of the RHSC, especially the subtile configuration. A module-on-board approach facilitated testing each large substrate.
- The subtile was designed to drop into an existing package used by Lockheed Martin, simplifying functional test of unique ICs.

How an MCM Approach Benefited this Development. Space systems require maximal economy of launch weight and power consumption. The MCM implementation of the RHSC replaced large brassboards, which themselves contained one or more MCMs. The closest off-the-shelf correlating to the RHSC before it was designed was a 10.8-lb computer, which was less mechanically robust and would have been disrupted by nuclear environments through which the RHSC could have operated.

The key result of the RHSC project was to press the limits of integration and complexity for a specialized application, including large-format MCMs, compound MCMs, and 3D MCMs

8.4.1.1.2 Advanced Instrument Controller (AIC)

Type of Design. Stand-alone low-to-medium performance, general-purpose processor designed for radiation-tolerant operation, with versatile interface/operating options. The AIC block diagram, shown in Fig. 8.41, involves a tightly coupled processor-memory-analog interface-chip combination. Through a tightly coupled MCM design approach, the AIC achieves a 3 g, 1.0- × 1.4-in. form factor and a 50-mW-power budget, and is designed to operate in 30,000-G environments at operating temperatures down to −130°C. The AIC was developed under NASA/USAF funding for the Deep Space II interplanetary probes attached to the Mars 98 mission.

Advanced Packaging Technology Used

- Few-chip, mixed-signal MCM system (1.0 × 1.4 in., 120 I/Os, plastic assembly)
- Tightly-coupled MCM design
- Plastic HDI patterned overlay MCM substrates
- Surface-mount components for trimming end performance
- In situ reprogrammability of memory

Die-level interconnections for plastic HDI are as in the standard HDI patterned overlay process. In the AIC, modules are sawed apart after carriers containing 6, 8, or 12 modules each are fabricated (like IC wafers). Most surface-mount components are then mounted and soldered, and

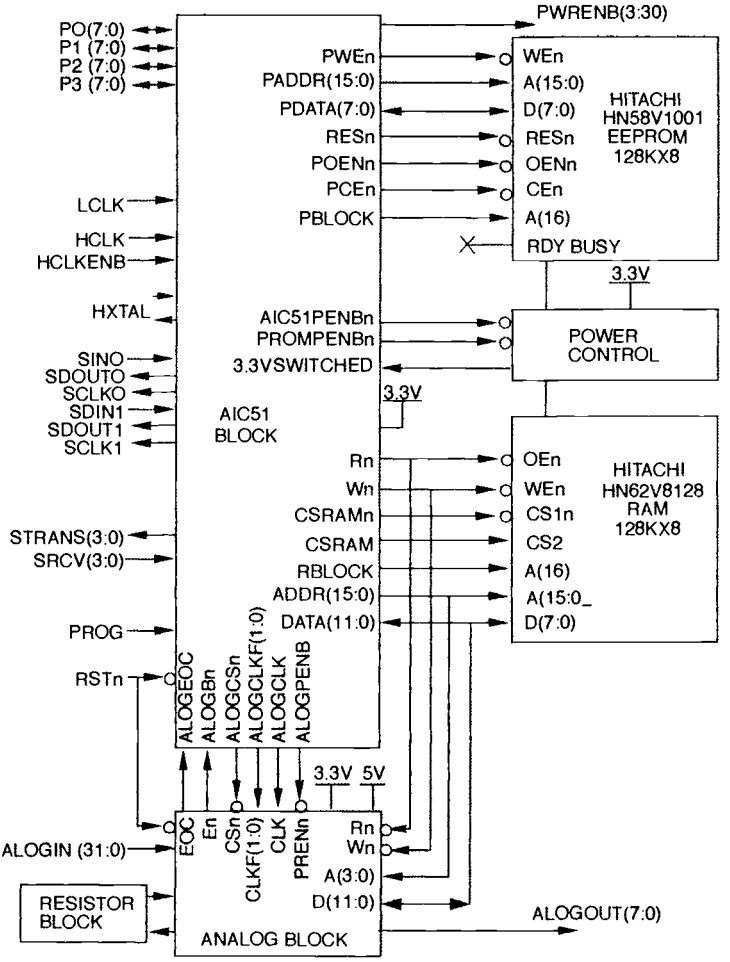


Fig. 8.41. Architecture of the AIC.

the balance are added after testing. In the initial modules built, modules were wire-bonded temporarily to small boards for evaluation, and then transferred to flight assemblies (e.g., another board), where wire-bonds are replaced by a flex-based lead attachment. Later generation AICs employ lead frames added during surface-mount attachment, which greatly reduces difficulty of testing and end use.

The AIC leaves hermetic assembly as an option, which has not been required by current users. The AIC can be placed within a kovar or ceramic package, if required, similar to traditional ICs and MCMs. AICs are being investigated for a number of applications with different end-packaging arrangements, ranging from using AIC as a subtile to fully enclosing the AIC within a small kovar package.

A Real Product or a Prototype for an MCM Technology? The AIC was an enabling application and provided a unique opportunity to do "intelligent things" with advanced concepts in plastic-based HDI technology. While ceramic HDI technology had received exposure to a number of flight projects, the plastic form of HDI had not been used in previous applications (since AIC, other flight experiments based on plastic HDI have been established). The appeal of the AIC, a

simple ultralow-power utility controller designed for space application, has not been limited to the original application. A number of other NASA and DOD programs are developing systems for flight with AICs or are evaluating AIC samples for potential use.

Unique Design Features

- CPU modified from an 8051 design
- 128 K × 8 SRAM and EEPROMs
- Analog application specific integrated circuit (ASIC)
- Large number of resistors bundled into a resistor-ASIC
- A p-channel FET, and other passive components

The AIC employs tightly coupled MCM design, which permits exploitation of

- Nonperipheral die area
- Smaller drivers on integrated circuits (sized for less than 10 pF drive)
- Large I/O between components within an MCM with relatively low I/O count
- Splicing in components from multiple processes

The AIC's CPU, for example, has more than 150 I/Os, while the MCM has only 120. The resistor ASIC reduced many fabrication uncertainties as well as design time for the analog ASIC, since the resistor values could be optimized independently, regardless of the sheet resistance values of the analog IC. The unwieldy number of resistors (~50) are absorbed within the MCM, freeing the user from the burden of more complex integration. In some applications, the AIC only needs applied voltage, as the collection of components (including even two oscillators) required for a minimal system are within the MCM. The AIC represents, therefore, a system-on-a-chip, albeit a relatively simple system. The ability of nonperipheral access was not exploited, however, as this was viewed as too traumatic a change in the current IC design/verification culture.

Some of the ICs used in the AIC are shown in Fig. 8.42. The CPU, built in the National Semiconductor 0.35-mm process, is obviously pad-limited [Fig. 8.42 (a)], based on the concentration of VLSI interconnect in the central region of the die. Had a distributed I/O array been used, the die size could have been substantially reduced. The CPU contains a barrage of user I/Os (including 32 discrete I/Os and 6 serial ports, along with a variety of power-management features and interfaces). Some of the interesting power-management features of the AIC include its ability to select from a palette of internal and user-provided oscillators and a separate copy of the entire data and address bus that is generated for the EEPROM (electrically erasable programmable read-only memory) to enable the AIC to physically remove power from the EEPROM when it is not in use without creating an undesired bus-loading condition for other components.

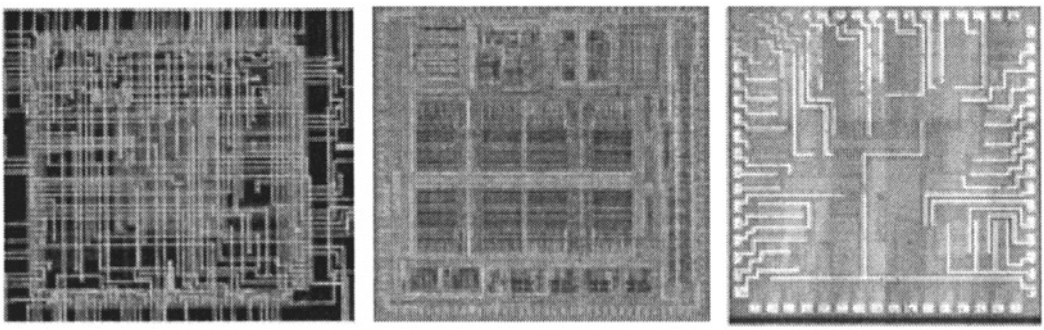


Fig. 8.42. Some of the ICs used in the AIC. (a) AIC51 CPU, shown here during an intermediate fabrication step (HDI traces visible over die and bond pads), (b) analog ASIC, (c) resistor ASIC.

The analog ASIC [Fig. 8.42 (b)] built in the Orbit Semiconductor 2-mm process contained approximately 70,000 components and implemented a surprising array of functions:

- 32 external A/D channels, 12-bit resolution
- 16 additional internal A/D channels, 12-bit resolution to monitor ASIC health and status
- 8 individually programmable DAC channels, 10-bit resolution (fed back to some of the internal A/D channels)
- Band-gap reference
- Proportional to absolute temperature (PTAT) thermal sensor

One of the most interesting features of the AIC is the in-situ reprogrammability of its program and data through one of the six AIC serial ports. The AIC, because of this feature, can be reprogrammed without disassembly. Furthermore, the AIC can literally be personalized with a variety of unique data, such as serial codes, calibration coefficients, even a reduced "traveler" containing process history. In the Deep Space II mission, the AIC is designed to function with discontinuous applied power, because of the high probability that the extreme cold will periodically render the battery temporarily nonfunctional. The AIC, by virtue of this capability, can display history-dependent behavior and be put to sleep for extended periods of time.

Lessons Learned

- The AIC did not exploit the ability to access nonperipheral bond pads on its ICs, which was fortunate at least at the time this decision was made, as it would have been difficult to pretest the die. The impacts of pretest (somewhat short of the KGD functional guarantee) can be profound, even with few-chip MCMs.
- At the time of this writing, the analog ASIC is not pretesting and is consequently the most common failure mode.
- Simple wafer probing of the CPU has kept the yield within tolerable bounds. Hence, tightly coupled approaches must be assessed carefully.
- It is believed that in time, a few-chip design can more advantageously exploit reduced drivers and nonperipheral distribution, in which case the MCM is viewed merely as a monolithic component.
- Similarly, the choice of a resistor ASIC was found to be much more costly than previously imagined. Newer versions of the AIC were designed without it, and a 75% cost savings and 12-week schedule savings were realized. This finding is disturbing, as the elimination of many passive components heuristically should reduce cost, indicating opportunities for improvement in the current component development infrastructure. Newer technologies with embedded passives could possibly be exploited.
- The most sobering impact that the tightly-coupled design has on a time-critical project is that it requires concatenation of the development of both the ASIC and the MCM, resulting in a very lengthy development cycle.
- In the AIC development, both the CPU and analog ASIC were first-pass successful, but the resistor ASIC and MCM were not. In the former case, no overall schedule impact was experienced, as the refabrication was effected while the CPU and analog ASIC were still in development. The MCM design error, however, resulted in a 4-month schedule impact. The schedule impact could have been doubled had the CPU contained flaws.
- The MCM design flaws, while preventable, were not caught by traditional tools.
- Conclusions are that while tackling both die and MCM designs, especially for a tightly coupled design, results in optimal size, weight, power, performance, and ultimately low recurring cost, the risks can be great. Mitigating this risk requires careful design and planning up front.

How an MCM Approach Benefited this Design. The MCM implementation of the AIC is shown in Fig. 8.43. AICs could not have been built without MCM technology. While not a high-performance device, the AIC represents a greater amount of silicon than a monolithic IC can accommodate, but not much greater. The use of MCM permits the independent optimization of separate components. For example, better EEPROMs in the future can replace the current ones, and enhancements can be phased in gradually. As monolithic silicon improves, the AIC can either shrink through the use of BGA/CSP approaches, or increase in functionality. When the current AIC can be rendered as a single chip, new AICs with added functionality will be realized such that the packaging remains just outside the realm of monolithic implementation. This is in direct contrast to the WSI research of the 1980s, in which WSI circuits could often be realized monolithically (with higher performance) during the horizon of a single project. On the other hand, designs such as AIC seem to offer the greatest possibility for success because they can stay just outside the reach of monolithic ICs.

The key result of the AIC project was to demonstrate the utility of a simple system-on-a-chip concept, made possible through a tightly coupled MCM design approach.

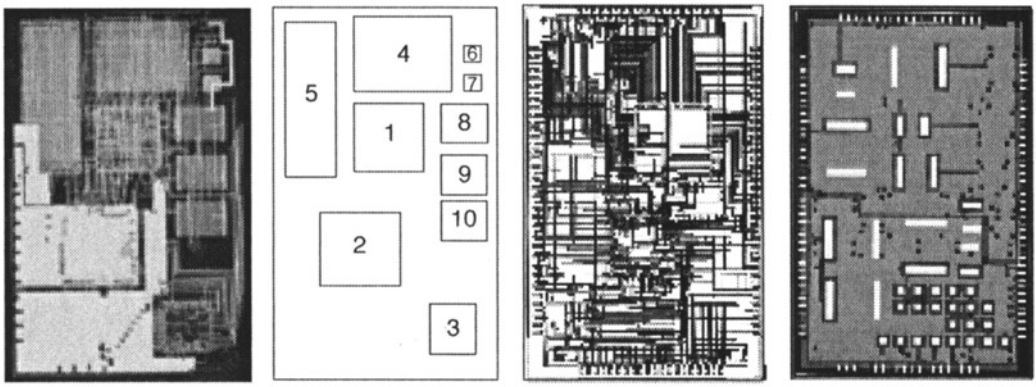


Fig. 8.43. AIC MCM. (a) AIC after first two levels of metallization; (b) legend: (1) CPU, (2) analog ASIC, (3) resistor ASIC, (4) 128 K × 8 EEPROM, (5) 128 K × 8 SRAM, (6–10) capacitors; (c) CAD files of some AIC metal and "drill" (via) layers; (d) AIC after HDI fabrication (components obscured by metallization and green solder mask coating).

8.4.1.2 Exploitation Techniques for Microinstrument Design

Just as a monolithic IC can be viewed as a synergistic arrangement of many individual transistors, microinstruments should be viewed as a synergistic arrangement of many individual components, not just electronics, but sensors and actuators. Sensors and actuators of an integrated microsystem correspond to the interface of physical domains, such as heat, light, sound,[119] radiation, and chemistry. This integration can be achieved monolithically or by hybridization, that is, the "magic of advanced packaging." Are aggressive miniaturization efforts always rewarding? It seems that in some cases physical size does scale and can accommodate miniaturization, but losses may not. The use of more sensitive physical effects may be required, or cleverer exploitations of designs pursued. Robustness is always desired, as it allows a greater diversity of operating conditions.

In this section, a few techniques that could be readily applied to constructing microinstruments are discussed. Constant-floor plan or quick-time reconfigurable designs allow more rapid mechanization of new concepts. Flex-based construction permits novel exploitation of flexible circuits to create 3D systems. Finally, some approaches for exploiting the high-performance nature of packaging to build better instruments are discussed.

8.4.1.2.1 Constant Floor Plan MCMs/Quick-time Reconfigurable Techniques

In a constant floor plan (CFP) MCM, preplaced components are arranged strategically in a substrate, but only partially interconnected, permitting an end user to complete the last stages of wiring and assembly for a particular application. The technique is particularly attractive in HDI processes, where a variety of components can be recessed within a substrate, leaving the entire module top surface available for adding components. The CFP MCM is a packaging analogy of a gate-array IC, where prefabricated transistor arrays are interconnected by completing the metallization system. The quick-time reconfigurable (QTR) approach is a variation on the CFP MCM in which a largely nonconnected MCM is programmed by adding a circuit board containing the final connection pattern and surface-mount components. In the QTR scheme, it is possible to exchange the circuit board quickly, creating a kind of plug-and-play system. QTR approaches are particularly convenient for complex sensors. The development of analog processing for a given sensor is necessarily custom because of the need to supply special timing, bias, amplification, and level-shifting networks for each variation. Furthermore, the resulting output signals vary in signal amplitude and format and have their own special timing relationships.

A conceptual example of a QTR system is shown in Fig. 8.44, involving a patterned overlay HDI module. Since patterned overlay MCMs feature planar surfaces, the introduction of a solder-bump array provides a compliant contact system for mounting a quickly customizable PWB. It is the introduction of the latter component that provides the QTR feature. Through custom patterning of commodity PWB technology (it is possible to fabricate them within 24 hours after transmittal of electronic design files), the final interconnection scheme is patterned, configuring bias, timing, filter, and amplification networks as needed for a given sensor. MEMS sensor and actuator arrays can also be implemented in a flexible and rapid manner. The use of semirigid flex PWBs also allows the integration of a final connector.

8.4.1.2.2 Flex-based Construction

In flexible circuitry, the "flatland" analogy of planar construction approaches can be, if not broken, at least bent. The example of a folded flexible MCM shown in Fig. 8.20(d) can be extended to many other concepts. For example, a three-axis inertial reference unit could be readily

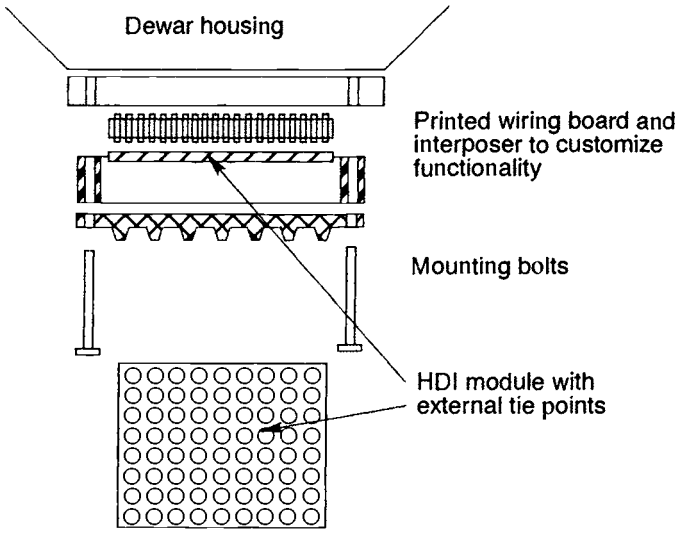


Fig. 8.44. QTR MCM.

constructed using a single-axis MEMS device built on three small MCMs, as shown in Fig. 8.45. The assembled MCM, fabricated as a continuous planar assembly, is readily converted into a three-axis system by folding the assembly like a box, around perhaps a "trueing" block, which could contain other electronics.

8.4.2 Highly Integrated Packaging and Processing

Traditional packaging hierarchy has, by comparison, seen little improvement because continual progress in feature size improvement of the IC takes pressure off striving for improvements. Most research in 3D packaging suffers from a lack of critical mass and applications pull, resulting in many impressive "hand-crafted" laboratory curios that lack acceptance. For the most part, 3D packaging research has centered around simple die stacks, with a considerable spread of substrate stacking approaches, most of which are implemented once with unclear benefit.

An AFRL initiative, referred to as the HIPP program, offers a fundamental reexamination of the packaging hierarchy and the successes and failures in 3D advanced packaging. In essence, it provides the implementation of a new hierarchy that complements the traditional one, but with important benefits to new technologies such as MCMs. The HIPP system offers a counterpoint to the inefficient frameworks in conventional packaging, which often cannot exploit the benefits of functions with multiple MCMs. Assemblies built in the HIPP approach can be merged into L2–L4 of the existing hierarchy whenever necessary and convenient. The HIPP framework, as an efficient MCM containment system, may not on the surface seem to achieve the impressive densities of laboratory-created 3D approaches. Spectacular gains of certain 3D approaches can be meaningless in complex system applications since those benefits cannot be accrued to the variety of component and circuit classes that make up an average system because of an intrinsic lack of "packaging services" needed in a complete system (e.g., thermal and electrical management). HIPP provides these services in a multitechnology framework, one that can allow any individual MCM to deliver the maximum benefit in most systems of reasonable complexity.

8.4.2.1 HIPP Assembly Structure

The HIPP program has sought packaging solutions more optimal at a system level, based on the concept of closely integrating a collection of various MCM substrates or other assemblies of identical size and conductor arrangement. Early artist concepts for these approaches are shown in Fig. 8.46. The first approach [Fig. 8.46(a)] involved the use of framed interposers and surface border interconnections, where contacts passed through entire substrates. Though efficient for bussed structures, this approach exacts a severe I/O penalty for complex substrate-to-substrate interconnections. For example, if the first substrate connects to the eighth with 80 signals, those 80 signals must be passed through substrates 2–7, whether or not those signals have a connection within

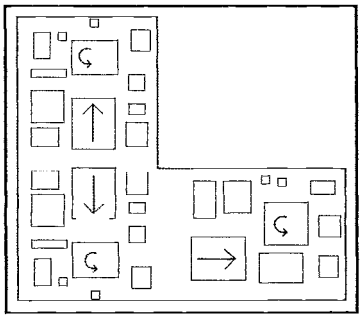


Fig. 8.45. Three-axis MCM built onto folded-flex HDI (shown before folding).

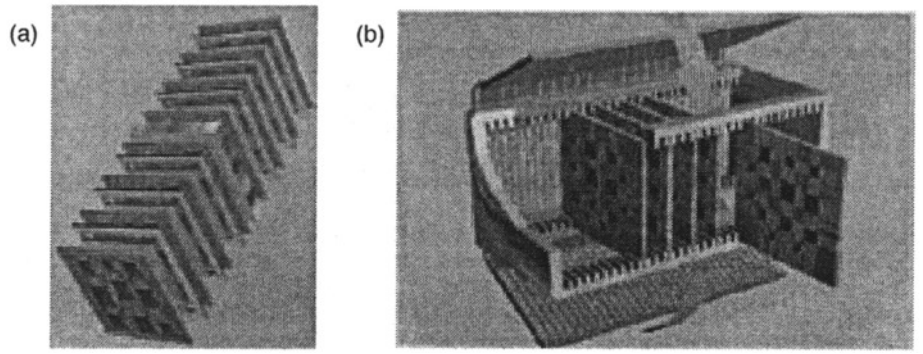


Fig. 8.46. Original HIPP concepts.

those substrates. At the random complexities represented by many heterogeneous systems, an unacceptably high amount of substrate real estate would be devoted to "pass-through" interconnect. The edge-interconnect system, shown in Fig. 8.46(b), provides an obviously more "agile" interconnection manifold, but requires tremendous edge-contact density on particular substrates, which could not be readily accommodated by the present state of the art (20–40 mils).

The requirements of candidate HIPP structures include:

- High I/O densities (up to 1000/layer)
- Heterogeneous signal composition (i.e., analog signals, digital, power, and microwave signals)
- Modularity and serviceability for layer repair and replacement
- Adequate power and thermal management
- Adequate I/O density at the second level of packaging
- Robustness for applications in harsh environments

Continued research led to a hybridization of the earlier conceptual approaches in which the theoretically high-contact densities of the first approach could be combined with the interconnection manifold agility of the second approach while addressing the basic requirements of a candidate heterogeneous 3D packaging system. Such a system, the baseline for a demonstration for the Discriminating Interceptor Technology Program (DITP), is shown in Fig. 8.47.

8.4.2.1.1 Segments

Figure 8.47 illustrates a many-layer 3D packaging approach that combines a number of segment entities into an assembly. The segments, which are the common and fundamental building blocks of HIPP, contain one or more MCMs or small circuit boards containing components. Systems, such as the DITP platform, can be partitioned into a number of segments, as shown in Fig. 8.47(a). In this case, the HIPP assembly baselined for DITP consists of approximately 16 segments (numbers reflect current order of segments from the front in the preliminary design):

- MSP (malleable signal processor) Subsystem 1
  - Sensor adaptation segment for passive focal-plane array sensor (2)
  - MSP 0.5 core segment (4)
  - MMGT (MSP management) segment, which contains MSP management processor and FI32 interface (5)
- MSP Subsystem 2
  - Ladar (light-based [laser] radar) adaptation segment containing nondigital interface (level shifters) (3)
  - MSP 0.5 core segment (6)
  - MMGT segment for second MSP core (7)

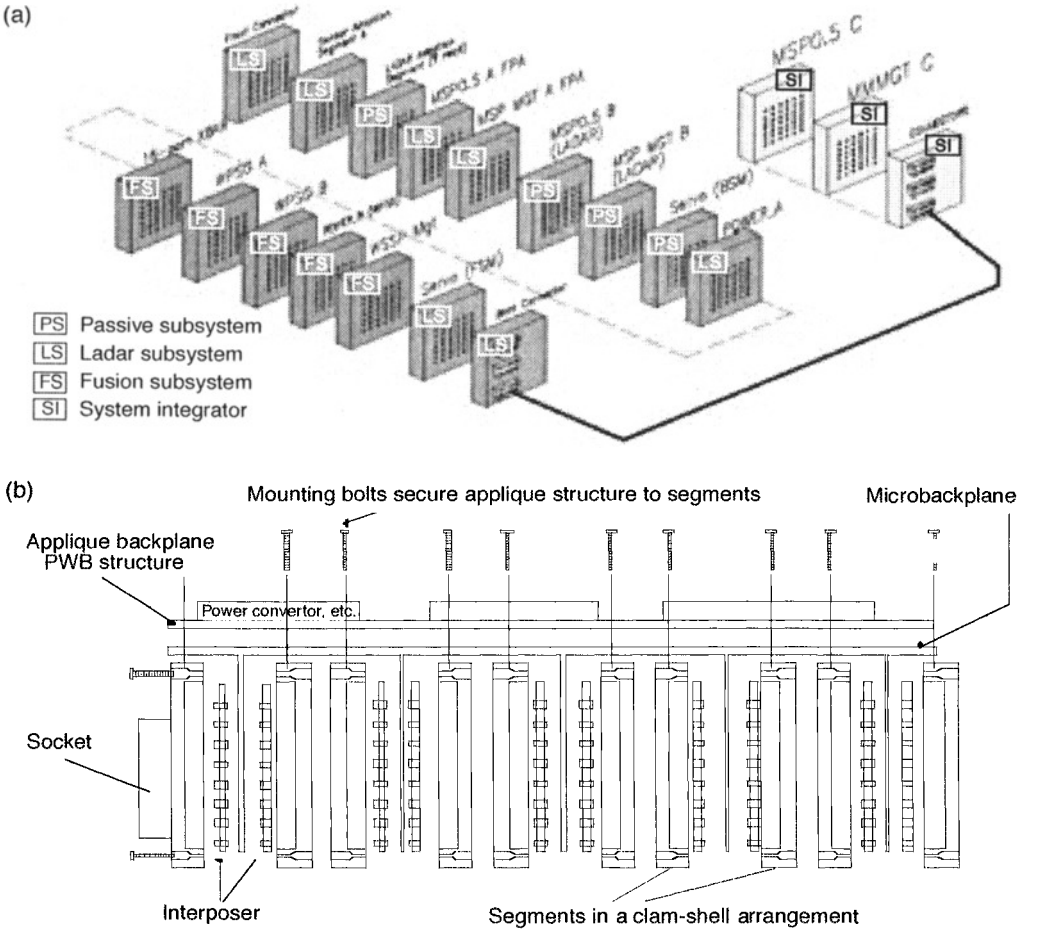


Fig. 8.47. HIPP baseline concept for DITP. (a) simplified physical representation (telemetry function external to sensor and fusion engine), (b) more detailed drawing illustrating interconnect features.

- Fusion Processing Subsystem
  - Myrinet crossbar (10)
  - Wafer Scale Signal Processor Segment (WPSG) No. 1 (11)
  - WPSG No. 2 (12)
  - Wafer Scale Signal Processor Management Processor (WMGT) (13)
- System/miscellaneous
  - Front (1) and back (16) connector segments
  - Power management layer A (9) for MSP subsystems and servo/guidance interface
  - Power management layer B (13) for Wafer Scale Signal Processor (WSSP) subsystems
  - Servo layer, which contains interfaces to communicate with servos, guidance (15)
  - Spare segment (8)

These 16 segments, referred to collectively as the Sensor and Fusion Engine (SAFE), have a common substrate size and I/O pad location. The physical size for substrates used in the DITP version of HIPP is 2 × 2 in., and up to 1000 I/Os can be accommodated on the surface of each segment.

8.4.2.1.2 Segment Layers

The contents of each segment can be completely different, and in fact the type of MCM technology used in each segment can be different so long as the segment definition is not violated. As such, layers do not necessarily need to be based on MCMs, but in fact could be single-chip packages, small PWBs, hybrids, or MCMs. In the terminology of HIPP, segments are said to contain one or more layers. Examples of possible layer arrangements within segments are shown in Fig. 8.48. Figure 8.48(a) illustrates a single-layer segment containing a single component mounted on a PWB. Figure 8.48(b) illustrates a single-layer segment containing one MCM. Figure 8.48(c) is an example of a segment with multiple layers, in this case two high-density interconnect MCMs. In principle, extremely dense MCMs could be used in multilayer arrangements, as shown here, to increase the volumetric densities of individual segments over that possible with a single-layer segment. Finally, Fig. 8.48(d) shows a single-layer segment in which a number of densely stacked 3D-chip configurations have been placed. In this manner, the HIPP packaging technology is versatile in that it can accommodate many existing modern forms of
single-chip, multichip, and 3D packaging.

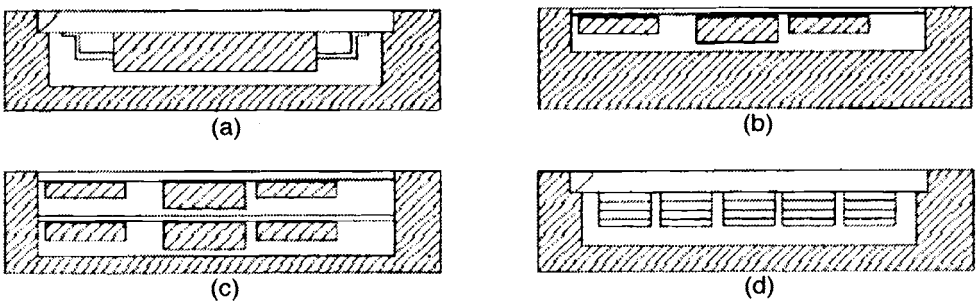


Fig. 8.48. Generic segment examples.

8.4.2.1.3 The Microbackplane and Assembly

As such, HIPP defines an efficient heterogeneous MCM-containment system. The deliberate stacking of segments is referred to as an assembly. Figure 8.47(a) gives a simplified, exploded picture of a HIPP assembly for clarity; Fig. 8.47(b) illustrates some of the special structures needed to integrate multiple segments into a complete assembly. These structures include the microbackplane, a number of interposers, an applique superstructure option for attaching more connectors or electronics, and hardware for secure segments to the microbackplane.

While each of the structures in Fig. 8.47(b) is essential, it is the microbackplane that uniquely defines the pattern of all interconnections in the DITP SAFE system (Fig. 8.49). The microbackplane is a compound flex system, based on a long manifold of multilayer copper-polyimide with orthogonal tabs of flex that address the face of every segment in the HIPP system. The microbackplane can be thought of as the "nervous system" of a HIPP assembly, and the design approach used for it combines the best elements of the two original HIPP concept designs shown in Fig. 8.46. Figure 8.47(b) illustrates the notion of clamshell mounting, a technique by which particular segments are mounted face-to-face through the microbackplane. Clamshell mounting allows a more intimate interconnection between two particular segments, which can serve to reduce complexity in the microbackplane.

Interposers form a springlike, compliant contact system, which mates the pinless conducting surfaces of segments to the tabs in the microbackplane. Such compliant inserts, which replace

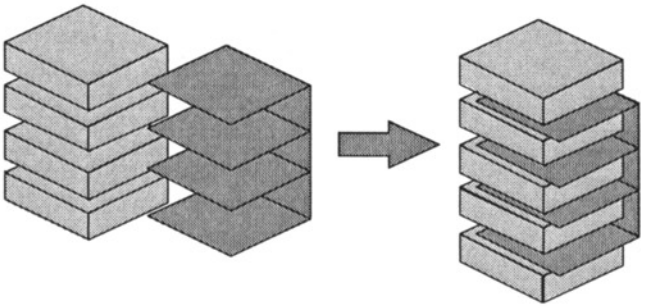


Fig. 8.49. Segment combination with microbackplane.

pins on ordinary packages, are necessary to ensure electrical continuity, exist between a large number of patterned conductors on flat surfaces. As shown in Fig. 8.50, an interposer is a compressible material that provides a conductive feedthrough, matching identical patterns on opposing surfaces. When the opposing surfaces are brought together, the interposer is compressed, forming electrical contact between the opposing surfaces. The compliance picks up any slack in irregularities between the otherwise flat surfaces to ensure a good contact. Since HIPP segments may contain up to 1000 I/Os, then so must the interposers. Since the pressure required to achieve the desired compliant travel range can be as high as 2 oz per contact, a significant amount of force could be required to tension the 16,000 total contacts possible in a DITP system. A "divide-and-conquer" approach is employed in the HIPP concept, involving localized tension of a smaller number of layers (usually one or two) to reduce the compressive requirements in the entire assembly. The concepts of segments, interposers, microbackplanes, and localized tensioning are illustrated in experimental assemblies shown in Fig. 8.51.

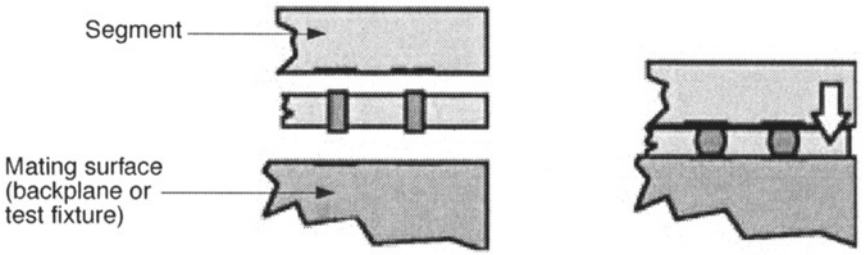


Fig. 8.50. Interposer detail.

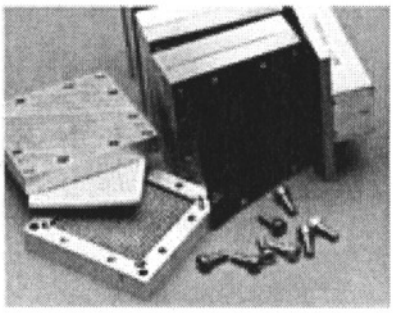


Fig. 8.51. Experimental HIPP assembly, exposing some segment details.

8.4.2.1.4 Power, Thermal, and Electrical Management

As a heterogeneous packaging system, the HIPP approach must deal with extremes in power management, thermal management, and signal integrity. Power delivery concepts under definition for HIPP are addressing the problem of delivery of 70–80 A of current at 3.3 V. Power delivery to systems with a large switched-signal content is a significant energy supply challenge. Also, the delivery of large amounts of current could require heavier metal structures than those available in a microbackplane. By contrast, only small amounts of current are required for analog sensors, but these current paths must have extremely low noise.

The thermal management in HIPP is hierarchical, based on first shuttling heat generated within each segment efficiently as possible to the outer edges of the segment walls and then coupling a second-level thermal management system. In the SAFE system, segment power dissipations range from about 1.5 W to 60 W per segment. Thermal transport in HIPP must occur laterally, parallel to the plane of the layers within the segments. This is because many HIPP assemblies are many-layer MCM assemblies, in which lateral transport is most important. Figure 8.52 illustrates the segment level thermal path. Segment thickness, segment wall thickness, and segment material selection can be based on local and global HIPP assembly needs. The second level of thermal management is application dependent, but must deal with power dissipation levels as high as 500 W for 16 segments. In the DITP SAFE assembly, as suggested in Fig. 8.52, lateral heat transport from segments into a phase change material is a second-level thermal management approach under consideration. For operation at longer intervals, a number of other options can be considered, ranging from heat sinks to heat pipes and liquid flow-through systems. Thermal management systems can more intimately link into segment walls, through texturing, flocking, insertion of flow-through channels, and other methods.

8.4.2.1.5 A New Packaging Hierarchy and Extensions of the HIPP Framework

HIPP establishes an alternate packaging hierarchy, one that is compatible with the existing hierarchy, but potentially much more efficient. L1 in the HIPP packaging hierarchy refers to internal layer composition, L2 is defined by layers within segments, L3 is defined by the segment itself, and L4 is defined as the HIPP assembly (of segments). Various forms of compatibility with the existing packaging hierarchy are readily achieved. For example, HIPP segments can be face-mounted onto PWBs through proper socketing or through conversion of the land grid array on the segment face to a ball grid array. Alternately, entire HIPP assemblies can be mounted onto a PWB, given the proper structural design. HIPP assemblies can be used to replace entire boxes; miniature connectors can be introduced on the front and back surfaces and onto the microbackplane itself.

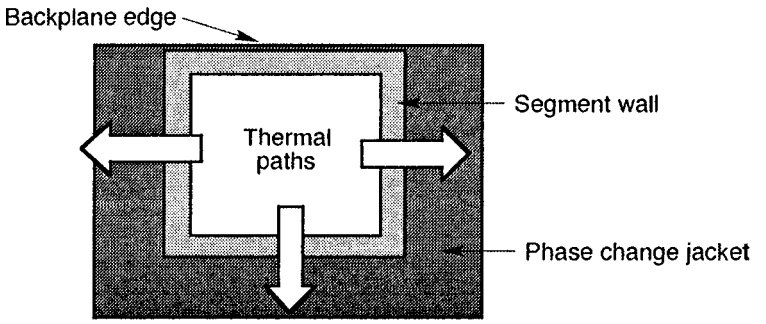


Fig. 8.52. Segment level thermal management.

HIPP offers a packaging system that meets the essential requirements of a 3D modular packaging system with high efficiency and flexibility. Segments are in essence interchangeable, reparable, and replaceable without great difficulty. In contrast to many 3D packaging approaches, HIPP can intermingle a great diversity of functional domains in electronics, providing in this way a great flexibility for system designers. Design guidelines under development will in time reduce to practice many aspects of the analysis needed to effectively use HIPP technology, such as electrical and thermal design rules and guidelines for exploiting CAD tools. It is believed that with these guidelines it will be possible to design HIPP-based systems with the ease and confidence one would use to design present-day VME or SEM-E (Appendix E of the military standard for standard electronic modules) board-based systems.

The HIPP-inspired packaging hierarchy may be further extended to include two additional levels: the cluster of assemblies (L5) and the cluster-stack (L6). Many potential concepts can be introduced to combine a number of HIPP assemblies into an efficient configuration that provides extremely high interassembly bandwidth and preserves good access of segment structures for thermal management. One such arrangement is shown in Fig. 8.53. Here, the cluster formation consists of a 2×2 arrangement of 16-segment assemblies [Fig. 8.53(a)]. A compound microbackplane [Fig. 8.53(b)] exploits short interassembly distances and affords a much higher interassembly bandwidth than nearly any other connector-based system, including present-day fiber optics. If the front and back of each assembly were fully exploited for I/O, the current HIPP standard would permit a signal transport of 8000 I/Os from the cluster. Two lateral edges of each assembly are available for thermal management, and the cluster could be completely encircled if necessary by an annular thermal management system. Figure 8.53(c) depicts a notional thermal management concept based on heat sinks.

Clusters of the nature described could be readily extended to form even more complex assemblies based on stacking clusters vertically to form cluster-stacks. The intercluster contacts could be accomplished with interposers, and system I/O and power delivery would be achieved through a connector system introduced at the top or bottom of the system. Such a stacked system has virtually no waste real estate; every part of the surface can be allocated to either thermal support or electrical transport (power and signal). In principle, the four-tier cluster-stack shown in Fig. 8.54 could aggregate well over one teraflop of processing within an approximately $5 \times 5 \times 4$ in. volume (exclusive of thermal management structures such as the heat sinks shown). This assertion is based on clusters of 16-segment assemblies, where each segment contains four layers of WSSP-based processing elements (each layer containing four WSSP components).

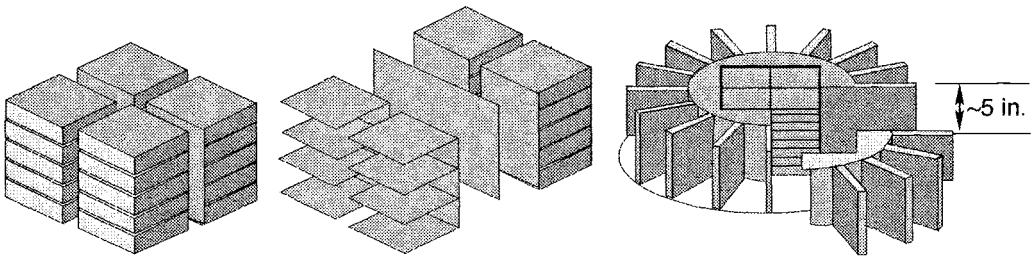


Fig. 8.53. Cluster arrangement of assemblies (extensions beyond present DITP embodiment).

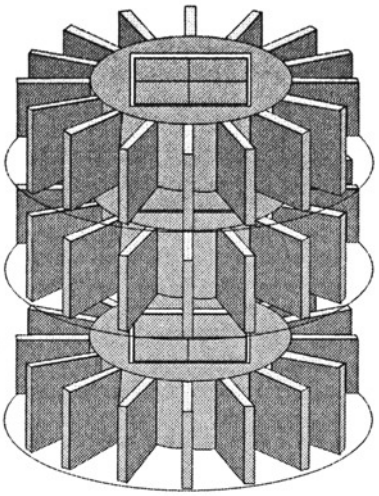


Fig. 8.54. Cluster-stack (teraflop concept).

8.4.2.1.6 Extensions of HIPP Through Improved Densities at L1

Beyond the extensions proposed for HIPP clusters and cluster-stacks, which provide a scheme for aggregating assemblies, the HIPP system can take full advantage of any new advances in microelectronics, MCMs, and 3D packaging technologies. HIPP can also take advantage of advances in connector, BGA, chip-scale packaging, and related technologies to increase segment I/O and system I/O densities. HIPP is then a framework for packaging, and new advances serve to accelerate further the density of this framework.

An example of such an accelerator is ultrahigh-density interconnect (UHDI). UHDI is centrally based around the idea of membrane-thin electronic MCMs. In the most aggressive form of this program, the electronic membranes are 0.002 in. thick, contain arbitrarily complex MCM circuits, and are flexible around structural contours or stackable into arbitrarily dense assemblies. The ideal UHDI system would use IC processes and designs optimized for this form of assembly, which could lead to significantly improved performance-to-power ratios (for example, 75% less power/MIP). In the simplest form, the first stepping-stone, existing ICs can be demonstrated functionally in substrateless versions of the HDI process (Fig. 8.55). Given that substrateless HDI is feasible, the remaining issues are solving crucial technical challenges in the four key UHDI research areas:

- Ultrathin semiconductor device processing
- Electronic membrane development and qualification
- 3D (membrane stacking) development
- Architectural optimization of 2D and 3D UHDI

Current AFRL-funding research is examining a number of these research areas. Recent experimentation on a limited scale has produced functional modules approximately 0.04 in. thick that contain four stacked substrates, compared with a normal single HDI module, which is normally 0.06 in. thick. This simple step alone quadruples the potential density of a HIPP segment based on HDI layer components. One potential application for the ultrathin UHDI technology is in congested MCM floor plans where planar arrangements of memory die can be replaced with compact tiles containing UHDI-based memory stacks. Such an approach can eliminate the space normally occupied by several components.

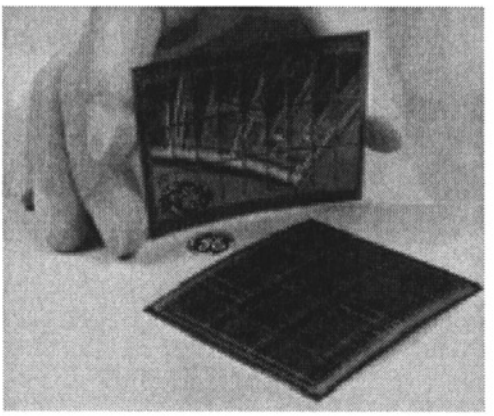


Fig. 8.55. Substrateless HDI technology.

8.4.3 A Third Generation of Advanced Packaging

If the ordinary hybrid microcircuits developed in the early days of space exploration can be considered a first generation of packaging, then surely the new emergent forms of MCMs in the late 1980s and early 1990s constitute a second generation. They differ not so much in role but rather in degree from their simpler precursors. While 3D approaches offer tremendous possibilities in extending the benefits of 2D MCMs in systems, they do not necessarily drive a change in the fundamental substrates. In fact, with approaches such as HIPP, first- and second-generation modules can be inserted with equal facility. Have we reached the end of improving the MCM, notwithstanding the obvious improvements gained, for example, by decreasing permittivity in the intermetal dielectric? Clearly not, given the trends in microcircuit technology. In fact, we believe it is possible to define a third generation of MCM technology, one that as before does not substantially differ in role but in degree. Coupled with the implied improvements in wiring density is the enablement of substantial improvements in 3D packaging by thinning. We introduce here the notion of hyperthinning in silicon, a domain in which silicon bends instead of breaking. The implications to conformal packaging, 3D systems, and integral passives are manifest. We believe the third generation of packaging will offer new challenges in design, both in complex homogeneous domains and mixed functional domains.

8.4.3.1 Density of Contacts and Wiring

Current MCM technologies are commonly limited to a 25–70-μm pitch. Technologies that can achieve a sub-10-μm pitch do so at the expense of performance, since in many such cases the dielectric and metal layers are thin, which gives rise to increased line resistance and capacitance. When the dielectric layers are not thin, then vias between two layers are often large to permit yieldable processes with good metal step coverages. What is needed in MCM approaches for the most advanced next-generation systems is a much greater line width for wiring density that has good electrical performance and permits high via and contact density. In this manner, both wiring and contact densities can be high enough to permit devices with many thousands of I/Os per cm^2 to be accommodated, consistent with the trends predicted by Rent's rule. The implications for patterned substrate designs are clear: only flip-chip approaches will suffice for high-density devices (devices that would force a perimeter wiring density below 35 to 40 μm). Such requirements are already met by advanced cryogenic hybrid detectors in which over 1 million contacts are made between a detector substrate (e.g., HgCdTe material) and a silicon readout IC at sub-40 μm using indium bumps.

It is important to note one impact of increased modular I/Os on the next level packaging systems, whether a 3D arrangement of MCMs or a tiling format onto boards. Precision alignment will be required at the next level of packaging. The alignment of modules at high density can be realized in two ways, using precision assembly approaches or using autonomous postassembly alignment, which we shall refer to as a two-phase connection approach. The former approach is tractable but highly undesirable for systems that would be fielded since servicing assemblies would require extremely specialized equipment (e.g., microscopes and micromanipulation/assembly equipment). In the case of a two-phase connector (shown in Fig. 8.56) an automatic alignment system affects a precision connection of two surfaces. The two-phase approach works by engaging coarse connection points under normal press fit connection tolerance (tens of mils). Then, after the first stage of connection, an active alignment system is activated that corrects for translation and angular misalignment. The precision sections of the assembly are then finally "docked." Such a two-phase concept could enable many thousands of I/Os between two surfaces to be efficiently interconnected.

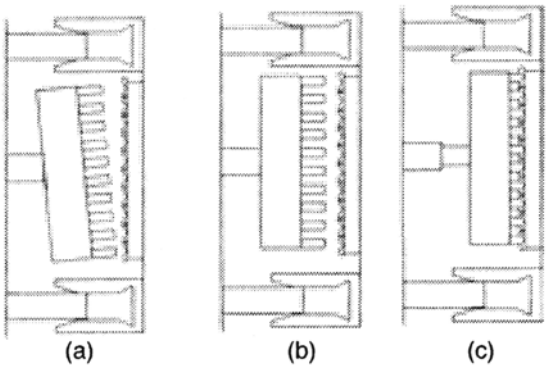


Fig. 8.56. Two-phase alignment sequence. (a) gross or coarse alignment after passive connection, (b) corrections made by active alignment system, (c) final connection made.

8.4.3.2 Hyperthinning and Its Implications

Typically semiconductor wafers are about 500 μm thick. For some packages, millions of devices are thinned in production to 0.007 in. In some 3D packaging approaches where thinning has occurred to 25–100 μm, silicon is observed to be very fragile and difficult to handle and process. At thinner extremes (< 25 μm), however, silicon becomes pliable (as shown in Fig. 8.57), completely changing the mechanical support/rigidity issues. This regime is referred to as hyperthin and creates a range of new possibilities in packaging.

The first obvious benefit is component density, particularly when patterned overlay technologies are used. With patterned substrate technologies, the height of the chip-attachment system (e.g., wire-bond loop height) will fundamentally limit density. In patterned overlay systems, the limiting factor is the height of the interconnect film, which is defined to be consistent with performance requirements. In other words, a 25–125 μm thickness in the patterned overlay sets the only limitation in how thin a hyperthin silicon MCM could be made. Layers this thin could be treated as electronic membranes, stacked like pieces of paper. If the layers could be permanently stacked, then it could be possible to laminate entire substrates together, forming connections in a manner analogous to interconnects between layers of the same MCM. Theoretically, especially for a system with low interconnect "intensity," such as a memory system, it is possible to form a

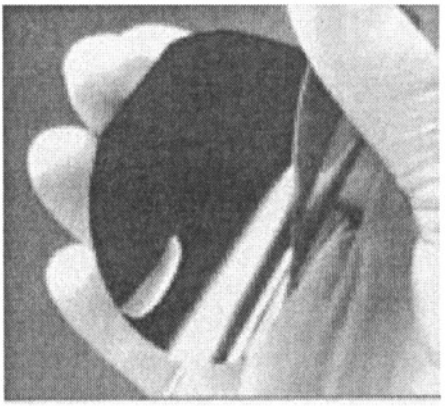


Fig. 8.57. Ultrathin silicon for specialty applications. (Courtesy Virginia Semiconductor.)

stack containing 500 MCMs/in., leading to a staggering 200 Gbits/cubic in. density based on 64 mbits/cm^2 in silicon memory technology

A second benefit of hyperthin silicon and overlay packaging is conformability, that is, the ability to bend substrates as though the flexible circuitry contained no components at all. As such, electronics could be formed like membranes and merged into structures of interest, planar and nonplanar, in ways never possible before. Computers could be wrapped around heat sinks, and distributed health monitoring systems "woven" into structures. The conformability raises interesting questions about the robustness of the packaging system. Does the silicon change its properties when bent? It has been shown that some insulators under pressure can become semiconducting, for example.[120]

A third speculated benefit of hyperthin silicon is improved radiation resistance. Since silicon substrates trap charge, which causes some radiation effects, by removing the substrate, radiation performance might be improved, which is a fundamental reason why silicon-on-insulator processes are of keen interest in radiation-hardened semiconductor research.

To be sure, the hyperthin silicon (or semiconductors in general) face significant challenges. In the case of dielectrically isolated silicon, a natural release layer may exist, permitting selective processing to remove the back of the die. Since most of the devices are built in bulk silicon, however, the hyperthinned approach is of little appeal unless a general thinning procedure can be found. Second, handling is a concern. If devices are thinned and then transferred to assemblies, the possibility of damage is great, and such an approach is likely to be very costly. One hope is that in using patterned overlay approaches, the die could be thinned *en masse* after they are bonded to the interconnection system. For HDI in particular, the ability to create substrateless MCMs (Fig. 8.55) could be particularly convenient in developing a system for creating hyperthin components without extensive special handling of individual components. Another concern in creating hyperthin MCMs is creating a compatible process for introducing capacitors and resistors. Some of the integral passive research may provide a solution, since discrete components do not necessarily lend themselves to thinning without destroying the components or their salient properties.

8.4.3.3 Next-Generation Design Approaches

The third generation of MCMs would place an increasing burden on the CAD infrastructure. Greater densities of interconnects and commingling of MCM and chip interconnect can create much more complicated design trade spaces. Exploiting fully the ultrathin systems when stacked

may promote specialty architectures, such as those explored in 3D monolithic wafer-scale integration.[121] In this approach, many parallel computers were partitioned into several subsections, each implemented on a different layer. Complete computers were formed only when all layers were stacked and interconnected. Finally, the ability to commingle functional domains remains a greater potential challenge in third-generation systems, given the significantly more complex signal integrity environment.

8.4.4 Advanced Multifunctional Structures

A sensible extension of packaging involves treating the structure of a system itself as part of the packaging solution. MFS refers to the general idea that structural members can serve other functions besides providing support. For example, a structure could be extended to serve the role of thermal management, signal and power distribution and/or generation, fluidic, and fiber-optic routing. While some examples of MFS can be found in ordinary life, no standard discipline exists for engineering MFS into systems.

Endowing structures with nontraditional functions is not an easy task. It is desirable to create a concept for MFS that promotes generality rather than inhibits it. Such generality promotes widespread use, which in time could lower cost through the economies of scale.

If MFS panels are to become the "LEGO™" blocks of future systems, then it is necessary to establish principles that permit a usable variation across a class of solutions for the things that MFS would functionally eliminate. For example, if an MFS panel were to replace power and signal distribution cables and harnesses, it would be necessary to either standardize the arrangement of signals on all panels, introduce concepts that allow changes to occur easily (e.g., reconfigurable interconnects), or both. The promise of MFS is realized when the technology can be shown to enable more rapid assembly of panels and more rapid configuration/reconfiguration of flight bus and payload systems. These benefits complement the potentially substantial reductions in size, weight, and power that could accrue when functions are handled by one set of structures as opposed to several.

What possibilities then exist for an MFS? A partial listing includes:

- Electrical power generation (embedded batteries and solar panels)
- Electrical power distribution (through high-power adaptations of flexible circuit technology)
- Electrical signal distribution (through multilayer flexible circuit technology)
- Microwave signal distribution (through low-loss, impedance control transmission line structures)
- Antennas (through surface-emitting dipole, microstrip patch antennas)
- Optical signal distribution (through embedded fiber-optic conduits/polymeric waveguides)
- Thermal management (through controllable heatpipelike structures within the MFS panel)
- Fluidic distribution (through standard distribution of embedded channels)
- Vibration control (embedded and distribution sensor, actuator, and control system)
- Distributed health and status monitoring (arrays of sensors in standard locations of each panel with built-in controllers and nonvolatile memory)

In each case, some concept of standardization and/or reconfiguration of channels is implied. Furthermore, in each case, a concept of socketing is implied. For effectiveness, socketing should be repeatable as necessary to permit rapid disassembly/reassembly.

8.4.4.1 Types of Electronics in MFS Systems

A possible organization of electronics subsystems in MFS-based spacecraft would divide the platform into bus-electronics and payload-electronics systems. Such a concept is illustrated in Fig. 8.58. In this approach, the bus electronics would be as invisible as possible, seemingly woven into

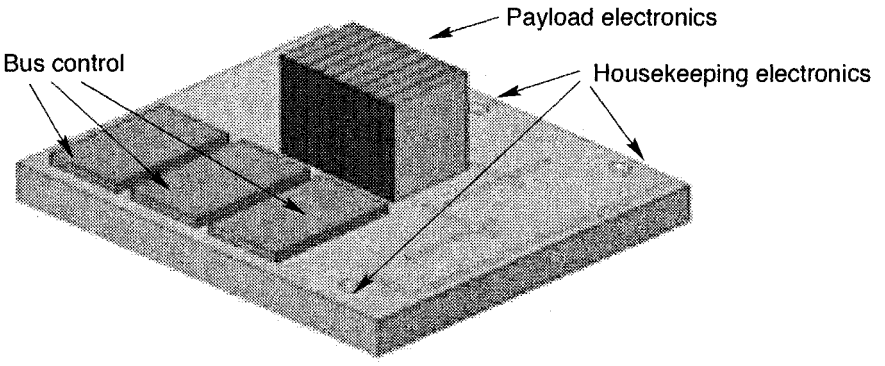


Fig. 8.58. Conceptual multifunctional structure panel.

the structure and putting little restriction on the placement of other functions. The bus electronics are broken into bus control electronics and distributed housekeeping electronics. The bus control electronics would deal with standard satellite control electronics such as a host processor, satellite communications, attitude control. The distributed housekeeping electronics, represented in Fig. 8.58 by a small low-power processor such as an AIC, would perform panel-specific sensor monitoring (e.g., pressure, temperature, dosimetry) and configuration management. It is envisioned that several AICs housed on each panel would perform these functions, working in concert with other panels as a self-organizing network, robust and tolerant of temporary or permanent failures of one or more nodes. The distributed housekeeping network would itself link to the spacecraft host processor through MFS interconnects within the panel.

Payload electronics are application-specific and introduced at specific docking points within the panel (shown in the center of Fig. 8.58). Here advanced packaging concepts such as HIPP can be used to implement, for example, a high-performance computation system or an integrated sensor payload. Specific mounting concepts must be developed to take full advantage of both the HIPP and MFS technologies, and intriguing possibilities exist, summarized in Fig. 8.59. For example, rather than employing a standard bolting arrangement [Fig. 8.59(a)], as one might do for a larger electronics box, it might be possible to employ a click-in-place system [Fig 8.59(c)] for a simpler attachment procedure to the MFS panel. The latter concept is particularly intriguing, as it allows for the possibility of in-situ placement/replacement of payload electronic assemblies. If the payload mounting areas are external to an MFS spacecraft, then the payload electronics can be serviced in orbit, without human intervention, creating the basis for a concept.

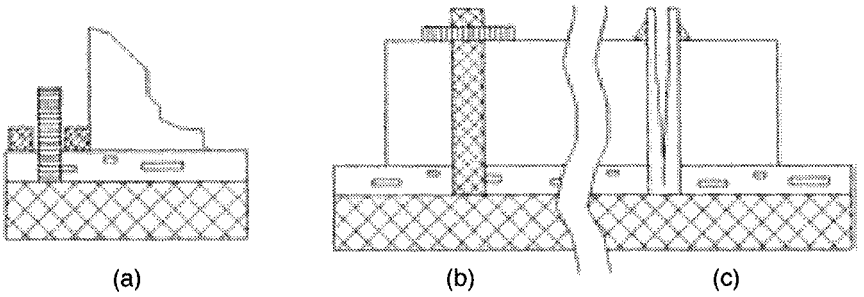


Fig. 8.59. Mounting systems. (a) bolt and bracket, (b) plate and cotter pin, (c) click-in-place assembly can be performed in flight.

8.4.4.1.1 Space Logistics Involving MFS and HIPP Technologies.

An overall, "hot-pluggable" electronics integration concept for in-orbit replacement of electronics on MFS panels is depicted in Fig. 8.60. Groups of layers or segments containing MCMs are combined with the necessary microbackplanes, shields, and protective layers to form an integrated HIPP assembly [Fig. 8.60(a–d)]. To effect a more intimate thermal management system, orthogonal fins protruding from the spacecraft panel could serve the dual purpose of improving thermal transport as well as providing a mechanical locking mechanism, similar to that shown in Fig. 8.60(c). The HIPP assembly is then integrated onto the spacecraft on the ground or in orbit [Fig. 8.60(e–f)]. To effect such a space logistics concept, improvements in connector technologies will be needed. Adaptation of existing interposer concepts could create an approach that automatically connects the payload to the spacecraft as the docking procedure is completed [Fig. 8.60(f)].

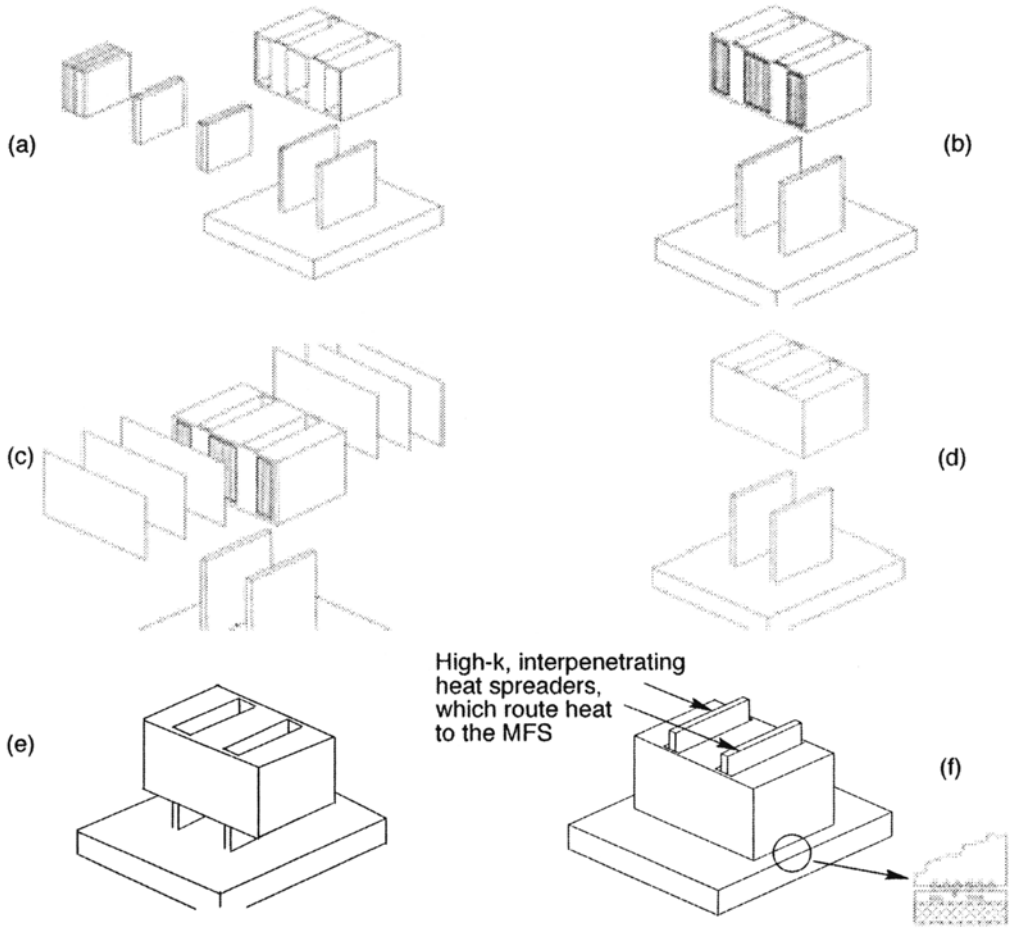


Fig. 8.60. MFS in-orbit "hot-plug" concept for space logistics. (a) HIPP assembly, exploded view, before integration; (b) HIPP assembly with integrated electronics; (c) addition of microbackplane, shielding, and protective layers; (d) a completed HIPP assembly, representing payload electronics for example, ready for integration onto panel; (e) partially integrated assembly; (f) fully integrated assembly, showing detail of underside flush-mount interposer system that electrically connects payload to MFS interconnection manifold formed on panel surface.

8.4.4.2 MFS Interconnection Manifolds

For purposes of affordability, the need to standardize MFS interconnection structures is manifest. Yet as previously discussed, the desired properties of interconnections are domain-specific. Hence, it is not possible to establish a "one size fits all" signal and power distribution scheme. As such, some regions of the MFS interconnection manifold might be optimized for power, some for digital, and still others for microwave and for analog.

8.4.4.2.1 Power Interconnection and Distribution within an MFS Panel

Power interconnections, which require low-resistance and sometimes high-current conductors, could be routed along gridlines, with branches from the grid to service particular points on the MFS panel corresponding to areas where bus electronics would be attached, mounting points for payloads. Figure 8.61 depicts a simplified grid. The metallization for the power grid would be robust enough for high-amperage power distribution, much heavier than needed for any other signal type. On some panels, where no special high-power electronics are mounted, the grid would simply "pass through" power to another panel. Each of the power "bars" would in fact be a collection of power conductors that can be separated for multiple voltage connections. Unused conductors on the power bars could be shorted together for lower loss power delivery. Discrete and specific regions of each panel could serve as tapping points for the various power lines passing through a panel. These regions, referred to as power service points (PSP), provide convenient supply points for electronics introduced *post facto*.

Power routing of the interconnections could be accomplished with three distinct methods. The first method, *a priori* routing, is reserved for the bus electronics, which are to be an intrinsic part of the panel. This type of power conductor routing could represent a deviation from the standard grid scheme, especially for distributed health and status monitoring networks, which are typically fairly low-power functions needed for less robust interconnection. The other two power routing methods rely on a number of power switch routing nodes (PSRN). PSRNs, in this concept, contain a number of links that allow "horizontal" power conductors to electrically connect to either "vertical" conductors (to map a voltage source to a load) or to another horizontal conductor (to increase current-handling capability of a particular voltage). In the second routing method, integration-point routing, the interconnecting links in PSRNs are "hard-wired," that is, prefabricated and

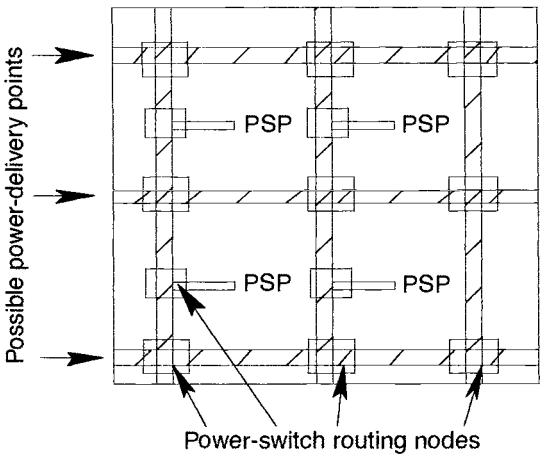


Fig. 8.61. Conceptual power routing manifold for an MFS panel.

installed during spacecraft integration. While this method preserves flexibility of MFS power routing until fabrication, it does not permit any flexibility after integration is completed.

This limitation is overcome by the third method, run-time routing, in which link connections are programmable within the panel and are reconfigurable even after integration and theoretically after launch. Configuration information on links for a panel can be stored in nonvolatile form within housekeeping electronics resident in the panel. At worst, the link elements that would be required would have to be nonvolatile and bistable themselves, and would necessarily have to preclude the possibility of glitching during spacecraft operation. The most significant barrier in the run-time routing concept, which enables plug-and-play spacecraft, is finding a bistable link-formation mechanism. Solid-state switches cannot be employed for the mechanism because such switches need to be at a particular threshold voltage with respect to power rails. As the rails are dynamic, it is impossible to guarantee that all switches can maintain the proper electrical relationship. Furthermore, power switches are subject to radiation effects, which can serve to increase on resistance. One technology not subject to this limitation is MEMS-based relays. In AFRL research, concepts for bistable MEMS relays (Fig. 8.62) suitable for such applications have been defined for inclusion in monolithic arrays and even within interconnections of an MCM. The number of such relays required is $(n^2-n)/2$ for nonblocking permutations of n different conductors, which is beyond reach of any non-MEMS relay technology. Fortunately, with MEMS-based switch approaches, it is possible to place a very high density of these switches within the space required of a PSRN.

Given the trends for continuing decreases in voltage for microelectronics, other issues remain for power distribution within a spacecraft, which makes the effects of I^2R losses caused by conductor resistance increasingly pronounced. Localized (point-of-load) power distribution will be realistically required at or below 3.3 V, even in small spacecraft. As such, some consideration is required for inclusion of such power converters directly within the MFS panels. Currently such converters are selected based on somewhat rigid specifications of input voltage, output voltage, and load conditions. With breaking improvements in radiation-hardened electronics, MEMS switches, integrated magnetics, and power-converter topologies, "smart power" is possible. Such smart-power converters could be reconfigured in-system, perhaps dynamically, to adjust for variations in voltage and load, considerably improve robustness and conversion efficiency.

In considering the issue of power distribution, it may make sense to also consider power generation/storage within the panel. In traditional spacecraft, power-combining electronics operate centrally and distribute power throughout the spacecraft. Other schemes based on a distributed power generation scheme should also be possible, eliminating a large amount of electronics devoted to centralized power management. The advent of the thin-film battery and solar-power technology might make possible the creation of panels that literally "carry their own weight" with respect to power generation, storage, distribution, and dissipation.

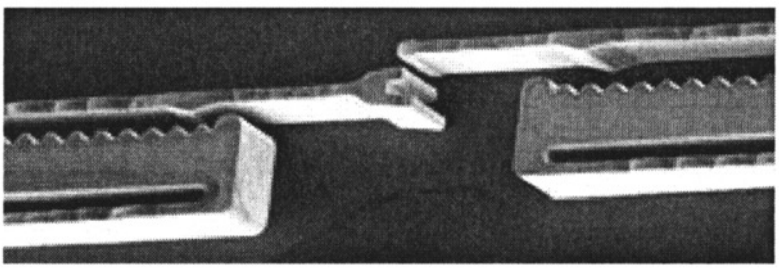


Fig. 8.62. LIGA-based bistable relay (courtesy Maj. John Comtois, AFRL).

8.4.4.2.2 Digital Interconnections within an MFS Panel

Digital connections within an MFS panel correspond to digital discretes, busses, and other general-purpose wiring. The digital interconnections are more tolerant of series loss and isolation in comparison to power interconnections, but unfortunately occur at a much higher density. It is envisioned that many hundreds of I/Os could be required for an average panel. As such, it is necessary to consider large groups of wires routed as in Fig. 8.61. In some cases, this can be done "virtually," that is, with FPGA or field-programmable interconnect devices. This approach assumes relatively low-signal frequency content, which is not true for many high-speed bus structures. When the interconnection distance exceeds a lumped element distance, transmission lines exist in the interconnect, requiring proper terminations. This situation is compounded considerably by the many naturally occurring discontinuities in the interconnection manifold. Once again, the use of MEMS switching devices could combat this complexity by effectively eliminating some of the transmission line stubs that would otherwise exist in the interconnection manifold.

8.4.4.2.3 Analog Interconnections within an MFS Panel

Analog signal types are classified as: low-frequency instrument (<1 MHz), high-frequency instrument (>20 MHz), and power analog. Interconnections that bear instrument-class analog signals have strict signal integrity requirements, chiefly in series attenuation and isolation from crosstalk. It is possible to form sophisticated shielded quasi-coax interconnections. And the quantity of such interconnections is higher than that required for power delivery, but not as high as for the complex digital portions of typical systems circuitry. Distributed, embedded analog signal capture nodes, which can be formed with a network of AICs laminated into the panel structure, greatly simplify the management of many dozens or hundreds of signal monitoring points. Such a network would localize AICs physically near the points where analog signals are generated, alleviating the need to protect long-distance analog signals from loss and crosstalk. Higher-performance analog measurement needs can be accommodated through higher-performance capture modules similar to AICs but with higher bandwidth. Here, transmission line effects could be problematic. Finally, power analog signals, associated with motor drive, require significantly larger conductor cross sections to minimize power delivery loss.

8.4.4.2.4 Microwave Interconnections within an MFS Panel

Microwave interconnections pose one of the greatest challenges in an MFS system, as microwave signals are subject to most of the previous concerns in signal integrity except for high wiring density. Impedance control for a fixed interconnect arrangement is difficult without the desired adaptive properties needed for a plug-and-play spacecraft. Finding switch configurations with low-loss and high bandwidth is not trivial, nor is designing the associated transition configurations to appropriate interconnection manifolds. Here, interconnections can generically be established in several arrangements for planar media (e.g., stripline, microstrip, coplanar waveguide, coplanar strip, and slotline). Figure 8.63 illustrates an evaluation approach for prospective MEMS switches that might operate at microwave frequencies.

If MEMS switches and flexible interconnection manifolds can be shown adequate for microwave applications, then some novel concepts might be exploited to more readily establish an adaptive system in which several interconnection paths could be formed for different applications using the same panel. It is possible, for example, that MEMS devices could create a tunable impedance manifold. Several concepts for this include the use of a transmission line in which a series of MEMS louvers adjust path impedances and implement MEMS-tunable stubs. The former concept, illustrated in Fig. 8.64, can tune the impedance of a transmission line with a dielectric of permittivity ε_1 by using small pieces of material attached to MEMS levels with higher

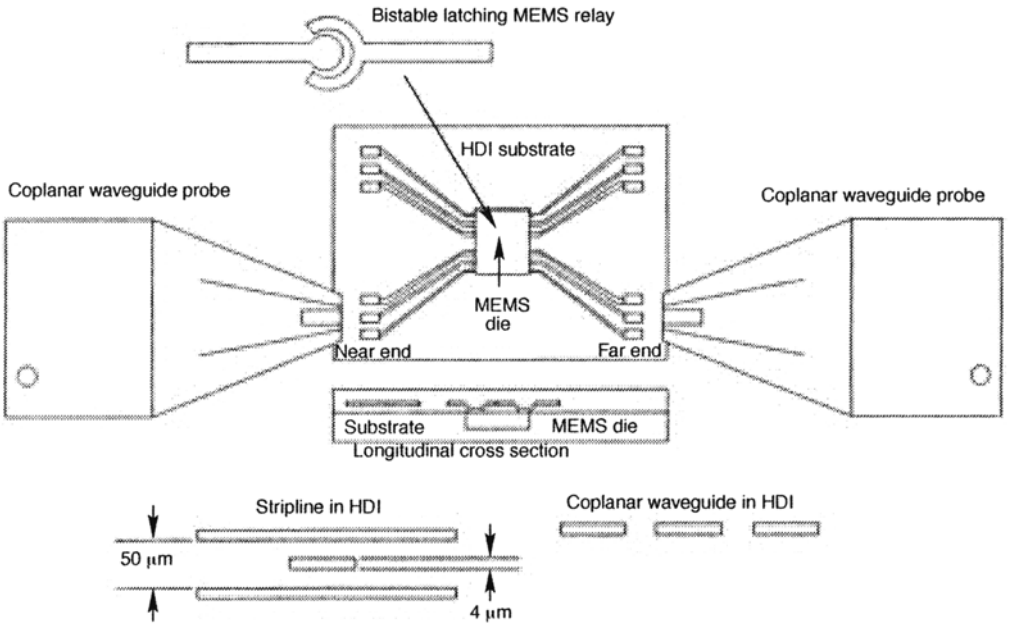


Fig. 8.63. MEMS switch test configuration.

permittivity (ε_2). Adjusting the positions of the MEMS louvers can alter the localized effective permittivity of the transmission line, permitting fine-scale impedance control. A second MEMS concept, that of tunable stubs, can employ relays similar to those shown in Fig. 8.61 to physically add or take away conductors and circuits, permitting some potentially useful adjustments in the microwave sections of a design.

8.4.4.3 Panel-to-Panel Attachment/Connection

The ability to attach multiple panels together and connect the various MFS features properly is an obvious requirement. Let us introduce the concept for a generic MFS "tile" of the right dimension to accommodate a wide range of possible satellite types and functions. The tile should have:

- Good structural characteristics
- Ability to form larger structures through tiling
- Ability to support assembly with more than one attachment angle to address vertices
- Ability to be stacked densely in a stowed configuration for mass deployment and in-space assembly

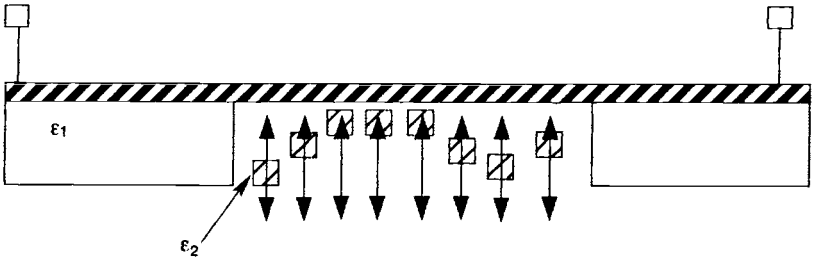


Fig. 8.64. MEMS-tunable transmission line.

A notional example of such an MFS tile is shown in Fig. 8.65. The particular geometry need not be hexagonal as chosen here, but should be chosen to permit flexibility in possible arrangements. The tile can be attached from any of the six edges shown; one mounting post or threaded insert is shown in the center. Larger planar arrangements are readily formed through juxtaposition as shown in Fig. 65(b), and payloads attached on a standard grid, as shown in Fig. 65(c).

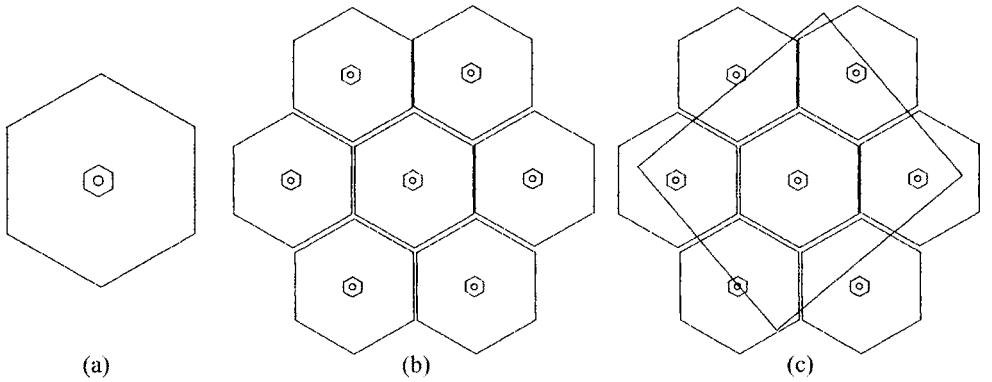


Fig. 8.65. MFS panel connections. (a) fundamental MFS tile, (b) tiles arranged to form larger planar structure with central payload attach point, (c) application of payload assembly that spans a number of the payload mounting points.

In this concept of panel-to-panel attachment, it would be important to support the ability to mount at certain angles to form boxlike and other polygonal shapes from which a system could be built. As shown crudely in Fig. 8.66, such a system would allow both planar and nonplanar engagement angles. Novel concepts for the joining process itself are suggested from other MEMS and non-MEMS sources. Large, coarse grids of panel-to-panel conductors correlate well with ordinary mechanical fitting tolerance and are readily accommodated by plug-and-socket arrangements. Dense conductor clusters, some optoelectronic and fluidic, and other connections will require greater precision in assembly, suggestive of a two-phase connection system as depicted in Fig. 8.57. The structural requirement for panel-to-panel attach will require good mechanical contact at several engagement angles. Perhaps this requirement can be addressed by special detented mechanisms involving shape memory actuation, which could be employed for the primary mechanical connection system. For ground assembly, more conventional techniques could be employed, but the big advantages are joining mechanisms in space for plug-and-play spacecraft.

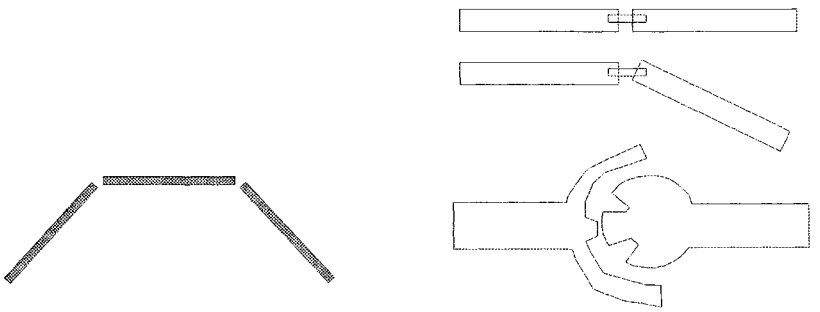


Fig. 8.66. Depiction of nonplanar engagement of panel and suggestion of detented attachment connection.

8.4.4.3.1 Extension to Space Logistics

If the several aforementioned elements of MFS design can be applied, then it is possible to consider satellites that can be serviced in orbit. Such a hypothetical repair operation is posed in Fig. 8.67. In this case, the faulty MFS panel [shown in Fig. 8.67(a) as a non-colored tile] is identified, and its edge connections are disengaged, permitting removal [Fig. 8.67(b)]. A known good tile is used for replacement and is inserted in the location previously occupied by the faulty tile [Fig. 8.67(c–d)]. Though a number of practical issues must be addressed to implement this space logistics concept, a number of powerful advantages clearly exist, such as spacecraft reuse (good tiles recovered from decommissioned spacecraft).

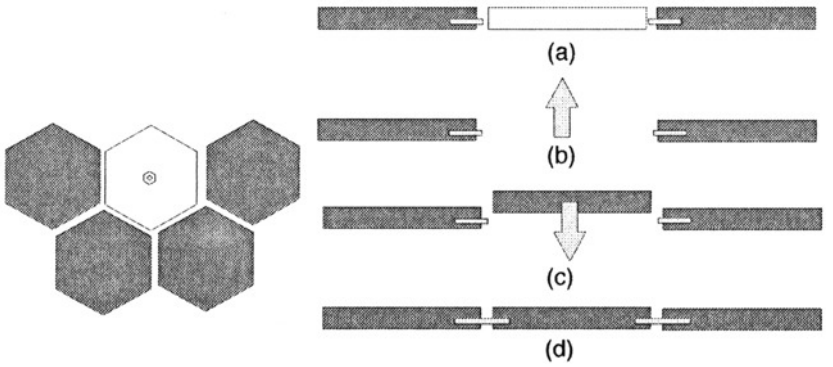


Fig. 8.67. Space logistics concept. Plan view (left), longitudinal view illustrating sequence (right). (a) Defective panel noncolored, (b) released from spacecraft, (c) new tile panel added, (d) completed replacement.

8.5 Conclusions

This chapter addresses the technologies of advanced electronics packaging. Packaging is primarily enabling to the extent that it allows the outside world to access functions contained within packages, recursively defined through a hierarchy that spans from the transistor to the system platform. Specific packaging concepts, especially those involving 2D and 3D MCMs, are discussed here as a practical introduction to the state of the art in electronics packaging. A review of the principles of package engineering includes discussion of the drivers for the materials and geometries of packages, substrates, and their respective configurations within systems, as well as the system design philosophy in which packaging is integrally considered. We present a number of advanced techniques and case studies in packaging and what we believe is an enabling heterogeneous 3D packaging concept, one that deliberately exploits the most promising microcircuit, MCM, and 3D packaging. This framework extends to the future of packaging—a "third generation" of packaging. Finally, in MFS, we consider both the challenge and the enabling benefits of approaches that can lead to LEGO™-like spacecraft, in which both ground and space assembly concepts are possible.

Packaging is the wrapper and mapper of systems at various levels. It is useful, in some respects, not for what it does but for what it does *not* do, such as *not* delaying or distorting signals, *not* restricting thermal transport, *not* contaminating sensors. By the same token, packaging is an enabler, as it enables a system to access the capabilities of disparate components. The goal of advanced packaging is to do this very well, even to the theoretical limits of what components can deliver.

8.6 References

1. M. J. Little and C. T. Moberly, *Signal Processing Systems Packaging-1*, Rome Laboratory Technical Report no. RL-TR-92-95 (April 1992).
2. R. Iscoff, "Wire Bonders—Stretched to the Max?" *Semiconductor Int.* **54** (2), 52–4, 56 (February 1993).
3. J. D'Ignazio, "Wirebonding's Reign Continues," *Semiconductor Int.* **19** (6), (June 1996).
4. R. Bidin, "High Pin Count Wirebonding: The Challenge for Packaging," *Solid State Technol.* **35** (5) 75–7 (May 1992).
5. Assembly Products, 97 East Brokaw Road, Suite 100, San Jose, California 95112-4209.
6. D. Maliniak, "MCMs Traverse the Cost Curve," *Electron. Design* **43** (13) (June 26, 1995).
7. J. Vardaman, president, TechSearch International, Austin, Texas (private communication, 1996).
8. "There's Magic in That Chip," *Electron. Products* (April 1995).
9. Harris Corp. literature on the Digital Drop Tuner (1995).
10. R. F. David, "Manufacturing Power Hybrid Circuits," *Electron. Packaging and Production* **36** (3) (March 1996).
11. "Module Provides Upgrade Path for Pentium Laptops," *Computer Design* (April 1997).
12. MicroModule Systems: http://www.mms.com/products/geminidoc/web_r20.htm.
13. nChip Corp., now Flextronics International Ltd., 2241 Lundy Ave., San Jose, CA 95131-1822.
14. *Microwave HDI Design Guide* (GE Corporate R&D Center, Schenectady, NY, 1995).
15. M. McNulty, *et al.*, "Microwave MCMs Using Low-Cost Microwave Chip-on-Flex Packaging Technology," *High Density Interconnect* **1** (1), May 1998.
16. *Chip on Flex High Density Interconnect Design Rules* (Revision 3) (GE Corporate R&D Center, Schenectady, NY, July 1997).
17. J. C. Lyke, "Two- and Three-dimensional High Performance, Patterned Overlay Multi-chip Module Technology," *Proceedings of NASA Technology 2002:Third National Technology Transfer Conference* Vol. 1 (December 1992), pp. 195–204.
18. J. D. Reed, *et al.*, "High Frequency IC to IC Signaling on Rapidly Prototyped Flip Chip MCM-D Substrate," *Proceedings of the International Conference and Exhibition on Multichip Modules and High Density Packaging* (Denver, 15 April 1998).
19. QTAI, MCC: http://www.mcc.com/projects/fmm/.
20. R. Iscoff, "Wafer Scale Integration: An Appraisal," *Semiconductor Int.* **7** (9), 62–65 (September 1984).
21. J. Banker, "Reducing Cycle Time Through Programmable Multichip Modules," http://www.pico-sys.com/red_cycl.htm.
22. "Advanced Packaging Gives SMT New Life," *Electron. Eng. Times* (6 September 1996).
23. S. Berry, "Reaching HDI Comfort Levels," *High Density Interconnect* **1** (1), 14–16 (1998)
24. "Component Packaging Technologies Enhance PCB Performance," *Electron. Packaging and Production* (January 1998).
25. A. Bindra, "TI Eyes New Package Type," *Electron. Eng. Times* (3 June 1996).
26. D. Strassberg, "More Pins and Less Space Beget New IC Packaging," *EDN* (25 May 1996).
27. "BGA Update," *SMT* (April 1998).
28. "Motorola Gate Arrays Span 12 K to 278 K Gates," *Electron. Eng. Times Product File.*
29. Y. H. Pao, *et al.*, "BGAs in Automotive Applications," *SMT* (January 1998).
30. P. Mescher, "The Evolution of BGA," *Advanced Packaging* (March 1996).
31. M. S. Cole and T. Caulfield, "Ball Grid Array Packaging," *EDN Products Edition* (15 August 1994).
32. T. LaMarche, "X Ray Closes the Inspection Loop," *Evaluation Eng.* (September 1993).
33. E. Williams and D. Duschl, "Effective Rework with Convective Tools," *SMT* (May 1998).
34. T. Costlow, "LSI Rolls Enhanced BGA," *Electron. Eng. Times* (6 March 1995).
35. "More BGA Choices Emerge," *Electron. Products* (April 1995).
36. J. Lipman, "New IC Packages Really Pack in the Leads," *EDN Europe* (1 September 1997).

37. R. DeJule, "High Pincount Packaging," *Semiconductor Int.* **20** (8), 139–140, 142, 144, 146 (July 1997).
38. R. A. Munroe, "Ball Grid Array Technology," *EDN Products Edition* (8 August 1997).
39. "Dimpled Ball Grid Arrays," *Semiconductor Int.* (June 1997). See also http://www.kyocera.com/kai/dbga.html.
40. R. D. Schueller, "Portable CSP," *Advanced Packaging* **7** (4), 28–30, 32, 34 (May 1998).
41. A. Seung-Ho and Y-S. Kwon, "Popcorn Phenomenon in a Ball Grid Array Package," *IEEE CPMT—B* **18** (3) (August 1995).
42. R. Ghaffarian, "BGAs for High Reliability Applications," *Electron. Packaging and Production* (March 1998).
43. G. Derman, "Socket Suppliers Target Emerging High-density World of Ball- and Land-grid Arrays," *Electron. Eng. Times* (January 1998).
44. R. Ghaffarian, "CSPs Assembly Reliability," *Advancing Microelectronics*, **24** (6) (November 1997).
45. T. DiStefano and J. Fjelstad, "Chip-scale Packaging Meets Future Design Needs," *Solid State Technol* (April 1996).
46. A. Bindra, "Flex Circuits Branch Out," *Electron. Eng. Times* (19 August 1996).
47. R. Ghaffarian, "Reliability of Chip Scale Packages," *EEE Links* **4** (1), (January 1998), publication of NASA/GSFC, http://arioch.gsfc.nasa.gov/32/Linkspg.html.
48. S. Greathouse, "Chip Scale Packaging Breaks New Frontiers," *Solid State Technol.* (March 1996).
49. J. Vardaman, "CSPs: Hot New Packages for Cool Portable Products," *Solid State Technol.* (October 1997).
50. N. Takahashi *et al.*, "Three-Dimensional Memory Module," *IEEE CPMT—B* **21** (1) (February 1998).
51. M. Donlin, "IC Packaging Moves to the Front of the Design Cycle," *Computer Design* (July 1994).
52. "Dense Memory Stacks Hold up in Space," *Military and Aerospace Electronics* (January 1997).
53. Cover, *Solid State Technol.* (October 1992).
54. Dense-Pac Microsystems, Inc., advertisement mailer (7321 Lincoln Way, Garden Grove, CA 92641-1428).
55. "Worldwide Connector Sales Continue to Grow," *Electron. Packaging and Production* **37** (9) (10 July 1997) and private communication with Ron Bishop of Bishop & Associates (July 1997).
56. R. R. Tummala and E. J. Pymaszewski, *Microelectronics Packaging Handbook* (Van Nostrand Reinhold, New York, 1989).
57. L. L. Moresco, "Electron. System Packaging: The Search for Manufacturing the Optimum in a Sea of Constraints," *IEEE Trans. Components, Hybrids, Manufacturing Technol.*, 13 (2), 494-508 (September 1990).
58. N. Weste and K. Eshraghian, Principles of CMOS VLSI Design (Addison-Wesley, New York, 1985).
59. M. A. Fury, "Emerging Developments in CMP for Semiconductor Planarization," *Solid State Technol.* (April 1995).
60. "A Look at the Worldwide IC Packaging Market," *Electron. Design* (24 June 1996).
61. H. B. Bagoklu, *Circuits, Interconnections, and Packaging for VLSI* (Addison-Wesley, New York, 1990).
62. J. Lyke and J. Tausch, "Development of a High Performance 800-pin Count Package for 3-D WSI Systems," *Proceedings of the 1992 Government Microcircuits Application Conference* (Las Vegas, Nevada, November 1992).
63. E. J. Pymazewski, private communication (1992). (See Ref. 55.)
64. "Effects of Interconnect Granularity," http://www.ai.mit.edu/projects/transit/rcgp/chapter1.6.1.html.
65. D. P. Seraphim *et al.*, *Principles of Electron. Packaging* (McGraw Hill, New York, 1989).
66. B. W. Kernighan and S. Lin, "An Efficient Heuristic Procedure for Partitioning Graphs," *Bell System Tech. J.* **49** (2), 291-308 (February 1970).
67. J. Lyke, "Efficient Heterogeneous Three-Dimensional Packaging at a System Level," *Proceedings of the 16th DASC Conference* (AIAA/IEEE) (Irvine, CA, 26–30 October 1997).

68. J. Butler *et al.*, "Adapting Multichip Module foundries for MEMS Packaging," *Proceedings International Conference and Exhibition on Multichip Modules and High Density Packaging—MCM '98* (Denver, Co, 15–17 April 1998), pp. 106–111.
69. R. S. Irwin, "Solder Self-Assembly for MEMS," *Proceedings of the 44th International Instrumentation Symposium* (Reno, NV, 3–7 May 1998).
70. P. Cook, "Silicon Micromachining Applied to the Management of the Thermal Environment in Wafer Scale Integration Technology," Master's thesis, Air Force Institute of Technology (1992).
71. P. Madden, "Microvias Take Center Stage," *Printed Circuit Design* **15** (2) 19–22 (February 1998).
72. V. Adams *et al.*, "Low Cost Packaging for Accelerometers," *Electron. Packaging and Production* (December 1993).
73. T. Sudo, "Present and Future Directions for Multichip Module Technologies," *IEEE J. Solid-State Circuits*, 30 (4), 436–441 (April 1995).
74. C. T. Gray *et al.*, *Wave Pipelining: Theory and CMOS Implementation* (Kluwer Academic Publishers, Boston, MA, 1994).
75. J. C. Lyke, "Silicon Hybrid Wafer Scale Integration Interconnect Evaluation," Master's thesis, Air Force Institute of Technology (December 1989).
76. G. Hutcheson, "Ten Trends Shaping the Next Ten Years," *Solid State Technol.* (May 1997).
77. R. E. Sigliano, *et al.*, "Multilayer Ceramics: a Revitalization," *Electron. Packaging and Production* (September 1996).
78. S. Chillara *et al.*, "Build-up Laminates Used in High Density Applications," *Electron. Packaging and Production* (August 1997).
79. G. L. Kmetz, "Designing Copper-Trace Resistors," *Electron. Design* (13 May 1996).
80. W. Ko and M. Pecht, "Humidity and Corrosion Analysis and Design," in *Handbook of Electronics Package Design*, edited by M. Pecht (Marcel Dekker, Inc., New York, 1991).
81. J. J. Licari and L. R. Enlow, *Hybrid Microcircuit Technology Handbook* (Noyes Publications, Park Ridge, New Jersey, 1988).
82. G. N. Ellison, "Thermal Analysis with Affordable Programs," *Proceeding of the 5th Annual International Electronics Packaging Conference* (1985).
83. S. Dakuginow *et al.*, "Thermal Measurement Standards for ASIC Packaging," *Semiconductor Int.* (June 1996).
84. C. P. Minning, "Thermal Management," in *Multichip Module Design, Fabrication, and Testing*, edited by J. Licari (McGraw-Hill, Inc., 1995).
85. B. Ozmat "Interconnect Technologies and the Thermal Performance of MCM," *IEEE Trans. Components, Hybrids, Manufacturing Technol.*, **15** (5), 860–869 (1992).
86. L. S. Fletcher, "A Review of Thermal Enhancement Techniques for Electron. Systems," *IEEE Trans. Components, Hybrids, and Manufacturing Technol.* **13** (4), 1012–1021 (December 1990).
87. D. B. Nor, "Mechanical Aspects of Multichip Module Reliability," *JOM* **44** (7), 29–35 (July 1992).
88. C. Libove, "Rectangular Flat-Pack Lids Under External Pressure," Rome Air Development Center Technical Report, RADC-TR-76-118 (May 1976).
89. J. H. Comtois, "Structures and Techniques for Implementing and Packaging Complex, Large Scale Microelectromechanical Systems Using Foundry Fabrication Processes," Ph.D. thesis (Air Force Institute of Technology, 1996).
90. R. E. Albano and J. P. Keska, "Is Design Realization a Process? A Case Study," *IEEE CHMT* **13** (3) (September 1990).
91. B. J. MacLennan, "Who Cares About Elegance?" University of Tennessee, Knoxville, Technical Report no. UT-CS-97-344.
92. J. D. Cho *et al.*, "High Performance MCM Routing," *IEEE Design and Test of Computers* **10** (4), 27–37 (December 1993).
93. Q. Yu *et al.*, "Algorithmic Aspects of Three-Dimensional MCM Routing," *Proceedings of 31st ACM/IEE Design Automation Conference* (San Diego, 6–10 June 1994).
94. D. Maliniak, "MCMs Traverse the Cost Curve," *Electron. Design* (26 June 1995).

95. Charles Stein, ARL/VSSE (private communication, October 1998).
96. N. Virmani, Nick and Kusum K. Sahu, "Reliability and Radiation Sensitivity of Plastic Encapsulated Microcircuits in space Applications," Report prepared for NASA Goddard Space Flight Center Report, Log no. EPG-007-93.
97. G. Rose *et al.*, "Fundamentals of Plastic Encapsulated Microcircuits for Space Applications," presentation by the NASA Parts Project Office (October 1994).
98. For example, http://misspiggy.gsfc.nasa.gov/og/ and http://msfcpec1.msfc.nasa.gov/eh12/outgas.html contain downloadable information.
99. A. Garrison "Case History—GGS Sensor Module Chip on Board Evaluation," NASA Goddard Space Flight Center Code 312 (1993).
100. D. Duston *et al.*, "COTS-grade Electronics Set to Escape Earthly Bounds," Military & Aerospace Electronics (December 1995).
101. M. Pecht, "Plastic Encapsulated Microcircuits: the Hardness Assurance Committee of the NASA/AFSMC Space Parts Working Group," briefing (Alexandria, VA, 20–22 May 1996).
102. S. Clark, *et al.*, "Plastic Packaging and Burn-in Effects on Ionizing Dose Response in CMOS Microcircuits," *Proceedings of IEEE Trans. On Nuclear Sci.* **42** (6) (December 1995).
103. C. P. Wong, "Understanding the Use of Silicone Gels for Nonhermetic Plastic Packaging," *IEEE CHMT* **12** (4) (December 1989).
104. M. J. Loboda *et al.*, "Manufacturing Semiconductor Integrated Circuits with Built-in Hermetic Equivalent Reliability," Proceedings of the 1996 IEEE ECTC (Orlando, FL, May 1996).
105. Air Force Research Laboratory contract F29601—Avanteco.
106. "Highly Integrated Packaging and Processing," Air Force Research Laboratory contract F29601-92-C-0137, Subtask 03-07.
107. F. Miller, "Space Qualifiable, Corrosion Resistant, Near Hermetic Plastic Packages," Air Force Research Laboratory contract F29601- 97-C-0046.
108. M. Merker, *et al.*, "Rad-Pak Radiation Shielding for ICs at the Package Level," Air Force Weapons Laboratory Report no. AFWL-TR-83-117 (April 1984).
109. A. P. Schmid, "Feasibility of Rad-Pak for VHSIC Packages," Air Force Weapons
110. Laboratory Report no. AFWL-TR-86-128 (April 1988).
111. A. P. Schmid *et al.*, "Fabrication and Testing of Rad-Pak IC Packages," Air Force Weapons Laboratory Report no. AFWL-TR-86-25 (April 1988).
112. Space Electronics, Inc., San Diego, CA, http://www.spaceelectronics.com/SpaceProd/Technologies/RadCoat.html.
113. D. G. Mavis, "Integral Shielding of Multichip Modules of Natural Space Reaiation Environments," Defense Nuclear Agency Report no. DNA-TR-96-18 (Mission Research Corporation through Air Force Research Laboratory June 1996).
114. S. Patel and I. Burgess, "Integrated Scan Techniques Ease Chip and Board Tests," *EDN* **41** (24) 129–142 (21 November 1996).
115. T. Storey, "Testing MCMs with Boundary Scan," *EE-Evaluation Eng.* (September 1994).
116. L. Peters, "A Better Method for Testing CMOS ICs," *Semiconductor Int.* (November 1991).
117. S. Ehlscheid, "A Practical method to Increase Test Coverage Using IDDQ," *EE-Evaluation Eng.* (August 1995).
118. *Plug and Play ISA Specification*, Microsoft and Intel Corporations, Version 1.0a (1994).
119. M. Wilinson, "Microsystem Integration Levels under Question," *Electron. Eng. Times*, 15 June, 1998.
120. D. C. Jiles, *Introduction to the Electronic Properties of Materials* (Chapman and Hall, New York, 1994).
121. M. Little and J. Grinberg, "The 3-D Computer: An Integrated Stack of WSI Wafers," in *Wafer Scale Integration*, edited by E. Swartzlander (Boston, MA, Kluwer Academic, 1989).

9

Micromachined Rate Gyroscopes

T. N. Juneau,* W. A. Clark,* A. P. Pisano,† and R. T. Howe†

9.1 Introduction

Modern aviation, space travel, and navigation would be impossible without gyroscopes. Gyroscopes are used to measure the rotation angles or rotation rates between a moving-body fixed axis and an inertially fixed axis. This allows estimation of pointing direction or vehicle heading. In this respect a gyroscope is akin to a magnetic compass, except that a gyroscope can reference any direction (not only North) and is not dependent on magnetic field lines. Whereas a compass may be inaccurate in the face of magnetic anomalies or helpless in space, a shielded gyroscope can provide measurement independent of outside magnetic environment. Conventional gyroscope technology ranges from high-end inertial navigation instruments to low-end consumer product sensors.[1] This diversity in performance and operational requirements has spawned a vast diversity in gyroscope technology. Spinning-wheel, vibrating-tuning-fork, solid-state laser, and magnetohydrodynamic gyroscopes are but a few examples of current technology. Unfortunately, what almost all high-end precision technologies have in common is high cost, large size, and appreciable power consumption. Micromachining has given rise to the possibility of revolutionizing the field by providing inexpensive, miniature gyroscopes with good performance.

Inertial sensors are crucial for spacecraft missions. Attitude correction, maneuver control, tumble recovery, health monitoring, and large structure stabilization can all be implemented using accurate inertial measurements from gyroscopes.[2,3] Inertial sensors allow stabilization and pointing of instruments such as cameras, antennas, solar panels, and detectors. High-performance inertial sensors provide backup and "fill in the gaps" during Global Position System (GPS) outages caused by physical or electrical interference, sun-sensor loss as a result of solar eclipses, and star-tracker down time between star acquisitions. As with all space technology, inertial sensor size, weight, and power consumption must be minimized. Every added watt of power requires added solar panel area, which then adds more weight and cost. The need for miniature sensors is intensified by the goal to design microspacecraft several orders of magnitude smaller and cheaper than present generation spacecraft, as exemplified by Galileo (2223 Kg). Many of these applications do not require inertial navigation grade sensors, but rather require tactical grade sensors. Gyroscopes with 1–10 deg/h bias stability and accelerometers with 10–100 μg stability are adequate.[2]

Conventional macrotechnology gyroscopes cannot meet future microspacecraft or nanosatellite requirements.[4] Although conventional technology can be utilized in large spacecraft, an appreciable price is paid in weight, power consumption, and cost. Even the new generation of miniature tactical gyroscopes weigh many ounces and fill cubic inches of volume. Adding support circuitry for operation, filtering, trimming, compensation, and analog-to-digital conversion multiplies the volume, the power consumption, and the cost many fold. In fact, current miniature mechanical and quartz gyroscopes,[5,6] with all mounting brackets and support circuitry included, cost thousands of dollars, require nearly 12 cubic inches of circuit volume, and have a substantial power draw. More accurate ring-laser gyroscopes and fiber-optical gyroscopes can be even larger

*Integrated Micro Instruments (IMI), Berkeley, California.

†Berkeley Sensor and Actuator Center (BSAC), University of California, Berkeley.

and more expensive. Further disadvantages of conventional gyroscope technology include modest shock survival, low bandwidths, and inadequate lifetimes. The fact that many of the stabilizing gyroscopes aboard the NASA Hubble Space Telescope had to be replaced testifies to the need for improved technology. In short, there exists a clear and present need for innovative gyroscope technology.

Recognizing this unfulfilled need, microelectromechanical system (MEMS) researchers have pursued the goal of devising a micromachined gyroscope. The inherent size, weight, and power advantages of MEMS should allow micromachined rate gyroscopes to fill the void left by conventional technology. Batch fabrication utilizing standard very large-scale integration (VLSI) compatible surface-micromachining techniques should yield at least an order of magnitude in price reduction. The solid-state sensor robustness to both shock and vibration makes micromachined designs even more attractive, especially in launch environments. Finally, the absence of bearings or friction surfaces should translate into greater lifetime and long-term stability. As is shown in Fig. 9.1, current micromachined gyroscope technology has modest performance; projected future performance lies in the widely useful tactical regime.

Even more impressive gains in economy, size, and performance can be achieved by integrating circuitry with the silicon micromachined sensors. Placing interface electronics, signal-processing circuits, and analog-to-digital conversion on chips allows improved noise performance, extreme miniaturization, and inexpensive manufacture. The need for a multitude of discrete components on large printed-circuit boards can be virtually eliminated. In the future, integration may apply to more than just circuitry. The mechanical sensors themselves may be integrated to form entire monolithic microsystems.

The true micromachine revolution may not emanate from the individual sensors themselves, but from the integration of many sensors into a microsystem. Orthogonal triads of accelerometers[7,8] and gyroscopes[9–12] have produced an inertial measurement unit (IMU) on a fingernail-sized microchip with signal-processing included. The micrograph in Fig. 9.2 shows a prototype IMU designed by Berkeley Sensor & Actuator Center (BSAC) researchers and fabricated by Sandia National Laboratories.[13] Offspring of this technology could revolutionize the navigation industry.[14-16] Every portable computer and cell phone may someday have the navigation

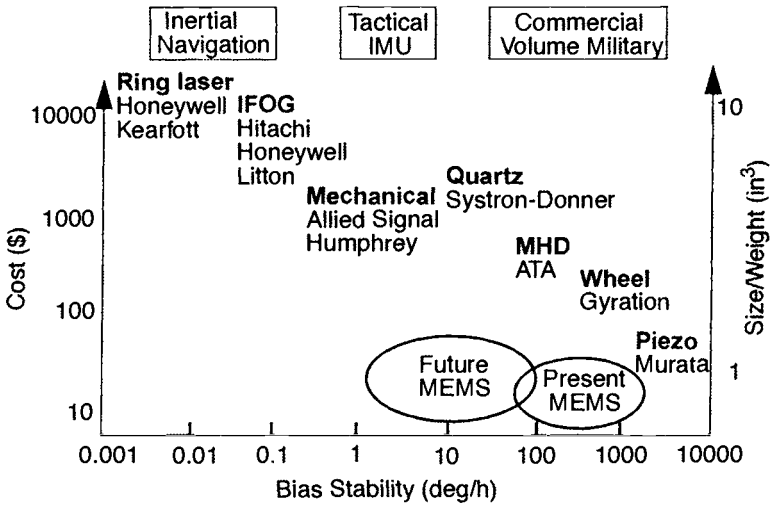


Fig. 9.1. Graph of price vs performance trade-offs with broad application areas. Size includes support circuitry. (IFOG: inertial fiber-optic gyro; MHD: magnetohydrodynamic).

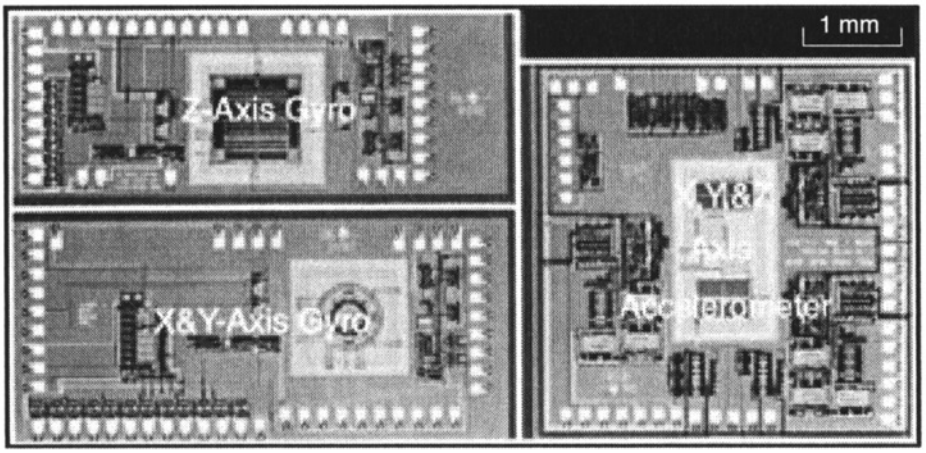


Fig. 9.2. Micrograph of the μIMU sensor array designed by BSAC and fabricated by Sandia National Labs. The *z*-axis gyro is at upper left; the *x*- and *y*-axis gyro is at lower left; *x*-, *y*-, and *z*-axis accelerometer is at right. This fingernail-sized chip is 9 × 5 mm.

capability of present-day aircraft. Total automobile safety and navigation modules may one day deploy multiple air bags, operate anti-skid braking, optimize ride comfort, and display maps and best routing to the driver. Even now, micromachined inertial sensor sales are experiencing explosive growth. Market forecasts[14,16] project 40–60% yearly growth, compounding to a $2-billion annual market by the year 2000. The space industry can directly leverage this technology base to vastly reduce costs while retaining performance. Combining other chemical sensors, physical sensors, and functionality could drastically alter the economics and capabilities of space science. Fleets of microsatellites smaller than a hockey puck may encircle the globe or explore the outer reaches of space.[4] Possibilities are endless, and the micromachine revolution has only just begun.

9.1.1 Survey of Micromachined Gyroscope Designs

As previously discussed, the principal use of gyroscopes is to measure orientation, heading, or pointing direction. Gyroscopes can be divided into two broad categories: one category measures orientation angles directly, while the other actually measures rotation rate. For this second type, true orientation angles must be estimated indirectly by integration. Free gyroscopes or spinning-wheel gyroscopes mounted on gimbals are included in the first direct-measurement category. These were the mainstays of aviation and navigation for decades.[17] However, the advent of inexpensive, miniature computers allowed the rise of strap-down navigation. The term strap-down navigation refers to inertial measurement units that have no gimbal-mounted gyroscopes, but rather use nongimbal rate gyroscopes. This computationally intensive method uses rate gyroscopes of the second broad category. In essence, the angular rate measurements from rate gyroscopes are integrated to estimate orientation angles. Although this approach requires more computation, advantages in gyroscope cost and reliability have allowed strap-down navigation to become dominant in many applications. Some applications, such as platform stabilization and missile guidance, require angular rate in addition to true orientation angle, rendering rate gyroscopes the natural choice.

Micromachine technology lends itself well to the fabrication of strap-down rate gyroscopes. Fabrication of rotary micromachines with long-lifetime, low-friction bearings is exceedingly difficult using micromachining, so free gyroscopes or spinning-wheel gyroscopes are rarely attempted. Instead, spring-mounted vibratory gyroscopes requiring no bearing and measuring an-

gular rate are the norm. A number of research centers are investigating silicon micromachined gyroscopes. Each group has different design philosophies and fabrication techniques. Most work has focused on variations of surface micromachining with structural films between 2 and 20 μm thick. The basic operating principle of all vibrating gyroscopes is based on the generation and detection of a Coriolis acceleration. For a Coriolis acceleration to be generated, a proof mass must be put in motion. Some of the typical proof masses for micromechanical gyroscopes are pictured in Fig. 9.3. These broad types include translational tuning fork, rotary disk, translational shuttle, structural mode ring, rotational X-shape, and translational X-shape. Because the automobile market offers massive sales volume and requires only low performance, most research has focused on micromachining automotive grade 1 deg/s gyroscopes.

Draper Laboratories was the first organization to develop a micromachined gyroscope.[18] Recent Draper designs include a tuning-fork gyroscope[19] developed for both the automotive and aerospace markets several years ago. The silicon-on-glass fabrication process used is not conducive to circuit integration, but did allow quick prototyping. However, Analog Devices Inc. (ADI) may soon be the first to actually commercialize a low-cost, automotive-grade gyroscope using a translational shuttle design.[20] Delphi Automotive Systems has developed a vibrating ring derived from the "wine glass" gyro for the automotive market.[21,22]This design has good attenuation of undesirable translation and rotational disturbances. However, this excellent robustness also results in relatively low sensitivity, making improvement to inertial grade performance difficult. Several Japanese manufacturers including Murata[23] and Matsushita[24] have published gyroscope designs involving both translational shuttle and translational X-shape designs. Again, these are aimed directly at the low-accuracy, high-volume automobile market. NASA Jet Propulsion Laboratory (JPL) produced a novel rotational X-shape gyroscope with a brass pillar for added inertia.[25] It has exhibited useful drift performance, but is not designed for mass production. Clark at BSAC has also developed a fully integrated, translational shuttle-type microgyroscope that measures rotation rates perpendicular to the substrate.[9,26] A similar shuttle design was later reported by Samsung.[27] This design ports naturally into the new high-aspect-ratio processes[28,29] made possible by deep-trench etcher technology and is the basis for commercial gyroscope development by Integrated Micro Instruments (IMI).[30] As Fig. 9.4 shows, the new high-aspect-ratio SOI (silicon-on-insulator)-MEMS process yields much thicker mechanical structures and an order of magnitude more capacitive sense area compared with traditional surface micromachining. The resulting increase in sensitivity is projected to yield performance in the 0.2–0.5 deg/$\sqrt{\text{h}}$ range. Although this is clearly not an exhaustive list of researchers, the most common types of silicon micromachined gyroscopes have been included.

The micromachined dual-axis rate gyroscope used as a design example for this chapter utilizes a rotational disk-type proof mass. This design is first mentioned in a 1992 internal BSAC

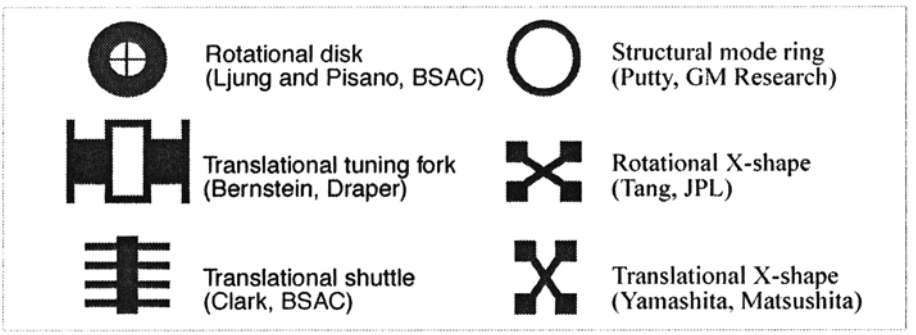


Fig. 9.3. Silhouettes of common micromachined designs.

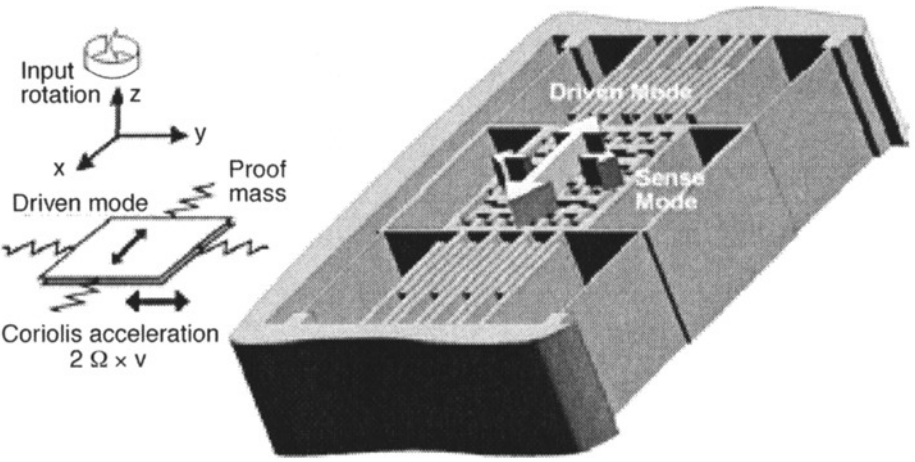


Fig. 9.4. Proposed z-axis gyroscope as implemented in the new SOI-MEMS technology by Integrated Micro Instruments (IMI).

document written by Ljung and Pisano.[31] This was chosen because it illustrates all the important features and design requirements common to the majority of the micromachined gyroscope designs mentioned above. A prototype has been fabricated and tested to compare with theory. The dual-axis rate gyroscope is one of the few micromachined gyroscope designs that measure angular rate about two orthogonal axes simultaneously. The symmetric circular design is the key to dual-axis operation. In contrast with two separate gyroscopes, the dual-axis rate gyroscope requires only a single set of drive and sense interface circuitry and does not require precise axial alignment of two separate proof masses. Furthermore, two separate gyroscopes would have two separate drive frequencies, which may result in corrupting electrical cross-talk.

9.1.2 Chapter Outline

The chapter is divided into six sections. The Introduction, Sec. 9.1, presents motivation for the research and reviews the state of the art. Section 9.2 describes the underlying theory of operation starting from gyroscopic dynamics. The next three sections detail aspects of the gyroscope design: Sec. 9.3 delves into the mechanical structure design; Sec. 9.4 explains electrical interface and signal processing; and Sec. 9.5 provides an overview of electrical and mechanical design trade-offs. The final section, 9.6, provides solutions to current problems and envisions future directions. Concepts presented throughout the chapter are directly applicable to most micromachined gyroscope design. The underlying gyroscopic dynamics, mechanical structure design, resonant drive design, Coriolis motion sensing, and signal processing are similar for all vibrating gyroscopes.

9.2 Principle of Operation

The classical mechanical gyroscope is composed of a spinning wheel or rotor that exhibits Coriolis acceleration because of conservation of angular momentum.[17] Unfortunately, fabricating microscopic, low-friction bearings needed for this classical approach is challenging. This chapter presents a rotor mounted on springs rather than bearings. Hence, the rotor motion is a counterclockwise and then clockwise angular oscillation rather than a constant spinning motion. This oscillatory motion still exhibits gyroscopic Coriolis motion that can be used to measure angular rate.

The dual-axis rate gyroscope was fabricated using the ADI BiMEMS surface micromachining process.[32] As Fig.9.5 shows, the micromachining is compatible with standard integrated circuit processing, so circuits are integrated with the mechanical structures on the same substrate. All

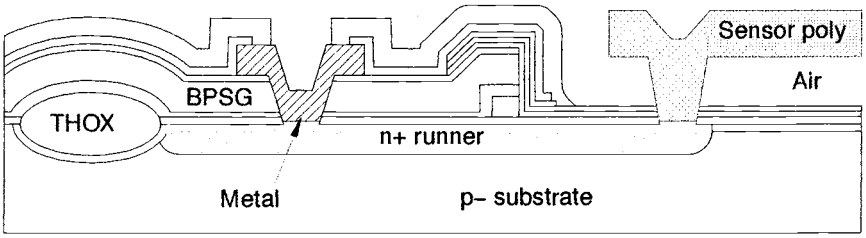


Fig. 9.5. Cross-section of ADI BiMEMS process showing interconnect from circuitry on the left to the mechanical structure on the right.

mechanical structures are fashioned from a planar, thin-film polysilicon layer. The inertial rotor is in fact a 2-μm-thick polysilicon disk. As depicted in Fig. 9.6, this inertial rotor is suspended 1.6 μm above the substrate by four symmetrically placed beams anchored to the substrate. These beams provide a torsional suspension allowing rotational compliance about all three axes. Fabrication begins with a 1.6-μm sacrificial silicon oxide layer applied to a bare silicon substrate. Holes are etched into this oxide to provide the anchor points to the underlying substrate. Next, the structural polysilicon 2 μm thick is applied over the sacrificial oxide and patterned. Finally, the sacrificial oxide is removed leaving the dual-axis rate gyroscope suspended above the substrate.

The basic operating principle of all vibratory gyroscopes relies on the generation and detection of a Coriolis acceleration. In gyroscopic dynamics, there is a distinct motion about all three orthogonal axes.[11] First, the proof mass is put into oscillatory motion about the z axis perpendicular to the substrate. Once in motion, the proof mass is sensitive to angular rates induced by the substrate being rotated by outside forces. This input rotation rate is on a second axis (say the x axis) perpendicular to the first drive axis. The input rate induces a Coriolis acceleration about the third axis (y axis), which is perpendicular to both the z-axis drive and the rate input x axis. This Coriolis acceleration induces a Coriolis motion with an amplitude proportional to the angular rate of the substrate. Detecting the Coriolis motion induced by the Coriolis acceleration allows the original rotation of the gyroscope to be inferred.

The key to dual-axis operation is the circular symmetry of the device. The mechanical sensor is identical about the x and y axis in the plane of the substrate. When the inertial rotor is

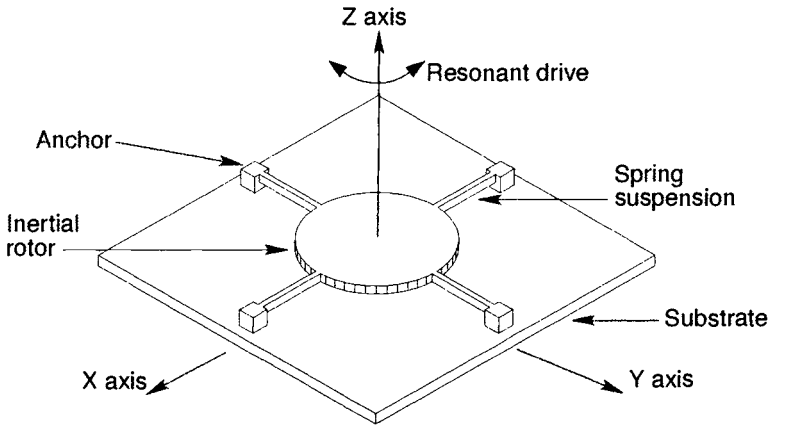


Fig. 9.6. Conceptual illustration of the underlying mechanical sensor element for the dual-axis rate gyroscope.

resonating, any rotation rate of the substrate about the x axis will induce a Coriolis angular acceleration about the y axis, which in turn induces a tilting oscillation of the rotor about the y axis. Because the mechanical gyroscope is symmetric in two orthogonal axes, any rotation rate about the y axis likewise invokes a tilting oscillation about the x axis, thereby allowing dual-axis rotation rate measurement. Hence the rotation rate about both these axes can be measured simultaneously, resulting in a dual-axis rate gyroscope. The axes are orthogonal, so the small tilting deflections associated with Coriolis motion will not mix or combine. However, the designer must be vigilant to ensure low cross-axis sensitivity. The conceptual illustration in Fig. 9.7 shows the out-of-plane tilting motion of the inertial rotor in response to a rotation rate input.

These intuitive dynamics are reflected in the simplified dynamics of gyroscopic motion. The linearized, first-order dynamics are described in Eqs. (9.1) and (9.2), where α_x and α_y are Coriolis accelerations for each axis, Ω_x and Ω_y are the input rotation rates to be measured, and $\dot{\theta}_z$ is the resonant drive angular rate.

$$x \text{ axis: } \alpha_x = 2\Omega_y\dot{\theta}_z \tag{9.1}$$

$$y \text{ axis: } \alpha_y = -2\Omega_x\dot{\theta}_z \tag{9.2}$$

The Coriolis acceleration is proportional to the product of the drive motion and the input rotation rate to be measured. The drive motion is an oscillation at the natural frequency ω_z of the drive axis. Therefore the tilting oscillations about the x and y axes are at the same frequency as the resonant drive frequency ω_z. These tilting oscillations are amplitude-modulated signals with amplitude proportional to the respective rotation rate inputs. Therefore, the rotation rate can be inferred by measuring the rotor tilt oscillation and demodulating this measurement at the resonant drive frequency. This is very much like amplitude-modulated (AM) radio. The dual-axis rate gyroscope oscillation drive frequency can be thought of as the radio station frequency, while the Coriolis motion is analogous to the modulated signal transmitted from the station. A listener uses a radio receiver to pick up and demodulate the station signal, leaving the original sound or music. For the dual-axis rate gyroscope, sense circuitry measures the rotor Coriolis tilt motion and demodulates, leaving a voltage signal proportional to the original angular rate input.

For an AM radio receiver to operate, the radio must be tuned to the proper frequency. A similar analogy exists for vibratory gyroscopes. Performance can be much improved by matching drive

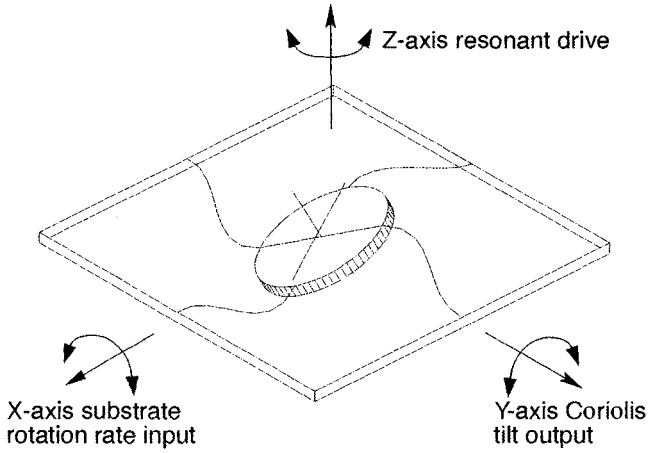


Fig. 9.7. Conceptual view of the Coriolis tilting motion invoked by rotation of the substrate.

oscillation frequency with the natural frequencies of both sense axes. This is because the gyroscope sense axes have large resonant peaks, so several orders of magnitude improvement in sensitivity is gained through matching modes. This is analogous to tuning a radio receiver until the best reception for a particular station is found. In a gyroscope, tuning using electrostatic force is used to "down tune" (i.e., reduce the natural frequency) of both sense axes natural frequencies until they nearly match the drive frequency. Because electrostatic force is always attractive, the electrical force between the rotor and the substrate-mounted electrodes is opposite the restoring force of the beam suspension. As the rotor moves closer to the substrate, the electrostatic force becomes stronger. Likewise, as the rotor moves away from the substrate, the electrostatic force becomes weaker. Tilting motion also exhibits the same behavior as the edge tilted closer to the substrate is pulled down, while the edge tilted away from the substrate has less force acting upon it. Hence, the electrostatic force acts like an unstable or "negative" spring working to force the rotor away from steady-state equilibrium. As will be discussed later, close mode matching can exhibit some disadvantages if the gyroscope is operated open-loop without force balancing. With the basic dynamics illustrated, the problem of actually setting the gyroscope in motion and sensing Coriolis tilting can be tackled.

There can be no rotation rate sensing unless the inertial rotor is driven into rotational resonance. This task is accomplished using a highly linear electrostatic comb drive.[33] Pairs of alternating differential combs, half drive combs and half sense combs, can be seen surrounding the rotor in Fig. 9.8. A voltage difference between the inertial rotor and the stationary drive combs around the rotor circumference induces attractive electrostatic forces on the rotor. By varying this voltage difference, alternating clockwise and counter-clockwise moments can be applied to the rotor, thereby exciting oscillation. The key to exciting oscillation is to use this drive moment to cancel out the natural air damping of the rotor. With no damping, the system becomes unstable and the inertial rotor self-oscillates at its natural frequency. A differential trans-resistance amplifier (i.e., current-to-voltage converter) measures rotor rotation rate using the sense combs and provides positive feedback to the drive combs to effectively cancel viscous damping.[34] Since the amplitude of resonance directly determines the scale factor, an automatic gain control loop is used to ensure constant oscillation amplitude.[35] The details of resonant drive and amplitude control are given in Subsec. 9.4.1.

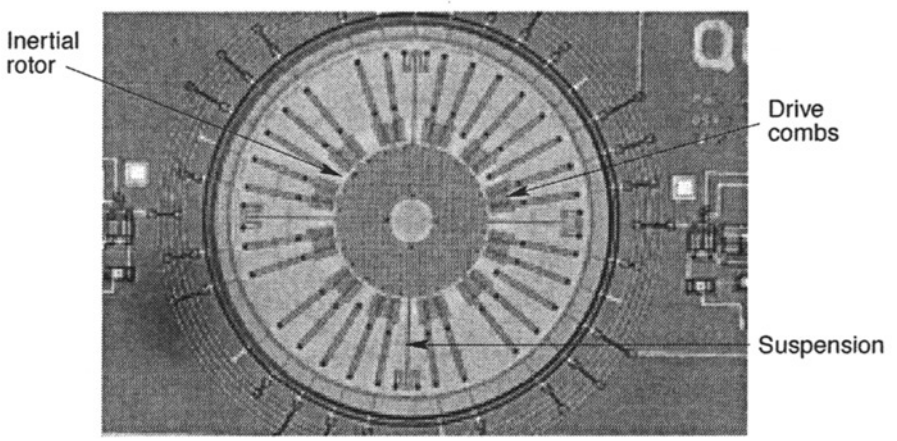


Fig. 9.8. Micrograph of the first generation dual-axis gyroscope fabricated by Analog Devices in the BiMEMS foundry.

With the rotor in oscillatory motion, the gyroscope is sensitive to substrate rotation rates. As previously noted, a rotation rate input induces a Coriolis acceleration, which in turn induces a tilting oscillation. Using a differential capacitance measurement scheme to detect the tilt oscillation amplitude, the original rotation rate input can be inferred. In Fig. 9.9 the rotor is shown with an underlying pair of quadrant-shaped electrodes. These electrodes form a capacitive divider with the inertial rotor. If the rotor tilts about the axis perpendicular to the page, then the capacitance of one sense capacitor increases while the capacitance of the other decreases.[36] This differential change in capacitance is detected via an integrator electrically connected to the inertial rotor in conjunction with a modulated sense voltage applied between the pair of quadrant shaped electrodes. If the sense capacitors are equal, then the modulation voltage causes charge to flow between the sense capacitors with no charge escaping through the integrator. If the sense capacitors are unbalanced because of inertial rotor tilt, then some charge flows through the integrator producing a measurement voltage proportional to tilt displacement and modulated at the sense-voltage frequency. Hence, the Coriolis motion is measured, and the desired input rotation rate measurement can be derived through signal processing.

Dual-axis operation is achieved by placing four quadrant-shaped electrodes beneath the rotor. This allows Coriolis tilt detection about both orthogonal axes. A different voltage modulation frequency is used for each pair of sense electrodes, thereby allowing the signal-processing circuitry to discriminate between the two axes. Separate demodulation circuits for each sense axis provide two voltage outputs proportional to the two angular rate inputs. The output voltages must be demodulated twice: the first demodulation removes the sense modulation frequency, and the second removes the drive resonant frequency leaving the base band rate input measurements.

The micrograph in Fig. 9.8 is of a dual-axis rate gyroscope designed at BSAC and fabricated in the ADI BiMEMS foundry, and will be used as a design example. Table 9.1 lists the key structural dimensions, dynamic model parameters, and noise performance.

9.3 Mechanical Structure

9.3.1 Simplified Gyroscope Dynamics

Intuitive dynamics discussed earlier are reflected in the simplified dynamics of gyroscopic motion. The full dynamical equations of motion are nonlinear and coupled,[17,37] and thus not

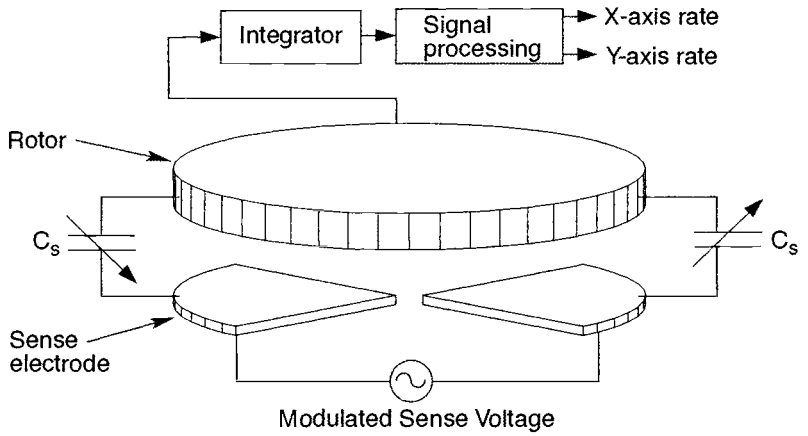


Fig. 9.9. System-level schematic of the Coriolis tilt motion sense electronics for a single axis. Note only two of four quadrant electrodes are shown.

Table 9.1. Final Design Parameters

Design Parameters	ADI Closely Matched Modes	ADI Unmatched Modes
Outside rotor radius R	150 μm	150 μm
Inside rotor radius R_i	50 μm	50 μm
Suspension beam length L	180 μm	180 μm
Suspension beam aspect ratio b/h	0.90	0.90
Number of drive comb gaps N	384	384
Drive mode frequency ω_z	28.3 kHz	28.2 kHz
Drive/sense mode mismatch	2%	8.5%
Quality factor Q	960	950
Experimental random walk noise	0.08 deg/s/$\sqrt{\text{Hz}}$	0.7 deg/s/$\sqrt{\text{Hz}}$

suitable for design synthesis. For this reason, approximations must be made. These approximations lead to a linearized, partially decoupled set of equations of motion,[38] which can be used for system design and performance optimization.

9.3.1.1 Linearized Equations of Motion

Derivation of simplified dynamical equations begins with key assumptions regarding the underlying kinematics and dynamics. The four-beam suspension is designed to attenuate translational accelerations by placing translation natural frequencies far from rotation natural frequencies. Hence, the system can be satisfactorily modeled using only angular rotation coordinates with the assumption that the inertial rotor stays centered. Although the dual-axis rate gyroscope is sensitive to angular accelerations about all axes, these terms can be ignored since they are small and typically far below the Coriolis motion bandwidth of interest. In addition, more terms can be ignored since the drive-axis angular rate thoroughly dominates the dynamics, and therefore, all centripetal and most Coriolis terms other than those of interest may be dropped. Details regarding these approximations can be found in Ljung.[39] This leaves the linearized set of equations (9.3)–(9.5), which represent the rotational dynamics about all three orthogonal axes.

$$I_{xx}\ddot{\phi} + C_{xx}\dot{\phi} + K_{xx}\phi = I_{zz}\Omega_y\dot{\theta} \tag{9.3}$$

$$I_{yy}\ddot{\psi} + C_{yy}\dot{\psi} + K_{yy}\psi = -I_{zz}\Omega_x\dot{\theta} \tag{9.4}$$

$$I_{zz}\ddot{\theta} + C_{zz}\dot{\theta} + K_{zz}\theta = M_z \tag{9.5}$$

The tilt of the inertial rotor is represented by ϕ about the x axis and ψ about the y axis. The rotation angle of the spinning rotor about the z axis is represented by the angular coordinate θ. The rate inputs to be measured are Ω_x about the x axis and Ω_y about the y axis. Constant coefficients I_{ii} model the moment of inertia, C_{ii} model viscous damping, and K_{ii} model suspension spring rate about each respective axis. The drive torque or moment about the z-axis M_z is chosen such that θ oscillates at the z-axis natural frequency ω_z. Note the gyroscope duality is visible in the mirror image symmetry of the two sense axis equations of motion Eqs. (9.3) and (9.4). The Coriolis accelerations, the right-hand terms in Eqs. (9.3) and (9.4), are proportional to the product of the drive angular rate and the input rotation rate to be measured. These accelerations cannot be

measured directly, but rather the resulting angular deflections are measured. The mechanical sensitivity relating rate input to deflection amplitude is discussed next.

9.3.1.2 Mechanical Sensitivity to Angular Rate Input

Mechanical sensitivity is of paramount importance because it directly determines the minimum detectable signal in the given interface electronic noise. Electronic noise will limit both the smallest possible rotor-tilt deflection that can be detected and the measurement resolution. Hence high sensitivity is desirable for low minimum detectable signal. The mechanical sensitivity relating Coriolis rotor-tilt amplitude to substrate angular-rate input for each sense axis can be derived from Eq. (9.3), with the final form shown by Eq. (9.6) for the x axis. Assuming the x and y axes are nearly identical, then each axis will have equivalent magnitude sensitivity, as approximated in Eq. (9.6).

$$\frac{\|\phi\|}{\|\Omega_y\|} \cong \left\| \frac{2\theta_0 \omega_z}{\omega_x^2 + \dfrac{j\omega_x \omega_z}{Q_x} - \omega_z^2} \right\| \tag{9.6}$$

This is the standard second-order frequency response, with θ_0 the resonant drive amplitude, ω_z the resonant drive frequency, ω_x the sense axis frequency, and Q_x the quality factor or inverse damping coefficient. The sensitivity equations can be used to improve sensitivity and hence noise performance. Sensitivity is proportional to drive amplitude θ_0, so larger drive oscillation amplitude is advantageous. Also, Fig. 9.10 shows a clear maximum sensitivity when the sense and drive frequencies are matched, as discussed in Subsec. 9.3.1.3.

9.3.1.3 Improving Sensitivity Via Mode Matching

Since the rate sensor is operated in partial vacuum (50-100 mtorr), the quality factors Q_x and Q_y are generally above 1000. Thus, a large resonant peak exists, which provides the possibility of excellent gain when drive and sense frequencies match. Sensitivity is plotted against sense axis natural frequency in Fig. 9.10. Unmatched natural frequencies give poor signal gain; whereas, exactly matched natural frequencies give maximum gain. To improve sensitivity, the four-beam suspension is designed to match sense and drive modes closely. However, there are drawbacks to matching-mode frequencies, especially if the device is to operate open-loop.

Under open-loop operation, matching natural frequencies closely forces the inertial instrument designer to confront the underlying trade-offs between noise performance and other performance criteria. For example, the cross-axis sensitivity of the dual-axis rate gyroscope degrades with closely matched modes. Another consideration is that gain and phase both change dramatically at the resonance peak, so the scale factor is not constant over any appreciable bandwidth. Furthermore, natural frequency drift can cause both substantial scale-factor and output-phase changes. These consequences limit practical mode matching for *open-loop* operation to no closer than 5–10%.

Closed-loop feedback is often used to enhance performance beyond the limitations of open-loop operation. Feedback typically improves scale-factor stability, linearity, bandwidth, and operating range. A well-designed feedback loop will not significantly alter noise performance. Thus the noise advantages of mode matching can be retained, while the dynamic advantages of closed-loop feedback can be added. In this case, null position moment balancing would be used. This technique balances all external moments and Coriolis accelerations, thereby nulling rotor tilt position. The output voltage measurement is then the feedback voltage required to balance the Coriolis moment. It is projected that closed-loop moment balancing of the dual-axis rate gyroscope

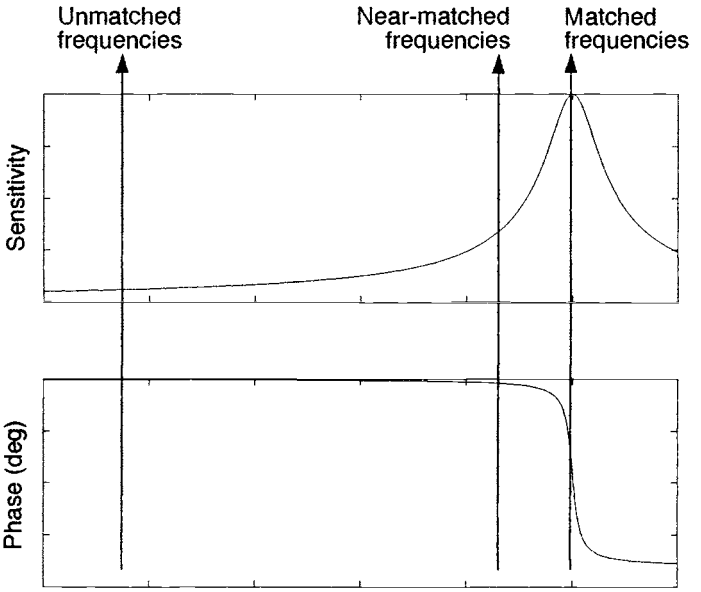


Fig. 9.10. Bode plot showing simulated sensitivity and output phase vs sense-axis natural frequency.

will alleviate many scale-factor and phase difficulties while also reducing cross-axis sensitivity. Many present micromachined gyroscopes operate open-loop because of simplicity and stability.

If sense frequencies are to be matched or nearly matched to drive frequencies, then the entire mechanical structure must be carefully designed. The suspension beams could be designed such that all mode frequencies are nearly matched, but process variation always results in a random mismatch of perhaps 2–20%, depending on the process and gyroscope design. Instead, the mechanical suspension was designed for coarse frequency tuning while using electrostatics for fine tuning after fabrication. Fine tuning using electrostatic force can be used to down tune both sense-axis natural frequencies until they nearly match the drive frequency. The principal drawback of this method is that mechanical shock survivability is reduced because of the electrostatic attraction on the rotor caused by tuning voltages.

9.3.2 Modeling of Mechanical Structure

The equations of motion use lumped parameters for moment of inertia, suspension stiffness, and damping. These parameters must be derived from mechanical sense-element dimensions and material properties. The parametric derivations were intentionally kept simple for two principal reasons. First, a complex derivation including all possible effects becomes intractable and useless as a design tool. Second, parameter errors due to manufacturing process variations often outweigh second-order theoretical effects. Hence, a balance must be struck between accuracy and usability.

9.3.2.1 Rotor Moment of Inertia

The total system moment of inertia is the sum of the rotor moment of inertia and the moment of inertia of the electrostatic combs. Although the notion is counter-intuitive, the gyroscope has better performance if the center is hollowed out. Leaving the center hollow does in fact reduce performance-enhancing parameters such as the moment of inertia and sense-capacitance area, but only slightly. That is because moment of inertia is proportional to the fourth power of radius, and sensitivity is proportional to radius cubed. On the other hand, increased rotor radius increases

surface area. Larger surface area implies larger parasitic capacitance, which increases the input-referred electrical noise. Larger surface area also results in larger electrostatic forces impinging on the rotor, which may pull the rotor toward the substrate. Rotor mass also increases with larger radius. This lowers translational natural frequencies, which in turn diminishes both shock and vibration survival. Hence, an optimum exists: balancing performance-improving inertia and sensitivity with performance-degrading parasitic capacitance and mass. This optimization encompasses both mechanical and electrical system aspects of the design, and is discussed in Sec. 9.5.

The exact integral for moment of inertia is complicated as a consequence of the etch holes and gaps between resonant drive combs. Instead, estimates based on effective polysilicon density were used. The effective density correctly discounts the rotor polysilicon density for release etch holes and the comb density for comb air gaps. For example, in the ADI process rotor density, ρ_e was discounted to 92%, and comb density ρ_{ce} was discounted to 25%. As Eq. (9.7) suggests, the total z-axis moment of inertia is the sum of rotor inertia I_{zr} and drive comb inertia I_{zc}. Using the outer radius R_o and inner radius R_i in conjunction with structural polysilicon thickness h, and rotor polysilicon effective density ρ_e, the rotor moment of inertia about the drive axis I_{zr} can be derived. Likewise, the electrostatic comb moment of inertia I_{zc} can be estimated using the inner comb radius R_{ci}, and outer comb radius R_{co}, and the effective comb density ρ_{ce}.

$$I_{zz} = I_{zr} + I_{xc} \tag{9.7}$$

$$I_{zz} = \frac{\pi}{2}\rho_e h(R_o^4 - R_i^4) + \frac{\pi}{2}\rho_{ec} h(R_{oc}^4 - R_{ic}^4) \tag{9.8}$$

Similarly, the moment of inertia about the sense axis can be calculated. As Eq. (9.9) shows, the sense-axis moment of inertia is exactly half the drive-axis moment of inertia. Typical z-axis moment of inertia estimates for ADI designs were 5×10^{-18} kg/m.

$$I_{xx} = I_{yy} = \frac{1}{2} I_{zz} \tag{9.9}$$

9.3.2.2 Quality Factor and Rotor Air Damping

Estimating damping or quality factor is important for designing resonant drive circuitry. The system damping is composed of both structural damping intrinsic to the mechanical element and air damping provided by the surroundings. Although the dual-axis rate gyroscope is operated in intermediate vacuum (50–100 mtorr), the air damping is still dominant. For small motions, the air damping torque is approximately linear with respect to rotor rotation rate, so linear damping coefficients are used for dynamical modeling. The damping may be represented in nondimensional form by using quality factors. The quality factor Q_z for the drive mode is much higher than that of the sense modes Q_x and Q_y because the sense modes exhibit squeeze film damping while the drive mode does not.

Unfortunately, the quality factors are not simple to calculate. The fluid interactions are complicated by the fact that the assumption of viscous continuum breaks down in intermediate vacuums. Very approximate methods exist in the literature[40] for calculating the drive-axis quality factor using effective viscous damping coefficients. On the other hand, no analytical solution exists for squeeze film damping of a circular rotor riddled with etch holes. Some recent researchers have investigated squeeze film damping of a rectangular plate,[41–43] but all results are through numerical computation of specific cases, which cannot be generalized to the current gyroscope problem. However, experimentation revealed an approximate empirical relationship of sense axis Q being an order of magnitude lower than drive axis Q. The experimental values measured at 60 mtorr were approximately 900 for the sense-axis mode and 30,000 for the drive-axis mode.

9.3.2.3 Suspension Spring Design and Resulting Spring Constants

The suspension design must satisfy several key design constraints. First, any mechanical structure deformation caused by fabrication-induced residual stress gradients must be accommodated by the suspension. This was particularly true for gyroscopes fabricated by ADI because of the relatively large stress gradients present in the structural polysilicon. Although recently developed surface micromachine processes make flat polysilicon, new silicon oxide-aluminum processes[13] still have serious warpage. Second, all natural frequencies, both rotational and translational, must be above a given threshold for shock and vibration robustness. This work tried to keep all frequencies above high audio frequencies of 10–20 kHz, with the preference that the rotational frequencies be lower than the translational frequencies.

Finally, to enhance sensitivity the suspension was designed for close but not exact matching of the rotational drive mode and both rotational sense modes to enhance sensitivity. Ideally, the sense modes would be slightly higher than the drive mode so that electrostatic down-tuning can equalize all rotational natural frequencies. That the sense-axis moment of inertia is half the drive-axis moment of inertia implies that the sense-axis spring constant must be approximately half the drive-axis spring constant for effective mode matching. The following subsections lay out the fundamental considerations and governing equations constraining beam suspension design.

9.3.2.3.1 Inside vs Outside Suspension

The overriding design consideration that must be tackled is whether to place the suspension inside or outside the rotor, as shown in Fig. 9.11. Placing the suspension inside the rotor can reduce sensor area and force all translation modes to be higher relative to the rotational modes. Unfortunately, the inside suspension can be disastrous if the residual stress gradient causes the polysilicon structure to warp into a convex shape with the center at the high point. This would force the outer edge of the gyroscope to drag on the substrate and thereby halt its motion. For this reason, outside suspension had to be used for the ADI fabricated gyroscopes.

9.3.2.3.2 Rotational Spring Constant Derivation

The linear spring constants were derived using simple Euler beam bending theory. A spring constant represents the linear relationship between angular rotor displacement versus suspension torque and is accurate for the case of small deflections. All spring constants were assumed uncoupled for this simplified analysis. The spring constants can be found by calculating the torque required to deflect the rotor a given angular displacement θ. Each beam can be thought of as a

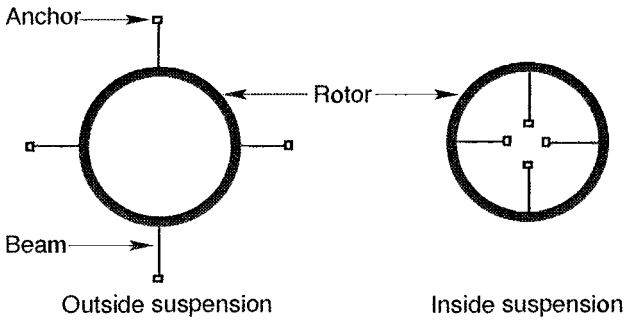


Fig. 9.11. Conceptual illustration of both inside and outside suspension designs. Inside suspension provides more compact designs, but excessive residual stress gradient can force the use of an outside suspension scheme.

cantilever with a force and an angular moment applied at the end attached to the rotor. For a given rotor angle, the beam must satisfy both a deflection and a slope constraint at the end attached to the rotor. Hence, there are two linear equations (one for end deflection and one for end slope) with two unknowns (force and moments applied by the rotor). Converting the force and moment applied upon the beam end to a total moment about the rotor center provides the spring constant for a single beam. The resulting moment about the rotor center for a single-beam M_R is the sum of the single-beam moment M and the single-beam force F multiplied by the radius from beam end to rotor center R.

$$M_R = M + F \times R \tag{9.10}$$

Simply multiplying the spring constant calculated for a single beam by 4 gives the total angular spring constant for all four suspension beams in the drive mode. This is mathematically represented by Eq. (9.11), where K_{zz} is the total drive-axis spring constant and K_{bz} is the spring constant for a single beam.

$$K_{zz} = 4K_{bz} = \frac{4M_R}{\theta} \tag{9.11}$$

The sense-mode spring constants K_{xx} and K_{yy} can be approximated by twice the single-beam spring constant K_{bx} as in Eq. (9.12). This is because only two beams exhibit bending, while the other two beams exhibit weak torsion, which produces negligible torque.

$$K_{xx} = K_{yy} = 2K_{bx} \tag{9.12}$$

In order to match mode frequencies, the sense-axis spring constant must be half the drive-axis spring constant. This requirement can be proved by assuming the natural frequencies ω_x, ω_y, and ω_z are equal and then substituting expressions based on total spring constants and moments of inertia.

$$\omega_x = \omega_y \approx \omega_z \tag{9.13}$$

$$\sqrt{\frac{K_{xx}}{I_{xx}}} = \sqrt{\frac{K_{yy}}{I_{yy}}} \approx \sqrt{\frac{K_{zz}}{I_{zz}}} \tag{9.14}$$

Now the total spring constants can be replaced by single-beam expressions, while z-axis moment of inertia can be replaced by twice the x-axis moment of inertia, and both the x- and y-axis moments of inertia are assumed equal.

$$\sqrt{\frac{2K_{bx}}{I_{xx}}} = \sqrt{\frac{2K_{by}}{I_{yy}}} \approx \sqrt{\frac{4K_{bz}}{2I_{xx}}} \tag{9.15}$$

$$K_{bx} = K_{by} \approx K_{bz} \tag{9.16}$$

Therefore, all frequencies are ideally matched if all *single*-beam spring constants are identical. The sense-axis spring constant (with only two beams in bending) is then naturally one-half the drive-axis spring constant (with all four beams in bending). In practice, process variation forces electrostatic tuning to be used, so in fact the sense x- and y-axis spring rates will be designed slightly larger, as Subsec. 9.3.2.3.3 regarding frequency ratios shows. However, setting all single-beam spring constants *approximately* equal is a good rule of thumb when attempting to match all mode frequencies.

Serpentine or folded suspensions may be used to conserve die area and provide some stress relief (see Fig. 9.12). If the serpentine beams are approximately the same length, then a simple approximation exists. Because each beam is nearly identical in both dimension and end conditions, each beam must exhibit identical deflection and slope change. Hence, a twofold suspension can be simplified into a single-beam problem with the beam undergoing half the total deflection and half the slope change of the total suspension. Generalizing this concept to multiple folds simply requires a single beam to exhibit deflection and slope divided by the number of folds.

Previous assertions and the following calculations rest on several key assumptions for an idealized gyroscope. First, all deflections are assumed small enough to remain in the linear range with beam elongation being negligible, so Euler beam theory is applicable. Second, all beams are identical in dimension and material properties. Third, the rotor can be considered absolutely rigid compared with the suspension beams, and hence rotor bending is negligible. This assumption is reasonable because the rotor is almost two orders of magnitude wider than the individual beams. Finally, although large residual stresses are present after fabrication, the serpentine suspensions provide some stress relief. The stress relief was intentionally implemented to lower natural frequencies and allow all critical natural frequencies to be dependent on beam dimensions.

Calculation of the spring constant for outside beams begins with the formulation of the deflection equation (9.17) and deflection slope equation (9.18) at the rotor-attached end of each beam.[44] If the rotor rotates through an angle θ, then the beam end attached to the rotor must have slope θ and deflect a distance $R_o\theta$. Both equations depend on the unknown force F and moment M applied by the rotor. Since Euler beam theory is purely linear, the deflection and slope induced by both applied force and moment can be combined linearly.

$$\frac{FL^3}{3EI_{bz}} + \frac{ML^2}{2EI_{bz}} = \frac{R_o\theta}{N} \tag{9.17}$$

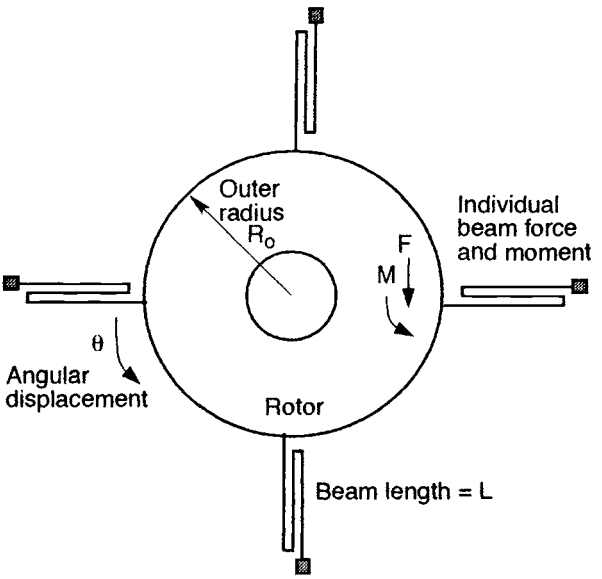


Fig. 9.12. Top view of polysilicon rotor showing dimensions and variable definitions for an outside suspension design.

$$\frac{FL^2}{2EI_{bz}} + \frac{FL}{EI_{bz}} = -\frac{\theta}{N} \tag{9.18}$$

Beam length L, outer rotor radius R_o, beam cross-sectional moment of inertia I_{bz}, number of serpentine folds N, and modulus of elasticity E all play a role. Although the thin-film polysilicon can exhibit large residual stresses, these are not included because all suspension designs implement stress relief. Incomplete tensile stress relief would result in higher spring constants than calculated.

The drive and sense axes may have different cross-sectional moments of inertia to intentionally mismatch natural frequencies. This is accomplished by changing the aspect ratio, which is defined as the ratio between beam height h and beam width b. The corresponding beam moments of inertia can be calculated as in Eqs. (9.19) and (9.20).[44]

$$I_{bz} = \frac{hb^3}{12} \tag{9.19}$$

$$I_{bx} = \frac{bh^3}{12} \tag{9.20}$$

Solving these two simultaneously results in Eqs. (9.21) and (9.22) for drive and sense-mode spring constants in terms of material and geometric parameters. Interestingly, increasing the radius with respect to suspension beam length results in dramatically higher rotational stiffness.

$$\text{Drive: } K_{zz} = 16\frac{EI_{bz}}{LN}\left\{3\left\langle\frac{R_o}{L}\right\rangle^2 + 3\left\langle\frac{R_o}{L}\right\rangle + 1\right\} \tag{9.21}$$

$$\text{Sense: } K_{xx} = K_{yy} = 8\frac{EI_{bx}}{LN}\left\{3\left\langle\frac{R_o}{L}\right\rangle^2 + 3\left\langle\frac{R_o}{L}\right\rangle + 1\right\} \tag{9.22}$$

9.3.2.3.3 Ratio Between Drive Mode Frequencies

As stated previously, the ratio of sense-spring constant to drive-spring constant should be half for successful mode matching. The fact that polysilicon deposition sets the beam height h while lithography sets the beam width b makes manufacturing well-matched modes extremely difficult. To overcome this difficulty, sense and drive modes are coarse-tuned using beam geometry and then fine-tuned using electrostatics. The only rotation mode frequency tuning that can occur is down-tuning of the sense-axis rotation mode. For this reason, it is advantageous to purposely mismatch modes with the sense mode 5–20% higher than the drive mode. This allows electrostatic down-tuning to equalize modes. The as-manufactured frequency ratio without electrostatic tuning can be shown to depend only on beam cross-section moment of inertia, which in turn is dependent only on the aspect ratio of the beam.

$$\frac{\omega_{Drive}}{\omega_{Sense}} = \frac{\sqrt{\frac{K_{zz}}{I_{zz}}}}{\sqrt{\frac{K_{xx}}{I_{xx}}}} = \frac{\sqrt{I_{bz}}}{\sqrt{I_{bx}}} = \frac{b}{h} \tag{9.23}$$

The ideal target mode-matching ratio between sense axis and drive axis can be selected using the aspect ratio b/h. For example, the ADI BiMEMS process had a fixed polysilicon thickness or height of 2 μm, so the beam width was made 1.8 μm to attempt a sense mode mismatch of plus 10%.

9.3.2.3.4 Translational Spring Constant Derivation

With all rigid rotational mode frequencies investigated, attention is now focused on the translational modes. Ideally, the translational mode frequencies will be much higher than the rotational mode frequencies. Higher frequencies attenuate rectilinear accelerations such as shocks and vibrations, which will alter scale factor or interrupt operation. In addition, the structure must have strength enough to withstand electrostatic forces due to sensing voltages. As discussed later in this section, it is the z-axis translation mode frequency that dictates both the maximum modulation voltage on the Coriolis sense electrodes and the maximum electrostatic frequency tuning range.

The z axis, which is perpendicular to the substrate, is the most critical for surface micromachines. One of the serious problems plaguing early surface micromachining was stiction between the substrate and the free-standing polysilicon structure. Unfortunately, the thin-film polysilicon used in a typical surface micromachining process provides little stiffness in the z-axis direction to prevent the mechanical structure from contacting the substrate. This tendency is further aggravated by the low-damping, intermediate vacuum environment in which the dual-axis gyroscope operates. Once in contact, the structures often adhere to the substrate, rendering the device inoperable. Many methods including the addition of small bumps to reduce contact area and antistiction monolayers have been used, but a stiffer z-axis mode increases the shock input needed to instigate contact between structure and substrate. Subsection 9.4.2 points out that external z-axis accelerations should not produce false Coriolis outputs caused by the differential sensing design. However, changing rotor height caused by accelerations will change the scale factor.

The x-axis and y-axis translation modes parallel to the substrate are naturally stiff because these involve beam compression and not beam bending. However, the z-axis mode is essentially a pure bending mode and therefore must be calculated. The exact same approach used for calculating the rotational spring constant can be used for calculating the translational spring constant K_{Tz}. The same assumptions regarding linearity, rotor rigidity, identical beams, and negligible residual stress effects used to calculate rotational spring constants are also assumed for the z-axis spring constant calculation. The identical beam assumption is important because this implies symmetry, which requires that the rotor remain parallel to the substrate. Beam-end displacement is replaced by the rotor axial displacement z. Beam-end slope is constrained to zero because the rotor is parallel to the substrate and does not deflect appreciably because of relative rigidity. A typical calculated value for the spring translational constant using Eq. (15-24) was 1 N/m for the ADI devices.

$$K_{Tz} = \frac{48EI_{bz}}{NL^3} \tag{9.24}$$

The actual z-axis translation mode frequency can be calculated as follows, with Ma representing the total rotor mass.

$$\omega_{Tz} = \sqrt{\frac{K_{Tz}}{Ma}} \tag{9.25}$$

A typical calculated value was 12 kHz for the ADI fabricated devices. This implies that the ADI devices should withstand a static 9000 g acceleration before touching the substrate. In general, suspension schemes outside the rotor will have much lower z-axis translation mode frequencies for given angular drive and sense mode frequencies compared with suspension schemes inside the rotor. In this case, a more robust outside suspension was used because of residual stress warpage.

9.3.3 Improving Sensitivity Via Electrostatic Frequency Tuning

Underetching, overetching, and polysilicon thickness all change suspension beam aspect ratio and thereby alter the frequency matching. Electrostatic tuning must be used after fabrication to equalize mode frequencies. The main drawback of this electrostatic tuning approach is a decrease in maximum mechanical shock survivability if the frequency tuning force is significant. Understanding this effect is important because all voltages including sense and drive voltages down tune natural frequencies.

9.3.3.1 Electrostatic DownTuning of Sense Mode Frequencies

The rotor is essentially a parallel plate suspended by the beam springs above the substrate. Any voltage applied between the rotor and the four quarter-pie shaped electrodes beneath the rotor will result in an electrostatic force. Because the electrostatic force as chosen is always attractive, the electrical force between the rotor and the substrate-mounted electrodes acts opposite the restoring force of the beam suspension. As the rotor moves closer to the substrate, the electrostatic force becomes stronger. Likewise, as the rotor moves away from the substrate, the electrostatic force becomes weaker. Tilting motion also exhibits the same behavior as the edge tilted closer to the substrate is pulled down, while the edge tilted away from the substrate has less force acting upon it. Hence, the electrostatic force acts like an unstable or "negative" spring working to force the rotor away from steady-state equilibrium. For small motions, the forces may be linearized about the rotor equilibrium position. The total x-axis torsional spring constant $\hat{K}_{xx}$ is the sum of the positive mechanical constant K_{xx} and the negative electrostatic spring constant K_{ex} as shown in Eq. (9.26). Equation (9.27) shows that the y-axis torsional spring constant $\hat{K}_{yy}$ has the same formulation as a result of symmetry.

$$\hat{K}_{xx} = K_{xx} - K_{ex} \tag{9.26}$$

$$\hat{K}_{yy} = K_{yy} - K_{ey} \tag{9.27}$$

This shows that sense springs and therefore sense natural frequencies can be down tuned. Since the effective electrostatic spring can never be positive, the sense frequencies must be designed via beam aspect ratio to be higher than the drive frequency. After fabrication, the voltage applied between the sense electrodes and the rotor can be adjusted to down tune the sense frequencies until they nearly match the drive frequency.

The linearized spring constant, K_{ex}, is defined as the derivative of the electrical torque M_x with respect to rotor tilt angle ϕ. This can be rewritten in terms of the physical parameters—the applied voltage V and the second derivative of each quadrant sense-electrode capacitance C_s. The linearized spring constant for the x and y axes are shown in Eqs. (9.28) and (9.29). Calculation of the sense-capacitance partial derivatives is left to Subsec. 9.4.2.3. The calculation is strongly dependent on the air gap spacing between rotor and substrate. Since the gap decreases with applied voltage, the second derivative is also a function of applied voltage.

$$K_{ex} = \frac{\partial M_x}{\partial \phi} = \frac{\partial^2 C_s}{\partial \phi^2} V^2 \tag{9.28}$$

$$K_{ey} = \frac{\partial M_y}{\partial \psi} = \frac{\partial^2 C_s}{\partial \psi^2} V^2 \tag{9.29}$$

9.3.3.2 Tuning Range Set by Pull-down Voltage

Equations (9.28) and (9.29) show that the total spring constant can be down-tuned to any value simply by increasing the applied voltage. Physically, this is not the case, since there is a limit to how high the voltage can be set. As the voltage is increased, the rotor will lower toward the substrate in the z-axis direction until the suspension force exactly balances the electrostatic force. However, there exists a voltage, termed the pull-down voltage, V_p, beyond which the suspension z-axis restoring force is too weak to balance the electrostatic force. At the pull-down voltage the electrostatic force overpowers the z axis, restoring force of the springs, and the rotor snaps down to the substrate.[45] As a result, the z-axis translational mode frequency sets the maximum tuning voltage, which in turn sets the maximum rotational sense mode tuning range. In fact, less than 10% x- and y-axis torsional sense-mode tuning was possible before the rotor was pulled to the substrate.

As shown in Fig. 9.13, there are two opposing forces acting on the gyroscope rotor. One is the suspension spring restoring force F_{Spring}, which works to keep the rotor at equilibrium, and the second is the electrostatic force $F_{Electro}$, which works to pull the rotor to the substrate. The strength of these forces can be approximated using the rotor distance from substrate z, the voltage applied between rotor and sense electrodes V_s, and physical characteristics of the rotor. Equation (9.30) shows the suspension spring force with the original gap g and the z-axis translation spring constant K_{Tz}.

$$F_{Spring} = K_{Tz}(g - z) \tag{9.30}$$

Equation (9.31) shows the electrostatic force impinging on the rotor in terms of rotor area A and the dielectric coefficient of vacuum ε. This relation assumes that the inertial rotor is ideally flat and that no outside accelerations are altering the rotor position.

$$F_{Electro} = -\frac{1}{2}\frac{A\varepsilon V^2}{z^2} \tag{9.31}$$

These two opposing forces can be plotted against rotor distance from the substrate z as in Fig. 9.14. Changing the voltage applied between the rotor and the sense electrodes shifts the electrostatic force upward for higher voltages. For low voltages, the force curves intersect at two points where the suspension and the electrostatic forces are equal. One of the equilibrium points is stable. The rotor stays at this point unless perturbed. The other equilibrium point is unstable. A rotor parked at this unstable point close to the substrate is susceptible to pull-down by the dominant electrostatic force. Any shock that forces the rotor within this range will halt operation. Above this critical distance, the rotor will snap back to the stable equilibrium point. At very high voltages, there is no force curve intersection, and the electrostatic force always dominates, pulling the

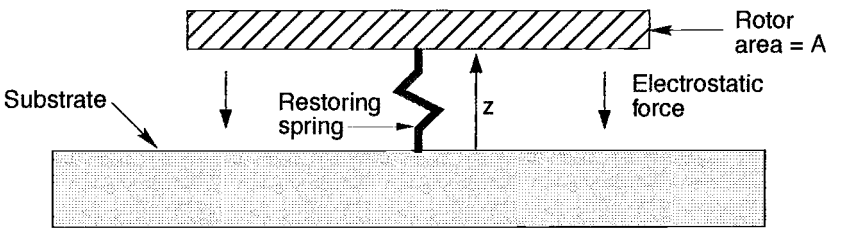


Fig. 9.13. Conceptual side view of the rotor acted upon by both the suspension-spring restoring force and the electrostatic attractive force.

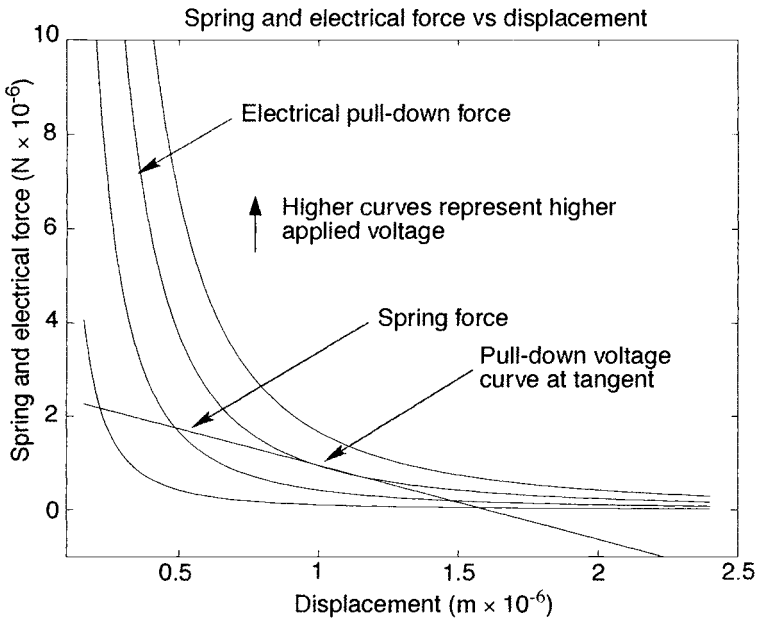


Fig. 9.14. Suspension-spring and electrostatic force curves plotted with respect to z-axis displacement.

rotor to the substrate. In between these two extremes lies a critical pull-down voltage where the suspension can just balance the spring. At the pull-down voltage, only one equilibrium point exists, and the two force curves are tangent. This results in a metastable equilibrium point with zero spring constant and a zero eigenvalue.

This critical pull-down voltage is the absolute maximum RMS voltage that can be applied between the rotor and the sense electrodes. Hence, the pull-down voltage limits the rotational x- and y-axis frequency tuning range as well as the maximum sense modulation voltage. This critical pull-down voltage can be calculated by solving two equations simultaneously for two independent variables. The first is the force balance equation (9.32), which expresses the equilibrium of equal and opposite forces—Eq. (9.33). This formulation assumes that the mechanical suspension is linear over the range of motion and that the rotor acts as an ideal parallel capacitor ignoring fringing fields. The rotor height above the substrate at pull-down is z_p, while the actual pull-down voltage is V_p.

$$F_{Spring} + F_{Electro} = 0 \tag{9.32}$$

$$K_{Tz}(g - z_p) = \frac{1}{2}\frac{A\varepsilon V_p^2}{z_p^2} \tag{9.33}$$

Equation (9.33) embodies the important observation that the two force curves are tangent at the pull-down voltage. Therefore, the slopes of the two curves are equal. This is shown by Eq. (9.34), and with substitutions, Eq. (9.35).

$$\frac{\partial F_{Spring}}{\partial z} = \frac{\partial F_{Electro}}{\partial z} \tag{9.34}$$

$$K_{Tz} = \frac{A\varepsilon V_p^2}{z_p^3} \tag{9.35}$$

All variables in the above equations (9.32–9.35) are known except for the pull-down voltage, V_p, and the rotor distance from the substrate when pull-down occurs, z_p. Solving the equations using substitution recovers the rotor distance above the substrate z in terms of the original as-fabricated gap g. Increasing voltage pulls the rotor down until the gap is two-thirds the original zero voltage gap. At this point, pull-down occurs. This critical distance ($z = 2/3\ g$) can be substituted back into the force balance equation, and the pull-down voltage can be derived as in Eq. (9.36).

$$V_p = \sqrt{\frac{8g^3 K_{Tz}}{27A\varepsilon}} \tag{9.36}$$

This pull-down voltage sets the absolute upper limit on voltage that can be applied for frequency tuning and Coriolis motion sensing. In practice, a lower voltage must be used, since at the critical pull-down voltage, the total spring constant is zero, and any perturbation will result in the rotor snapping down. It is therefore advantageous to choose a desired down-tuning voltage, V_d, below the critical pull-down voltage that provides stability, robustness, and survivability to outside accelerations. The ratio of electrostatically tuned versus the original manufactured z-axis translational spring constant K_{Tz} can be plotted as in Fig. 9.15 for the ADI design. The horizontal axis shows the voltage applied between rotor and sense electrodes. The top graph shows the z-axis translation spring constant drop with increased voltage until it reaches zero. At this zero point, pull-down occurs with applied voltage equaling approximately 2.4 V for the ADI designs. Experimental data showed that the actual pull-down voltage was in fact 3.4 V for ADI designs. The derivation gives an approximate lower estimate with residual stress, spring hardening, fringe

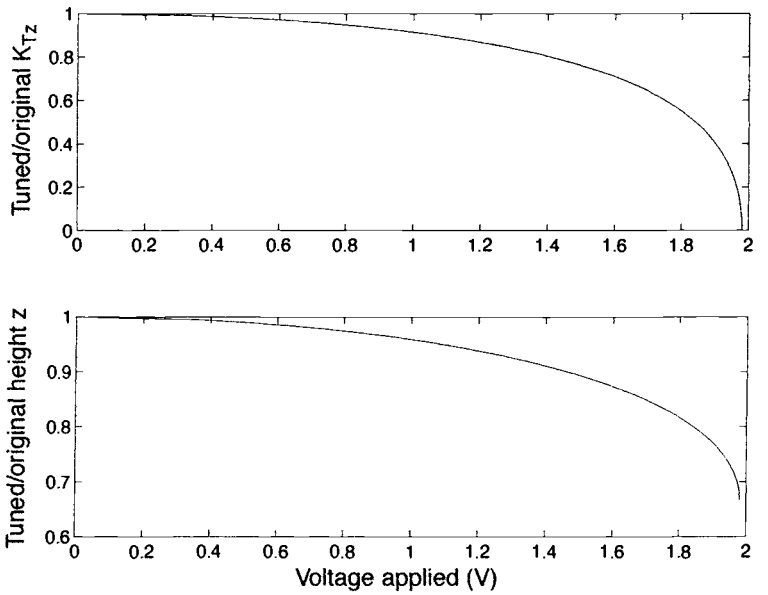


Fig. 9.15. The top graph shows the ratio between new and original translational spring constants vs applied voltage. The bottom graph shows new vs original rotor height vs applied voltage.

capacitances, and rotor warpage ignored. The lower graph shows the ratio between new gap and the original manufactured gap between rotor and substrate. As expected, the gap decreases with increased voltage as the spring deflects to balance the electrostatic force. Eventually, electrostatic force dominates, and pull-down occurs when the new gap measures two-thirds the original gap.

Using the desired voltage, V_d, required for the desired z-axis translational spring constant, the maximum tuning range of the x- and y-axis sense modes, $\hat{\omega}_{\text{sense}}$, can be calculated. This actually overstates the tuning range, as the sense voltage required for electrical interface circuitry must be added. However, this approximates the maximum possible frequency tuning range of the rotational spring constant, as given in Eqs. (9.37) and (9.38).

$$\frac{\hat{\omega}_{\text{sense}}}{\omega_{\text{sense}}} = \frac{\hat{K}_{xx}}{K_{xx}} = \sqrt{1 - \frac{K_{ex}}{K_x}} \tag{9.37}$$

$$K_{ex} = \frac{\partial^2 C_s}{\partial \phi^2} V_d^2 \tag{9.38}$$

9.3.4 Minimum Detectable Signal as a Result of Brownian Noise

Maximizing sensitivity by frequency tuning is important for performance, given there is electrical interface noise. However, even if the circuits were ideal and contributed no noise, an ultimate minimum detectable signal would exist, as dictated by Brownian noise.[46] Just as electrical resistors have thermal or Johnson noise, mechanical "resistors" or dampers have Brownian noise. In this intermediate vacuum case, air damping acts like a source of white noise force as air molecules randomly impinge on the mechanical structure. The ultimate limit in hard vacuum where air damping becomes negligible is set by the structural damping of polysilicon. In either case, the resulting angular torque or moment root power density M_n can be derived in terms of the damping coefficient C_{xx}, temperature T, and Boltzmann constant k_B as revealed in Eq. (9.39).

$$M_n = \sqrt{4k_B T C_{xx}} \tag{9.39}$$

This torque appears as a false Coriolis acceleration and is noise that sets the absolute minimum detectable rate signal for the gyroscope. Converting the Brownian noise torque into an equivalent false rate input is shown in Eq. (9.40) by setting the Brownian torque noise equal to the Coriolis acceleration term from Eq. (9.3), given in Subsec. 9.3.1.1.

$$I_{zz}\Omega_n\dot{\theta} = M_n \tag{9.40}$$

The expression for the drive oscillation rate $\dot{\theta}$, $\theta_0\omega_z Cos(\omega_z t)$, can be substituted. The sinusoid spreads the input rate signal over twice the original bandwidth, but careful demodulation aliases only 1/2 the noise power to base band. This allows the calculation of the equivalent false rate input caused by Brownian noise and is given in Eq. (9.42).

$$I_{zz}\Omega_n\theta_0\omega_z Cos(\omega_z t) = M_n \tag{9.41}$$

$$\Omega_n = \frac{\sqrt{4k_B T C_{xx}}}{I_{zz}\theta_0\omega_z} \tag{9.42}$$

This equation can be converted into a more useful form involving the quality factor Q as shown in Eq. (9.43).

$$\Omega_n = \frac{2}{\theta_0}\sqrt{\frac{k_B T}{I_{zz}\omega_z Q_x}\left(\frac{\omega_x}{\omega_z}\right)} \tag{9.43}$$

Despite operation in a moderate vacuum (50–100 mtorr), the main source of damping is still air viscosity and not structural damping. At lower pressures the damping would be reduced, and therefore Brownian noise would also be reduced. However, at very low pressures the structural damping of the polysilicon becomes the dominant effect, and no more improvement can be achieved without structural or material alteration. Actual gyroscope testing, at approximately room temperature with an operational pressure of 60 mtorr, resulted in a Q of approximately 1000. This resulted in an estimated Brownian noise level of approximately 0.5 deg/$\sqrt{\text{h}}$ for the first-generation dual-axis gyroscope fabricated by ADI. Unfortunately, the performance level never reached the Brownian noise level because of electrical noise from the interface circuitry.

9.4 Electrical Interface and Signal Processing

The dual-axis gyroscope requires both interface and signal processing circuitry. The overall system-level schematic is pictured in Fig. 9.16. First and foremost, the gyroscope must be driven into rotational oscillation with a constant amplitude. This is accomplished by the trans-resistance amplifier in combination with automatic gain control (AGC). Next, a capacitive measurement scheme must detect Coriolis motion in both orthogonal input axes and then produce useful electrical signals. High-frequency digital sense voltages applied to the quadrant sense electrodes in conjunction with an integrator attached to the rotor provide Coriolis sensing. These electrical sense signals must undergo signal processing to produce the final rotation rate output signals. Finally, all signal processing, modulated voltages, and AGC must be coordinated by a master clock. Therefore, a phase-locked loop is used to slave a voltage-controlled oscillator to the gyroscope drive frequency. Using digital circuit building block units, the required digital signal processing and modulation signals are created. The BiMEMS versions described in this chapter have all crucial circuitry on-chip. Later versions not discussed here, which were fabricated by Sandia National Labs, had all signal processing circuitry on-chip.

9.4.1 Resonant Drive

There can be no rate sensing unless the inertial rotor is in motion. To this end, the inertial rotor is driven into rotational resonance about the z axis. An electrostatic comb drive is used, as it has the

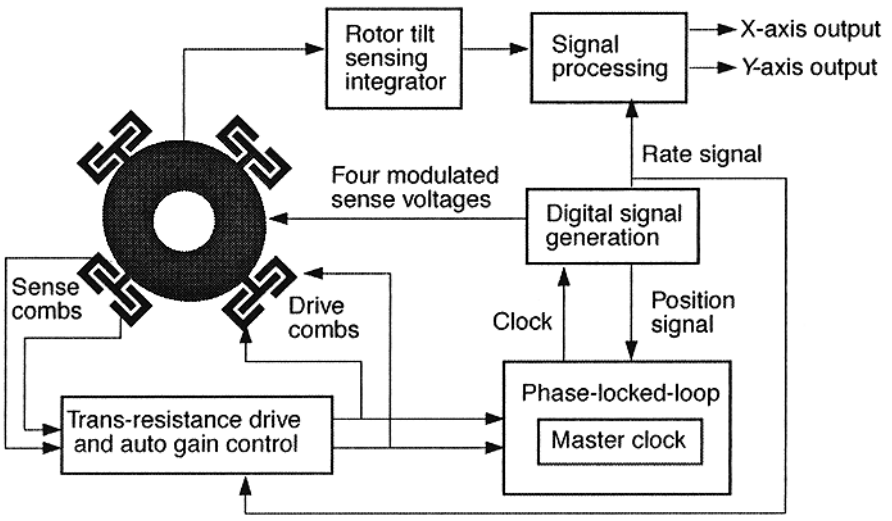


Fig. 9.16. System-level schematic showing entire dual-axis gyroscope electronics.

distinct advantage of remaining linear despite large displacements. As discussed by Nguyen,[34] a trans-resistance amplifier provides positive feedback, which essentially cancels viscous damping and thereby induces self-resonance. A phase-locked loop locks on to the electrical drive voltage to provide a clean signal for amplitude control and signal processing. The resonant amplitude is carefully controlled using automatic gain control because the scale factor is directly proportional to the resonant amplitude. The diagram in Fig. 9.17 depicts the entire drive system. Each of these important subsystems will now be discussed in turn.

9.4.1.1 Electrostatic Comb Drive

Comb electrodes are used both to impart a torque to the rotor and to sense the rotation rate of the rotor. One set of combs is dedicated to providing driving torque, while a second set is dedicated to providing rotation rate measurement. Electrostatic comb drive was chosen over other electrode configurations such as parallel plate because comb electrode sensing and forcing remain linear despite a degree or more of angular displacement.[33] This is vital for gyroscope actuation since the target angular displacement of one degree results in several micrometers of rotor motion

The drive subsystem is fully differential. In other words, half the drive combs point clockwise, while the other half point counterclockwise. Likewise, half the sense combs point clockwise, while the other half point counterclockwise. This provides a differential measurement of the rotation rate and a differential torque applied to the rotor. Differential comb-drive circuitry is crucial for functionality and performance. The usual arguments involving less susceptibility to outside interference and power-supply variations are important but are not the primary reasons for the differential design. The most important argument stems from the fact that the rotor is used as the sense node for Coriolis motion detection, as detailed in Subsec. 9.4.2. Therefore, the sensing measurement node is affected by the comb-drive voltages via the air gap capacitance between drive comb fingers and the rotor. This results in a drive-voltage feedthrough and an extraneous motion current in the critical Coriolis sensing circuitry. If these parasitic signals become large, then the Coriolis sensing circuitry can be corrupted as a result of distortion and amplifier saturation. Differential drive reduces the size of these parasitic signals, but they cannot be totally eliminated.

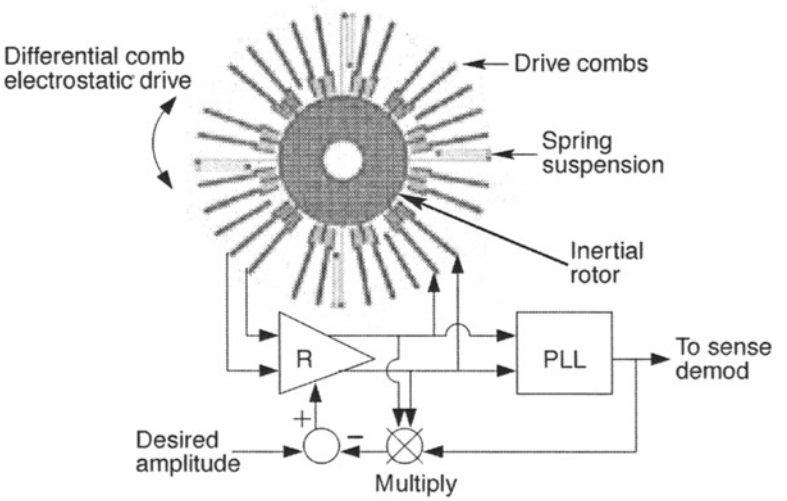


Fig. 9.17. System-level schematic of the trans-resistance resonant drive with automatic gain control plus phase-lock-loop for clean signal generation.

To ensure that the rotor will indeed oscillate, two key comb electrode parameters must be calculated. These parameters are the torque generated by a given drive voltage and the current output generated by a given rotation rate. The generated torque can be calculated by considering the potential energy E of the rotor and drive combs with voltages potential between them. This can be written as shown in Eq. (9.44) in terms of the voltages V_{cw} and V_{ccw} applied across the air gap capacitance between the rotor and the respective clockwise C_{cw} or counterclockwise C_{ccw} drive combs.

$$E = \frac{1}{2}(C_{cw}V_{cw}^2 + C_{ccw}V_{cw}^2) \tag{9.44}$$

Invoking the definition of potential energy, the resulting torque T can be calculated by differentiating the potential energy with respect to rotation position as shown in Eq. (9.45).

$$T = \frac{dE}{d\theta} = \frac{1}{2}\left[\frac{\partial C_{cw}}{\partial\theta}(V_{dc} + v_{cw})^2 + \frac{\partial C_{ccw}}{\partial\theta}(V_{dc} + v_{ccw})^2\right] \tag{9.45}$$

The dependence upon the square of voltage can be linearized by assuming a large constant dc voltage V_{dc} and smaller, variable ac drive voltages v_{cw} and v_{ccw}. Noting that the counterclockwise and clockwise combs are identical except for physical direction, the torque T in Eq. (9.45) can be simplified. As the rotor rotates θ, the clockwise combs increase in capacitance while the counterclockwise combs decrease in capacitance. Hence, the derivative of air-gap capacitance with respect to θ is equal and opposite for the two sets of combs. Further simplification results from using a differential drive signal such that $v_{cw} = -v_{ccw}$. The opposite voltages and opposite partial differential yield cancellation and the torque simplifies to Eq. (9.47).

$$T = \frac{1}{2}\left[\frac{\partial C_{cw}}{\partial\theta}\left(V_{dc}^2 + 2V_{dc}v_{cw} + v_{cw}^2\right) + \frac{\partial C_{ccw}}{\partial\theta}\left(V_{dc}^2 + 2V_{dc}v_{cw} + v_{cw}^2\right)\right] \tag{9.46}$$

$$T = 2\left[\frac{\partial C_{cw}}{\partial\theta}V_{dc}\right]v_{cw} \tag{9.47}$$

The air-gap capacitance can be estimated using the parallel-plate approximation. The total plate area is determined by rotor and comb finger overlap, while the capacitor gap is simply the distance between comb fingers and the rotor. This ignores fringe capacitance and therefore results in a conservative estimate of torque. The air-gap capacitance is estimated in Eq. (9.48), where N is the number of air gaps on clockwise combs, ε is the permittivity of vacuum, h is the polysilicon height, d is the comb air gap between rotor and stationary combs, R_i is the radius from gyroscope center to the inner most air gap, R_o is the radius from gyroscope center to the outermost air gap, and θ_0 is the nominal angle of overlap.

$$C_{cw}(\theta) = \frac{N\varepsilon h}{d}\left(\frac{R_o + R_i}{2}\right)(\theta_0 + \theta) \tag{9.48}$$

Taking the partial derivative with respect to rotation angle (Eq. 9.49) and substituting back into the torque equation (Eq. 9.47) finally reveals the torque generated by a given drive voltage in Eq. (9.50). The key to comb-drive linearity is that the partial derivative of overlap capacitance with respect to rotation angle be constant. Assuming adequate time overlap in travel, drive torque has no dependence upon the rotor angular position.

$$\frac{\partial C_{cw}}{\partial\theta} = \frac{N\varepsilon h}{d}\left(\frac{R_o + R_i}{2}\right) \tag{9.49}$$

$$T = 2\left[\frac{N\varepsilon h}{d}\left(\frac{R_o + R_i}{2}\right)V_{dc}\right]v_{cw} \tag{9.50}$$

The output current generated by the rotor rotation rate is derived from basic principles. The starting relationship is the simple linear capacitor equation (9.51), which relates the charge on the clockwise drive combs Q_{cw} with the applied voltage. In this case, the capacitance is the overlap capacitance between the rotor and stationary sense combs, and the voltage is the dc sense voltage. The voltages and comb dimensions are assumed to be identical for both electrostatic sense and drive combs.

$$Q_{cw} = C_{cw}V_{dc} \tag{9.51}$$

Since current is the time derivative of charge, the output current can be determined by taking the total time derivative of the charge stored as a result of the voltage applied between the rotor and stationary sense combs as shown in Eq. (9.52). Counterclockwise and clockwise currents are opposite, so the differential current i is twice the clockwise current alone.

$$i = \frac{dQ}{dt} = 2\left[\frac{\partial C_{cw}}{\partial\theta}V_{dc}\right]\dot{\theta} \tag{9.52}$$

As shown previously, the partial derivative of the overlap capacitance with respect to the rotor rotation angle is constant. By fixing the sense voltage, V_{dc}, constant in time, the output current is directly proportional to rotor rotation rate $\dot{\theta}$.

9.4.1.2 Positive Feedback for Resonance

Positive feedback is used to induce rotor self-oscillation at the drive-mode natural frequency. This technique is used in many oscillators, including quartz crystals and micromachined tuning forks.[35] In short, the positive feedback cancels the inherent viscous damping of the rotor. This leaves the system unstable, and the rotor begins to self-oscillate. In this case, the feedback path starts with the sense combs generating a current proportional to rotation rate. This current is converted to a voltage by a trans-resistance amplifier. Finally, the voltage is applied to the drive combs, which induce rotor rotation. The trans-resistance approach was chosen over lower noise designs such as the Pierce oscillator scheme because the trans-resistance method was previously demonstrated by Nguyen[34] and allows variable gain.

The key to successful operation is ensuring that the combs and trans-resistance amplifier are designed with enough feedback loop gain to overcome viscous damping. The calculation begins with Eq. (9.53), which represents the obvious requirement that the viscous damping in the drive axis must be canceled by the resonant drive torque.

$$T = C_{zz}\dot{\theta} = \frac{I_{zz}Q_{zz}}{\omega_z}\dot{\theta} \tag{9.53}$$

Now the resonant drive torque must be calculated given a rotation rate. As stated previously, the rotation rate generates a sense-comb output current, which is converted to a voltage by the trans-resistance amplifier and then fed back to the drive combs. Assuming the rotor resonant frequency is well below the first pole of the amplifier and that feedthrough is minimal, the transfer functions are all frequency-independent algebraic expressions. Substituting the comb transduction expression and adding trans-resistance ohmic gain R_Ω, the torque can be shown to be as in Eq. (9.54).

$$T = 2\left[\frac{\partial C_{cw}}{\partial\theta}V_{dc}\right]R_\Omega i = 4\left[\frac{\partial C_{cw}}{\partial\theta}V_{dc}\right]^2 R_\Omega\dot{\theta} \tag{9.54}$$

Replacing the resonant torque in Eq. (9.53) with the above expression gives the combined design equation for comb structures and trans-resistance amplifier. The equation can then be solved for the resistance ohmic gain R_Ω in Eq. (9.55).[34] Equation (9.56) recasts Eq. (9.55) by including geometry terms, finger numbers, and physical constants for inspection of design trade-offs.

$$R_\Omega = \frac{C_{zz}}{4\left[\frac{\partial C_{cw}}{\partial \theta} V_{dc}\right]^2} \tag{9.55}$$

$$R_\Omega = \frac{C_{zz}}{4\left[\frac{N\varepsilon h}{d}\left(\frac{R_o - R_i}{2}\right) V_{dc}\right]^2} \tag{9.56}$$

9.4.1.3 Trans-Resistance Amplifier Design

The trans-resistance amplifier must satisfy several design criteria for successful operation and improved performance. As mentioned earlier, the trans-resistance amplifier must have adequate gain at the rotor drive natural frequency to induce self-oscillation. This implies that a high dc gain with the first electrical pole well above the drive-mode natural frequency is required. It is also important to have minimal phase lag at the drive frequency. Since the trans-resistance output is used to synchronize the master phase-lock-loop (PLL) clock, any phase lag will reduce performance of the final signal processing. As will be discussed in Subsec. 9.4.1.6, this phase lag will exacerbate offset drift resulting from rotor wobble or quadrature error. Typically, higher gain results in lower bandwidth and greater phase lag, so there is a maximum practical trans-resistance gain for any given technology. The gain may also be limited by destabilizing capacitive feedthrough from drive to sense combs. An important feature of the trans-resistance amplifier is that a variable gain must be provided. This is required for oscillation amplitude control.

9.4.1.4 Resonant Amplitude Control Via AGC

Once the rotor is self-oscillating, oscillation amplitude will grow unchecked until a nonlinearity fixes the amplitude. This results in a very large, distorted, and uncontrolled amplitude. Because the gyroscope mechanical scale factor is directly proportional to oscillation amplitude, it is of paramount importance to accurately fix the oscillation amplitude constant. Hence, AGC is needed to measure the amplitude and adjust the trans-resistance amplifier gain. The underlying principle of AGC is that if the oscillation amplitude is above the desired level, then the trans-resistance gain is decreased. Likewise, if the oscillation amplitude is below the desired level, then the trans-resistance gain is increased. The feedback system that accomplishes this task is the subject of Fig. 9.18.

The trans-resistance amplifier produces a sine wave with an amplitude proportional to physical rotor oscillation amplitude. This sinusoid is converted into a usable dc voltage, which is proportional to the oscillation amplitude, by first chopping it with the PLL generated rate signal and the low-pass filtering. The resulting dc voltage becomes proportional to the oscillation amplitude and is compared with the desired amplitude voltage. The error signal from this comparison is fed back into the variable gain trans-resistance amplifier to correct the oscillation amplitude. This synchronous demodulation requires a digital signal generated by the PLL.

9.4.1.5 Phase-lock-loop Lock-in to Resonance

Both Coriolis signal processing and AGC require demodulation with the resonant drive signal. Analog multiplication of critical signals with the resonant drive signal is not preferable because of the small signal size and the large noise inherent to trans-resistance amplifiers. PLLs are often

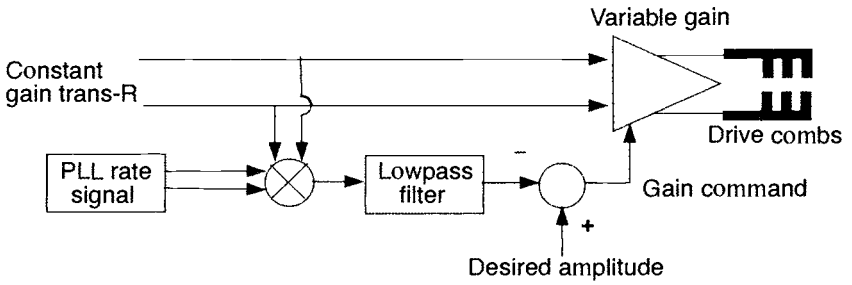


Fig. 9.18. Overview diagram of the AGC circuitry used to fix the rotor resonant amplitude constant.

used to surmount these difficulties and provide a clean digital signal.[47] The heart of the PLL is the voltage control oscillator or VCO, which produces a clean digital output squarewave at the same frequency as the incoming trans-resistance signal. In this design, the VCO is used as the master clock to produce all modulation and demodulation signals. Using digital logic, a rate signal is produced that is in phase with the driven rotor angular rate. This signal is used for automatic gain control and Coriolis signal demodulation. A second digital signal termed the position signal is at the drive natural frequency, but is 90 deg out of phase with the trans-resistance drive and can be used for the PLL as well as quadrature error cancellation. Finally, two high-frequency signals used for Coriolis motion sensing are also generated, as shown in Subsec. 9.4.2.

9.4.1.6 Quadrature Error

As with all rotating bodies, the dual-axis gyroscope rotor will exhibit some wobble caused by rotor moment of inertia imbalance as well as suspension asymmetry and other imperfections. This wobble is in phase with rotor position, so it is at the exact same frequency as the Coriolis tilt motion. The only difference between the desired Coriolis motion and the corrupting wobble motion is that they are 90 deg out of phase. Therefore, theoretically the wobble can be removed using demodulation with the orthogonal PLL-generated rate signal. In practice, however, the rate signal will exhibit some phase lag. This allows a component of the wobble motion to be perceived as a false Coriolis motion. This false signal is termed quadrature error. Typically there might be only 1–5 deg of phase lag, but because the wobble is a gigantic motion compared with the minute Coriolis tilt, the quadrature error can be significant. Rough experimental measurements show the wobble to be approximately 100–500 ppm (parts per million) of the drive motion, which translates into a quadrature error of 2–10 deg/s if electronic phase lag is merely 1 deg. The quadrature error appears as a dc offset, which theoretically could be calibrated out. Yet the quadrature error magnitude depends on second-order manufacturing imperfections and as such may tend to cause drift. Rate gyroscope measurements are often integrated to estimate angular heading. Therefore, any drift in offset causes an unbounded angular heading error, which increases linearly with time. The designs tested thus far have attempted to overcome quadrature error and the associated offset through careful trans-resistance amplifier design, but future designs may rely on directly canceling wobble.[48]

Any mechanism that alters phase relations will accentuate quadrature error, including distortion due to nonlinearities. Since the electronics are not perfectly linear, all electrical interface circuitry may exhibit distortion. Distortion may reduce the phase difference between the Coriolis and the wobble signal, thereby corrupting the output signal and making it difficult to separate the Coriolis signal. In particular, odd order harmonic terms in electronics gain stages alter the wobble phase such that the Coriolis signal is corrupted. The wobble motion is massive compared to the Coriolis motion, so odd order distortions should be avoided at all costs.

9.4.2 Coriolis Motion Sensing

With the resonant drive system fully explained, attention is turned toward the actual sensing of the Coriolis tilt motion. This requires a capacitive measurement scheme to transduce the mechanical motion into a change in the capacitance. Interface electronics to detect the change in the capacitance, and signal processing circuitry to create a final output rate-measurement voltage.

9.4.2.1 Capacitive Detection of Coriolis Motion

It was previously shown that a rotation rate input induces a Coriolis acceleration, which in turn induces a tilting oscillation. Using a differential capacitance measurement scheme to detect the tilt oscillation amplitude, the original rotation rate input can be inferred. To this end, four quadrant electrodes are patterned using $n+$ diffusions underneath the rotor. One pair of diametrically opposed electrodes provides x-axis tilt sensing, while the second pair provides y-axis tilt sensing. Figure 9.19 shows the rotor with only one pair of quadrant electrodes, for simplicity. These electrodes form a capacitive divider with the inertial rotor. If the rotor tilts as shown in Fig. 9.19, then the capacitance of one sense capacitor increases while the capacitance of the other decreases. This differential change in capacitance is detected via an integrator electrically connected to the inertial rotor in conjunction with a modulated sense voltage applied between the pair of quadrant electrodes.

The dual-axis rate gyroscope is equally sensitive to rotation rates about both the x and y axes. Differentiating between x- and y-axis input rotation rates is accomplished by differentiating between the orthogonal x- and y-axis inertial rotor tilt oscillations caused by the Coriolis acceleration. Because all tilt-oscillation detection is done using a single integrator that is electrically connected to the structure, electrical differentiation between x- and y-axis tilt oscillation is accomplished by using a different sense-modulation frequency for each axis. Separate demodulation circuits for each axis provide two separate output signals proportional to the two orthogonal rotation rate inputs (Fig. 9.20.).

As the multiple signal-frequency bands shown in Fig. 9.21 suggest, choosing sense-modulation frequencies demands great care. Clearly, the sense-modulation voltages should have a far higher frequency than the inertial rotor resonance to avoid mixing with drive feed-through, double frequency motion current, and any drive signal distortion. Since the integrator output signals for each axis are actually double modulated, the detection signals for each axis appear as double sidebands spaced equally distant about each respective modulation voltage frequency. The

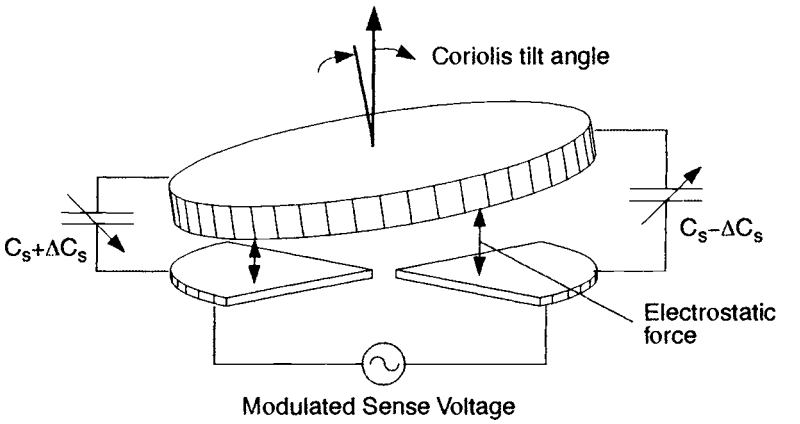


Fig. 9.19. Side view showing rotor with underlying diffusion sense electrodes for one sense axis.

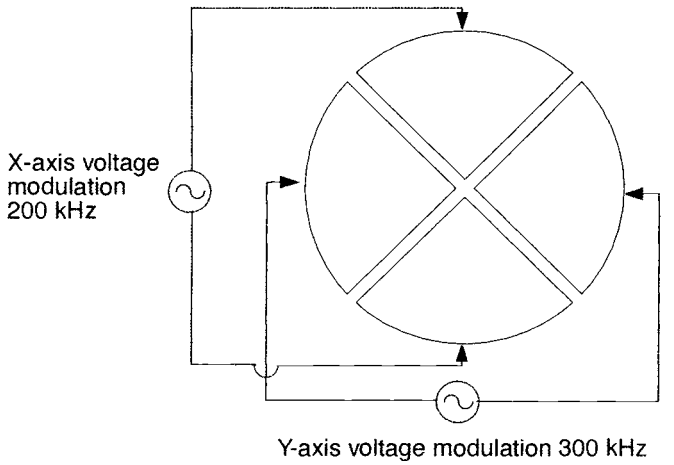


Fig. 9.20. View of the quadrant sense electrodes beneath the inertial rotor. Note the different modulation frequencies for each axis

frequency difference between each side band and the original sense modulation voltage is equal to the drive frequency (28 kHz in this case) because the tilt oscillation induced by Coriolis acceleration is at the drive frequency. These side bands should never mix, so the sense voltage modulation frequencies have a minimum separation of greater than twice the drive resonant frequency. In addition, higher harmonics and intermodulation resulting from distortion can interfere with signal purity, thereby further constraining frequency choice. The sense-modulation frequencies 200 kHz and 300 kHz were chosen because they satisfied all these requirements and were within the electrical circuitry bandwidth. Both of these sense-modulation voltages are created from the high-frequency VCO signal generated in the phase-locked-loop. The base VCO signal is approximately 2.4 MHz during normal operation. This high-frequency signal is divided to produce the sense-modulation signals. For example, dividing the VCO signal by 12 yields a 200 kHz signal, while dividing by 8 yields the 300 kHz signal.

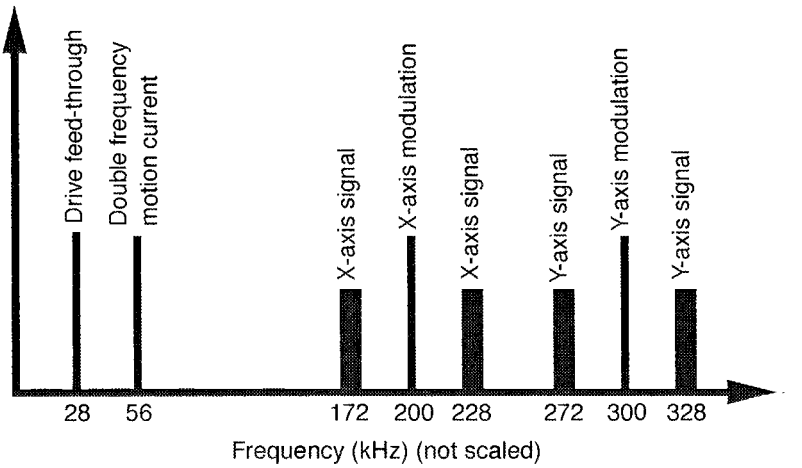


Fig. 9.21. Conceptual power spectral density of all important dual-axis rate gyroscope signals. Note signal frequencies are chosen to avoid overlapping between each sideband, and the original sense-modulation voltage is equal to the drive.

9.4.2.2 Interface Circuitry and Signal Processing

The capacitance change resulting from rotor tilt is measured using an integrator. The integrator holds the rotor at a constant voltage rendering the rotor as a virtual ground. If the sense capacitors are equal, then the modulation voltage causes charge to flow between the sense capacitors with no charge escaping through the integrator. If the sense capacitors are unbalanced because of inertial rotor tilt, then some charge flows through the integrator producing a measurement voltage proportional to tilt displacement and modulated at the sense-voltage frequency. The integrator is realized by placing a small integrating capacitor of approximately 50 fF in feedback around a high-gain operational amplifier. The front-end dc voltage level was set by placing a subthreshold MOSFET (metal-oxide-silicon field-effect transistor) and diode in parallel with the integrating capacitor. This provided a high enough input impedance for effective capacitive measurement, while providing a low enough resistive path to ground for setting of the rotor bias voltage.

As shown in Fig. 9.22, the voltage output from the integrator must be demodulated twice to recover the desired voltage output signal. The first demodulation removes the sense-voltage modulation frequency leaving a voltage proportional to the inertial rotor tilt position. The second demodulation removes the inertial rotor drive frequency, leaving a baseband voltage signal proportional to the rotation rate input. The second demodulation uses the PLL generated rate signal, which is in phase with both the trans-resistance drive signal and the desired Coriolis signal. In practice, this double demodulation was accomplished by premultiplying the digital demodulation signals and then using a single chopping of the integrator signal and a low-pass filter.

Because all tilt oscillation detection is completed using a single integrator electrically connected to the structure, electrical differentiation between x- and y-axis tilt oscillation is accomplished by using a different sense-modulation frequency for each axis. Two nearly exact copies of the demodulation signal-processing circuitry are integrated on-chip. The only difference between the circuit copies is that one demodulates with the x-axis sense-modulation frequency, while the other demodulates with the y-axis frequency. Demodulation circuits for each axis provide two output voltage signals proportional to the two orthogonal rotation rate inputs.

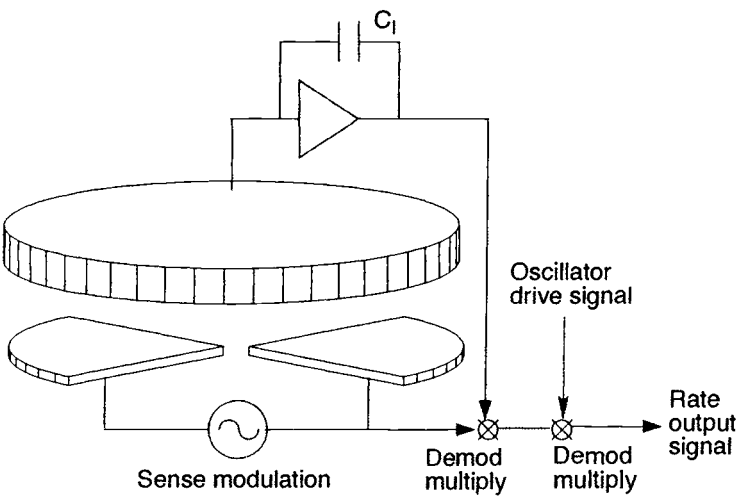


Fig. 9.22. Side view and schematic showing the sense integrator and the signal processing required to produce final rotation rate measurement for one axis.

9.4.2.3 Electrical Sensitivity to Rotation Rate Inputs

The electrical sensitivity to rotation rate input relates the voltage output to a rotation rate input. The calculation of this quantity can be separated into two distinct sensitivities. Mechanical sensitivity and electrical sensitivity to rotor displacement combine to give the electrical sensitivity to rotation rate inputs. It is important to find the combined sensitivity because mechanical sensitivity or electrical sensitivity to Coriolis displacement alone can be misleading. (See Fig. 9.23.)

The mechanical sensitivity was calculated previously, so only the electrical sensitivity to rotor displacement need be addressed here. The electrical sensitivity to rotor displacement relates voltage output to rotor angular deflection. The x and y axes are identical and ideally decoupled, so only the x-axis electrical sensitivity will be explored. The calculation begins with the change in sense capacitance, given an angular displacement of the rotor. Assuming a small angular displacement ϕ, the electrode sense capacitance can be approximated by the first two terms of the Taylor series for sense capacitance evaluated at zero tilting. Any positive angular deflection of the rotor causes one capacitor C_{sp} to increase in capacitance, while the other capacitor C_{sn} to decrease in capacitance. Thus the two separate sense capacitances can be written as Eqs. (9.57) and (9.58).

$$C_{sp} = C_{s0} + \frac{\partial C_s}{\partial \phi}\phi \tag{9.57}$$

$$C_{sn} = C_{s0} - \frac{\partial C_s}{\partial \phi}\phi \tag{9.58}$$

Because the integrator attempts to hold the rotor voltage constant, any imbalance in sense capacitors caused by rotor displacement will force charge onto the integrating capacitor. The foundation for determining this sense charge is the relationship between charge on a capacitor given capacitance and voltage. The charge in question is the sense charge q on the rotor.

In this case, the capacitance is the capacitance between the two electrodes and the rotor (C_{sp} and C_{sn}), while the voltage is the modulated sense voltage V_s, assuming the rotor is held at ground. The two other sense electrodes that measure rotation about the orthogonal y-axis direction can be ignored, because they exhibit no net change in capacitance resulting from rotation about the x axis. The sense voltage is differential, so the sense voltage on each sense capacitor is equal and opposite, as depicted by Eq. (9.59).

$$q = C_{sp}V_s - C_{sn}V_s \tag{9.59}$$

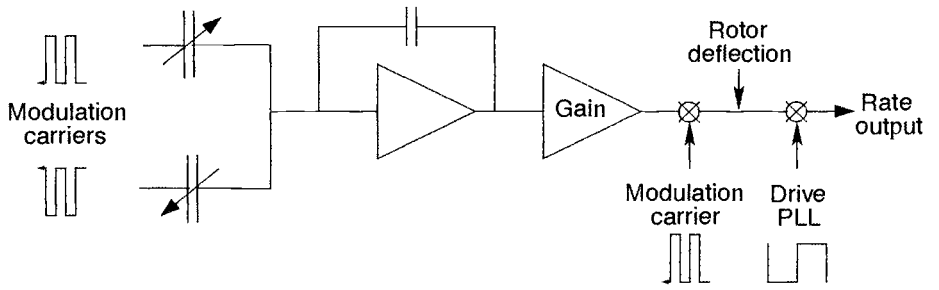


Fig. 9.23. Schematic showing circuit equivalent of the entire interface circuitry and signal.

Substituting the Taylor series approximation of the sense capacitance, Eqs. (9.57) and (9.58), into Eq. (9.59) gives the much simplified total charge on the rotor due to the x-axis sense voltages in Eq. (9.60). Note the constant sense-capacitance terms are canceled.

$$q = 2\left(\frac{\partial C_s}{\partial \phi}\phi\right)V_s \tag{9.60}$$

Therefore, the current flowing through the integrator is simply the time derivative of the charge on the rotor as defined in Eq. (9.61).

$$i = \frac{dq}{dt} = \frac{d}{dt}\left[2\left(\frac{\partial C_s}{\partial \phi}\phi\right)V_s\right] \tag{9.61}$$

The integrator, of course, integrates this current onto a capacitor C_I, which thereby produces a voltage. The voltage out of the integrator V_{int} can be written as shown in Eq. (9.62) and finally as in Eq. (9.63). This assumes the frequencies of the modulation sense voltage are high enough such that the interface circuitry behaves as an ideal integrator.

$$V_{int} = \frac{1}{C_I}\int_t i \tag{9.62}$$

$$V_{int} = \frac{2\left(\frac{\partial C_s}{\partial \phi}\phi\right)V_s}{C_I} \tag{9.63}$$

The voltage output of the integrator undergoes several transformations, including voltage gain, two demodulations, and final filtering. Both the demodulators and the filter have individual gains. To simplify representation without losing any important nuances, the effects of these final stages can be captured in an effective gain A_e. The final form of the electrical sensitivity to rotor displacement is given in Eq. (9.64).

$$\frac{V_{out}}{\phi} = \left[2\left(\frac{\partial C_s}{\partial \phi}\right)\frac{V_s}{C_I}\right]A_e \tag{9.64}$$

Now electrical sensitivity to rotor deflection can be combined with mechanical sensitivity to reveal overall electrical sensitivity.

The combined mechanical system, sense electrodes, and integrator can be summarized in Eq. (9.65) with sense voltage V_s, integrating capacitor C_I, and electrode capacitive sensitivity $(\partial C_s)/(\partial \phi)$.

$$\left|\frac{V_{out}}{\Omega_y}\right| = \left|\frac{\phi}{\Omega_y}\right|\left|\frac{V_{out}}{\phi}\right| = \left|\frac{2\theta_0\omega_z}{\omega_x^2 + \frac{j\omega_z\omega_x}{Q_x} - \omega_z^2}\right|\left[2\left(\frac{\partial C_s}{\partial \phi}\right)\frac{V_s}{C_I}\right]A_e \tag{9.65}$$

It is clear that increasing the sense voltage V_s between the rotor and the quadrant sense electrodes will increase overall sensitivity. Unfortunately, this sense voltage also results in an electrostatic force which tends to pull the inertial rotor down to the substrate. The critical pull-down voltage limits the maximum applicable sense voltage and is proportional to the natural frequency for z-axis translation perpendicular to the substrate. Conventional wisdom would suggest that mechanical sensitivity should be maximized by lowering the rotational natural frequencies. However, lower rotational natural frequencies result in lower z-axis translational frequency. For this

reason, the sense voltage for lower frequency suspensions must be reduced, nearly eliminating all mechanical sensitivity gains. In general, the integrator capacitor is set by minimum process capability or minimum size to avoid amplifier saturation caused by feedthrough and motion currents. That leaves the derivative of sense capacitance with respect to tilt angle as an important factor for improving sensitivity, which is discussed next.

So far, the electrical sensitivity has been documented in general terms. In order to find the numerical electrical sensitivity of the dual-axis rate gyroscope, the change in capacitance for a given change in rotor tilt angle $\partial C/\partial\phi$ must be derived. To estimate this quantity, first the actual capacitance between the rotor and a quadrant electrode is calculated, given a small x-axis tilt angle ϕ.

Ignoring fringe fields, a closed-form integral approximation can be found, but it is rather complex and unwieldy. However, the assumption of small tilt angles justifies linear approximation using a Taylor series expansion. The capacitance sensitivity to tilt angle can be written as shown in Eq. (9.66). The key results to notice are the cubic dependence on radius and the square dependence on the height of the rotor above the sense electrodes g. Thus capacitance sensitivity can be greatly enhanced by increasing rotor radius or by decreasing gap size. The sensitivity is additive, so calculating the differential for a solid rotor of outside radius R_o and then subtracting out a smaller inner radius R_i hollow center will give a hollow rotor design. The resulting equation also depends on the dielectric constant for air ε and the air gap g between the rotor and the quadrant electrodes.

$$\frac{\partial C_s}{\partial \phi} = \left(-\frac{0.471\varepsilon}{g^2}\right)(R_o^3 - R_i^3) \tag{9.66}$$

The calculated change in sense capacitance, given a change in rotor angle for the ADI fabricated gyroscope, was 5.3 pF/rad.

9.4.2.4 Electrical Noise Resolution Limit

Although the absolute minimum detectable signal is set by Brownian motion of the rotor, the actual minimum resolution of the dual-axis gyroscope is limited by electrical noise in the interface circuitry. Specifically, the input referred voltage noise at the front end of the sense integrator was calculated to be the dominant noise source. In this case, the dominant noise originates from the thermal noise of the polysilicon interconnects rather than the inherent noise voltage of the input MOSFET pair. There are many other noise sources, including trans-resistance drive noise capacitively coupling onto the structure, PLL phase noise allowing corruption of the Coriolis signal by quadrature error, and electrical noise in the generated sense modulation voltages, to name a few. The entrance points of these noise sources into the sensing system are shown in Fig. 9.24. However, these sources were calculated to be below the dominant resistive interconnect source. Later electrical stages in the Coriolis signal path add negligible noise as the signal magnitude is greatly increased by the initial integrator gain.

The minimum resolution can be found by equating the voltage output resulting from electrical noise with the voltage output of a hypothetical white spectrum angular rate input W_n. Integrating this hypothetical input noise power over the bandwidth will give the minimum detectable signal of the gyroscope. Converting a voltage output to a hypothetical rate input is completed easily by dividing the noise density of the output voltage by the electrical sensitivity as in Eq. (9.67). Results may vary by factors of 2, depending on the nature of the demodulation circuitry.

$$\Omega_n = \frac{v_n}{\left|\frac{V_{out}}{\Omega_y}\right|} \tag{9.67}$$

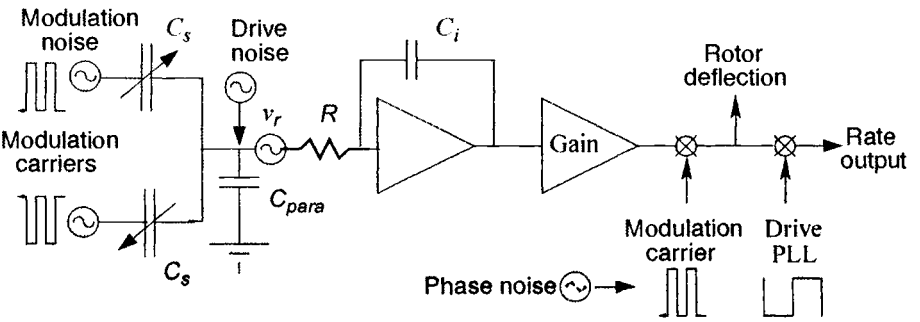


Fig. 9.24. Schematic diagram showing the dominant electrical noise source with all interface and signal processing circuitry degraded by this noise.

With electrical sensitivity already calculated in the previous subsection, the voltage noise output can be derived. Lemkin has shown[49] that for his accelerometer designs the dominant noise source using the ADI BiMEMS technology is the thermal noise generated by the resistance of the interconnect between mechanical structure and interface circuitry. The same result was found for the dual-axis rate gyroscope. This resistance includes both the polysilicon suspension beams, which the signals must flow through, and the signal runners from suspension anchor to the integrator input transistors. For reasonable transistor sizes and bias currents, the noise contribution of the interface amplifier is relatively small compared with the interconnect resistance-based noise. The combined structural and interconnect resistance was between 8 kΩ and 12 kΩ. The schematic in Fig. 9.24 shows the sense circuitry with the resistance and equivalent interconnect noise source v_r in front of the interconnect resistance R. The parasitic capacitance C_{para} between the rotor and the substrate sense electrodes also has a direct effect on noise performance.

Lemkin showed that the sense circuit schematic can be rearranged into the simplified equivalent circuit.[49] This allows direct calculation of the output voltage v_n by considering the v_r induced noise current, which is integrated on capacitor C_i, as shown in Eq. (9.68).

$$v_n = \frac{(2C_s + C_{para})A_e v_r}{C_i} \tag{9.68}$$

The thermal noise induced by any resistive load can be calculated[47] as in Eq. (9.69), where R is the interconnect resistance, T is temperature, and k_B is the Boltzmann constant. The typical noise calculated for ADI versions was approximately 11 $nV\sqrt{Hz}$.

$$v_r = \sqrt{4k_B TR} \tag{9.69}$$

The equivalent angular rate noise power density can also be calculated by substituting Eqs. (9.69) and (9.68) into the expanded Eq. (9.67). The result is shown in Eq. (9.70).

$$\Omega_n = \frac{(2C_s + C_{para})}{\left(\frac{\partial C_s}{\partial \phi}\right)V_s} \left| \frac{\omega_x^2 + \frac{j\omega_z\omega_x}{Q_x} - \omega_z^2}{2\theta_0\omega_z} \right| \sqrt{4k_B TR} \tag{9.70}$$

Equation (9.69) implies that equivalent angular rate noise caused by electrical interconnect resistance can be reduced by decreasing connection resistance R, decreasing all capacitances, increasing capacitive sensitivity, and increasing the sense modulation voltage V_s. The middle

term in the absolute value brackets models the mechanical dynamics. This is the inverse of the second-order dynamical model of gyroscopic motion. This term is minimized by matching sense and drive mode frequencies and increasing resonant drive amplitude θ_0. Assuming poor frequency matching, the ADI version rate noise power density was calculated to be 0.09 deg/s/$\sqrt{\text{Hz}}$. Improving the matching between drive- and sense-axis frequencies reduces the noise density. Actual experiments yielded noise levels several times higher than that calculated. Lower drive oscillation amplitude, higher interconnect resistance, and inexact capacitive sensitivity estimation are among the most likely contributors to error.

9.5 Design Trade-offs

The important aspects of the mechanical and electrical design have been described in the previous sections. Now the task remains to combine all these aspects into an optimized system that will perform to specifications within the constraints of process capability. This treatment will only consider the broad design decisions based on inertial rotor outside radius, rotor inside radius, and suspension beam length.

First, the performance specifications and robustness criteria must be considered. One obvious constraint stems from the need to put the inertial rotor in motion. Therefore, the gyroscope must be designed such that the trans-resistance amplifier with maximum resistance R_Ω can induce drive oscillation. The maximum trans-resistance value depends on transistor speed, transistor gain, and feedthrough. This suggests that the amplifier maximum gain is dependent on circuit technology, and optimization can be completed separately from mechanical design. The gyroscope clearly cannot operate if the rotor is pulled to the substrate by electrostatic voltages. To ensure survival during start-up transients and possible electrostatic discharges (ESD), a minimum pull-down voltage V_{pull} should be established. Survival following mechanical shock can be enhanced by determining a minimum natural frequency for all modes. This also aids in operational robustness to outside accelerations and vibrations. The minimum frequencies are either the drive frequency ω_z or the z-axis translation frequency ω_{Tz}.

Other design constraints follow directly from performance specifications. The most obvious is the minimum detectable signal, but there are several other interrelated factors. For example, the electrostatic tuning range should be chosen in an effort to provide frequency matching despite process variations. Ideally, the tuning range will be broad enough to allow mode matching within a matching percentage tolerance despite suspension beam height variance Δh and beam width variance Δb. Closely connected to the tuning range is the ratio between sense and drive mode frequencies. Mode separation should be wide enough to ensure the sense mode is above the drive mode frequency regardless of process variation. Yet the mode separation should not be so wide that frequency tuning cannot match modes. Finally, all of the above constraints should be satisfied efficiently through the use of minimum die area.

The outer rotor radius has paramount effect upon gyroscope performance and operation. A large radius obviously provides greater capacitive sensitivity, thereby improving electrical noise performance. A larger radius improves sense-axis frequency tuning range and so implies a better noise performance. Another benefit might be increased moment of inertia as this tends to lower the natural frequencies and increase sensitivity. On the other hand, there are many drawbacks to increasing the rotor radius. Aside from increased die area, the larger rotor area has several detrimental side effects. A larger area exposed to the substrate and sense electrodes produces higher parasitic capacitance on the rotor, which in turn increases the noise level. More area lowers pull-down voltage, which then requires lower sense modulation voltages thus reducing overall sensitivity. A larger rotor naturally implies a larger mass, which reduces translational natural

frequencies. The large rotor requires a more powerful drive to overcome viscous friction. Finally, in technologies with residual stress gradients, a larger rotor leads to exacerbated warpage. In general, the requirement for higher noise performance must be balanced with the drawbacks of using a larger radius rotor.

Many drawbacks plaguing a larger rotor radius can be partially surmounted by hollowing out the rotor center or, in other words, increasing the inside rotor radius R_i. It is true that hollowing out the rotor center will reduce capacitive sensitivity. However, the vast majority of sensitivity is derived from the outer rotor edge where the sense electrode area and the motion during tilting are greatest. Thus, hollowing out the center reduces sensitivity only slightly while imparting many crucial advantages. This is represented mathematically by the fact that sensitivity is proportional to R^3, while rotor area and the related drawbacks are proportional to R^2. With lower rotor area, the rotor parasitic capacitance is lower, thus reducing noise. Reduced area increases pull-down voltage, which in turn increases both allowable sense modulation voltage and tuning range. The hollow center also reduces mass, which increases the natural frequencies of translational modes. There is, of course, a limit to how large a hole can be etched in the rotor center, for at some point the mechanical rigidity of the rotor will be compromised. Still, great benefit is reaped from removing the center of the inertial rotor.

The ratio between suspension beam length to rotor radius L/R embodies the complex relationship between the rotor and the beam suspension. Shorter beams, and thus lower L/R, represent a stiffening of the mechanical structure. Clearly both rotational and translational frequencies will be increased with lower L/R. Interestingly, in the design range of interest, translational frequencies increase more in comparison with sense natural frequencies. Hence, for a given sense mode frequency, the translational modes are higher. This provides better mechanical shock protection, higher pull-down voltages, and larger frequency tuning range. Disadvantages include the fact that shorter beams will exhibit nonlinearity for smaller oscillation amplitudes and have less capacity for stress relief.

9.6 Future of Micromachined Gyroscopes

The field of micromachining is young, and the field of micromachined gyroscopes even younger. Future improvement in design and integration will allow performance and functionality far beyond that attained to date. Innovative technologies such as high-aspect-ratio MEMS made possible by new deep trench etchers may allow vast improvements in performance. Future directions include adding closed-loop feedback, which should allow future designs to benefit from the performance advantages of mode matching without sacrificing cross-axis sensitivity and scale factor stability. Other future directions include quadrature cancellation to reduce measurement offset drift and increased circuit integration. In addition to integrating circuitry, multiple sensors themselves can be integrated on the same substrate. There are some drawbacks to higher levels of integration, including added system complexity and possibly lower silicon die yields as a result of increased area. However, the benefits of simplified packaging, increased robustness, increased performance, smaller size, and presumably reduced cost are great. The art of system design will play a crucial role in deciding which circuit elements to place on-chip with the microsensors and which circuit elements should be realized off-chip using ASIC (application specific integrated circuits) or software.

Micromachined gyroscope size, weight, power consumption, and cost should be orders of magnitude below conventional technology. This will allow the common consumer access to technology once reserved for aerospace and the military. Society as a whole will benefit from improved automobile safety, personal navigation, virtual entertainment, and advanced

manufacturing. More importantly, micromachined gyroscopes may be the enabling technology for previously undreamed of applications. With suitable modifications, this industrial and commercial technology can be incorporated into miniature spacecraft and satellites at low cost. The dream of developing relatively inexpensive microspacecraft or nanosatellites that fit in the palm of a hand has moved a step closer to reality.

9.7 References

1. *Evaluation of Inertial Technologies and Inertial Systems Market Potential-Fourth Edition*, R. G. Brown Associates Inc., Hillpoint, WI (1991).
2. G. N. Smit, *Performance Thresholds for Application of MEMS Inertial Sensors in Space*, The Aerospace Corp. Report no. ATR-95(8168)-2 (1995), pp. 45–64.
3. R. S. Goetz, *Market Assessment for Space-Qualified MEMS*, ATR-95(8168)-2, The Aerospace Corporation, El Segundo, CA, pp. 7–26 (1995).
4. E. Y. Robinson, *ASIM and Nanosatellite Concepts for Space Systems, Subsystems, and Architectures*, The Aerospace Corp. Report no. ATR-95(8168)-1 (1995).
5. "Quartz Rate Sensor," QRS-11, Datasheet, Systron Donner, Concord, CA (1997).
6. "G-2000 Gyroscope," Datasheet, Litton, Salt Lake City, UT (1997).
7. M. Lemkin, B. E. Boser, D. M. Auslander, and J. H. Smith, "Three-Axis Force Balanced Accelerometer Using a Single Proof-Mass," *Proceedings of the 1997 Int. Conf. on Solid-State Sensors and Actuators* (Chicago, IL, June 1997), pp. 1185–1188.
8. M. Lemkin, *et al.*, "Three-Axis Surface Micromachined Sigma Delta Accelerometer," *Proceedings of the 1997 IEEE Int. Solid-State Circuits Conf., ISSCC (San Francisco, CA, February 1997), pp.* 202–203.
9. W. A. Clark, R. T. Howe, and R. Horowitz, "Surface Micromachined Z-Axis Vibratory Rate Gyroscope," *Tech. Digest, IEEE Solid-State Sensor and Actuator Workshop* (Hilton Head, SC, June 1996), pp. 283–287.
10. T. Juneau and A. P. Pisano, "Micromachined Dual Input Axis Angular Rate Sensor," *Tech. Digest 1996, IEEE Solid-State Sensor and Actuator Workshop* (Hilton Head, SC, June 1996), pp. 299–302.
11. T. Juneau and A. P. Pisano, "Dual Axis Operation of a Micromachined Rate Gyroscope," *Proceedings of the 1997 Int. Conf. on Solid-State Sensors and Actuators* (Chicago, IL, June 1997), pp. 883–886
12. T. N. Juneau, "Micromachined Dual-Axis Rate Gyroscope," Ph.D. thesis, University of California, Berkeley 1997.
13. J. H. Smith, *et al.*, "Embedded Micromechanical Devices for the Monolithic Integration of MEMS and CMOS," *Proceedings of the IEEE Int. Electron Devices Meeting* (Washington, D. C., Dec. 1995), pp. 609–612.
14. *Microelectromechanical Systems (MEMS): An SPC Market Study*, Systems Planning Corp., Arlington, VA (1994).
15. A. Jones, *et al.*, "*Microelectromechanical Systems Opportunities. A DOD Dual Use Technology Industrial Assessment*," Office of the Director of Defense Research and Engineering Report, December 95, 1995.
16. C. Song, "Commercial Vision of Silicon Based Inertial Sensors," *Proceedings of the 1997 Int. Conf. on Solid-State Sensors and Actuators* (Chicago, IL, June 1997), pp. 839–842.
17. S. Merhav, *Aerospace Sensor Systems and Applications* (Springer, New York, 1996).
18. P. Greiff, B. Boxenhorn, T. King, and L. Niles, "Silicon Monolithic Micromechanical Gyroscope," *Proceedings of the 1991 Int. Conf. on Solid-State Sensors and Actuators* (San Francisco, 1991), pp. 966–968.
19. J. Bernstein, *et al.*, *A Micromachined Comb-Drive Tuning Fork Rate Gyroscope, Proceedings of the 1993 IEEE Micro Electro Mechanical Systems* (Fort Lauderdale, FL, February 1993), pp. 143–148.
20. J. A. Breen, "A Path to Low-Cost Byroscopy," *Tech. Digest 1998, IEEE Solid-State Sensor and Actuator Workshop*, (Hilton Head, SC, June 1998), pp. 51–54.

21. M. W. Putty, "A Micromachined Vibrating Ring Gyroscope," Ph.D. thesis, University of Michigan, Ann Arbor, MI (1993).
22. S. R. Zarabadi, *et al.*, "An Angular Rate Sensor Interface IC," *Proceedings of the IEEE 1996 Custom Integrated Circuits Conf.* (San Diego, CA, 1996), pp. 311–314.
23. K. Tanaka, *et al.*, "Vibrating Silicon Microgyroscope," *Tech. Digest of the 13th Sensor Symposium* (Murata Mfg. Co., Yokohama, Japan, 1995), pp. 185–188.
24. M. Yamashita, *et al.*, "An X-shaped Tuning Fork Type Resonant Gyroscope by Silicon Micromachine Technology," Matsushita Electric Industrial Co., Osaka, Japan.
25. T. K. Tang, *et al.*, "Silicon Bulk Micromachined Vibratory Gyroscope," *Proceedings of SPIE, The Int. Society for Optical Engineering: Space Sciencecraft Control and Tracking in the New Millennium,* Vol. 2810 (Denver CO, 1996), pp. 101–115.
26. W. A. Clark, R. T. Howe, and R. Horowitz, "Z-Axis Vibratory Gyroscope," Paper presented at *Micromachining Workshop III*, Anaheim, CA, September 1996.
27. K. Y. Park, *et al.*, "Laterally Oscillated and Force-balanced Micro Vibratory Rate Gyroscope Supported by Fish Hook Shape Springs," *Proceedings of the 1997 10th Annual Int. Workshop on Micro Electro Mechanical Systems. MEMS* (Nagoya, Japan, 1997) pp. 494–499.
28. T. Brosnihan, *et al.*, "Embedded Interconnect and Electrical Isolation for High-Aspect-Ratio, SOI Inertial Instruments," *Proceedings of the 1997 IEEE Int. Conf. on Solid-State Sensors and Actuators* (Chicago IL, June 1997), pp. 637–640.
29. K. Shaw and N. MacDonald, "Integrating SCREAM Micromachined Devices with Integrated Circuits," *Proceedings of the IEEE 9th Annual Int. Workshop on MEMS* (San Diego, 1996), pp. 44–48.
30. T. Juneau, *et al.*, "Commercialization of Precision Inertial Sensors Integrated with Signal Processing," Paper presented at *Sensors Expo '98*, San Jose, CA, 1998.
31. P. Ljung and A. P. Pisano, *Micromachined Vibrating Gyroscope*, Research report, Berkeley Sensor and Actuator Center, University of California, Berkeley (March 1992).
32. R. S. Payne, *et al.*, "Surface Micromachining: From Vision to Reality to Vision," *Proceedings of the 1995 IEEE Int. Solid State Circuits Conf.* (San Francisco, CA, February 1995), pp. 164–165.
33. C. W. Tang, "Electrostatic Comb Drive for Resonant Sensor and Actuator Applications," Ph.D. thesis, University of California, Berkeley 1990.
34. C. T. Nguyen, "Micromechanical Signal Processors," Ph.D. thesis, University of California, Berkeley, 1994.
35. T. A. Roessig, A. P. Pisano, and R. T. Howe, "Surface Micromachined Resonant Accelerometer," *Proceedings of the 9th Int. Conf. on Solid-State Transducers and Actuators* (Chicago, IL, June 1997).
36. P. B. Ljung, T. Juneau, and A. P. Pisano, "Micromachined Two Input Axis Angular Rate Sensor," *Proceedings of the ASME Int. Mechanical Engineering Congress and Exposition*, session DSC-16 (New York, NY, 1995), pp. 957–962.
37. L. Meirovitch, *Methods of Analytical Dynamics* (McGraw-Hill, New York, 1970).
38. L. Meirovitch, *Analytical Methods in Vibrations* (Macmillan Publishing, New York, 1967).
39. P. B. Ljung, "Micromachined Angular Rate Sensor," Ph.D. thesis, University of California, Berkeley, 1997.
40. M. K. Andrews, et al., "A resonant Pressure Sensors Based on a Squeeze Film of Gas," *Sensors and Actuators A* **A36**, 219–226 (May 1993).
41. J. B. Starr, "Squeeze-film Damping in Solid-State Accelerometers," *Tech. Digest, 1990 IEEE Solid-State Sensor and Actuator Workshop* (Hilton Head, SC, June 1990) pp. 44–47.
42. C. L. Chen and J. J. Yao, "Damping Control of MEMS Devices Using Structural Design Approach," *Tech. Digest, 1996 IEEE Solid-State Sensor and Actuator Workshop* (Hilton Head, SC, June 1996), pp. 72–75.
43. Y. J. Yang, S. D. Senturia, "Numerical Simulation of Compressible Squeeze-Film Damping," *Tech. Digest 1996, IEEE Solid-State Sensor and Actuator Workshop* (Hilton Head, SC, June 1996), pp. 76-79.
44. J. M. Gere and S. P. Timoshenko, *Mechanics of Materials* (PWS-Kent, Boston, 1984).

45. M. W. Putty, "Polysilicon Resonant Microstructures," Master's thesis, University of Michigan, 1988.
46. T. B. Gabrielson, "Mechanical-Thermal Noise in Micromachined Acoustic and Vibration Sensors," *IEEE Transactions on Electronic Devices* **40** (5), 903–909 (May 1993).
47. P. R. Gray and R. G. Meyer, *Analysis and Design of Analog Integrated Circuits* (Wiley, New York, 1993).
48. W. Clark, "Micromachined Vibratory Rate Gyroscope," Ph.D. thesis, University of California, Berkeley, 1997.
49. M. A. Lemkin, "Micro Accelerometer Design with Digital Feedback Control," Ph.D. thesis, University of California, Berkeley, 1997.

10

MEMS-Based Sensing Systems: Architecture, Design, and Implementation

S. T. Amimoto,* A. J. Mason,† and K. Wise†

10.1 Introduction

The microelectronics industry has grown tremendously during the past few decades, largely because of increasing demand for microprocessors and memory. Advancements in microelectronics technology continue to meet the strong demand for more sophisticated and lower-cost electronics. In some areas of the industry, low-volume, custom ASICs (application-specific integrated circuits) are commonly being built to fulfill many of the world's electronic needs.

Rooted in microelectronics technology, microelectromechanical systems (MEMS)[1] has created a new industry of integrated sensors and actuators. MEMS technology has lowered the cost of these devices and opened new markets for MEMS transducers,[2] which will be increasingly important in providing a means for capturing information from the physical world and converting it to digital form.[3,4] As the MEMS industry grows, the trend is to combine transducers with increasingly sophisticated circuits[5–7] to form "smart" sensors. The low cost of signal-processing electronics (microprocessors, digital signal processors, etc.) makes it possible to join this circuitry with sensors[8] to form complete microsystems.

Microinstrumentation systems[9] that combine sensors, actuators, and signal processing circuitry on a common substrate are, in fact, now in development. These microsystems form autonomous units capable of gathering nonelectronic information, transducing the data to electrical signals, processing the information, making decisions based on it, and finally passing it on to other electronic systems that gain intelligence from the process.[10] Specifically for space applications, these functions may be used to monitor the environment experienced by a launch vehicle or satellite during assembly, transportation to launch site, storage, launch, and orbit. This information may be critical to aid identification of failure modes or anomalies during a failure investigation.

This chapter examines a MEMS-based data-logging system for a space application and presents two approaches to the development of multiparameter sensor (MPS) microsystems. The two microsystems have been designed as subsystems in vastly different macrosystems, yet they share many component-level functions and have similar goals. The first microsystem uses a "top-down" design approach: requirements of the macrosystem are established at the start. A survey of commercial components that may be used in an MPS system is provided to help the developer make choices for implementation. An implementation of this top-down system is described.

The second microsystem takes a "bottom-up" approach: requirements are largely determined by the needs of the transducers.The following topics are also covered.

- The architecture of the microsystem
- A review of current applicable MEMS technologies
- Design, fabrication, testing and calibration of the bottom-up-design microsystem
- Challenges faced during the design of each microsystem as sample procedures for MPS system development

*Mechanics and Materials Technology Center, The Aerospace Corporation, El Segundo, California

†Department of Electrical Engineering and Computer Science, University of Michigan

These two microsystems span a range of state-of-the-art design approaches, from those based on commercially available components to those based on advanced MEMS technologies. Designers can use these systems as guides to choose components and gain insight into the design and integration of an MPS system.

10.2 MEMS-Based Monitoring System for Space Applications

The European Space Agency Round-Table discussion[11] concludes that incentives to use micro/nanotechnology derive from its ability "to meet new or expanded flight performance requirements, recover from failure or a serious mission degradation event, increase reliability, and reduce overall system cost." MPS systems, for example, offer significant potential for decreasing the failure rates of U.S.-launched vehicle systems. Between 1984 and 1994, failure rates of DOD and non-DOD launches were 4.7% and 8.9% for 106 and 90 missions, respectively.[12] Of the 13 failures, five were attributed to the propulsion system. An MPS system can supply critical information to reduce delays in identifying the cause of a failure and to increase reliability and cost reductions on future flights.

On a Titan IV, one of the largest and best-instrumented launch vehicles, only 75 independent sensors can be monitored. Bandwidth restrictions, no data-processing, wires to each sensor, lack of small-sensor-form factors, and weight present severe limitations in the number of sensor locations that can be monitored. Some engine compartments are not monitored because of their distance from a central controller and difficulty in wiring. In addition, changing a sensor location/channel precipitates a significant analysis and qualification process that can cost up to $500,000. Small, low-cost, wireless, unobtrusive sensors could easily be moved without incurring this expensive process. These sensors will perform a large number of measurements at different locations, each fully instrumented with a set of sensors, and will lead to faster flight characterization of vehicle design and configuration. If these sensors preserve critical timing or phasing information, this would unambiguously identify where and when an abnormal event may have occurred.

Because of the low number of sensors and the high number of different flight-assembly configurations, the Titan IV-class launch vehicle has not been fully characterized despite its 10 years of flight history. Historically, data from each flight have indicated that the design margin was exceeded, often precipitating a study of design margin or requalification and testing to ensure that later flights will be successful. A single set of flight sensors costs $4 million without installation; furthermore, the hardware weighs 170 lb.[13] The fact that technology development has caused launch vehicles to evolve at a rate faster than they can be characterized is not unique to Titan. Only eight sensors are in use for the Delta rocket program, which suffered a spectacular failure on 17 January 1997 on flight K126.[14] A sizable effort was initiated to investigate potential causes of the failure. If the rocket had been instrumented with more sensor capability, the investigation may have been less costly.

MEMS-based sensors offer many advantages for the design of a multiparameter system for a launch-vehicle application. In particular, a "peel-and-stick" MEMS-based sensor system would have significant advantages over current flight environmental monitoring systems, including the following.

- Reliability is increased by a higher level of integration with a smaller chip count.
- Redundancy is made possible by low-cost sensors.
- Small-form factors result from the small sensor size and integration of supporting sensor electronics.
- Decreased size also leads to lower power consumption because of lower stray capacitance and thus a smaller battery requirement.

- With wireless chip sets, radio frequency (RF) or optical communications can reduce installation costs associated with conventional cables, a significant portion of the overall installation cost.
- With a microprocessor at each node, a networked MPS system offers capabilities far beyond merely reporting and routing data. Networking when combined with distributed intelligence will provide data compression, anomaly reporting, control or commanding operations, power management, transitions from one network mode to another, and various services such as error correction, time synchronization, data analysis, self-test, and autonomy.

The low cost of MEMS-based sensors enables their proliferation in many instrument or monitoring systems. As a result, data collection and node management become a paramount topic of concern. In very large node systems the cost drivers may not be the MEMS sensors, but the network data acquisition/management systems. On the other hand, without networking and other service capabilities, such as reliable reporting of high data rates from sensors such as accelerometers and acoustical devices, there becomes a need to incorporate autoconfiguration/modification protocols, which may in itself impede the proliferate use of sensors within a single system.

Advantages of an MPS system can be extended to military and commercial flights:

- It could readily carry out housekeeping on the Space Station by monitoring temperature, pressure, oxygen levels, and vibration and by detecting chemical vapors. (See Chapter 11).
- Monitoring Delta and other space-vehicle engines and satellite components in transport or storage could significantly improve reliability and increase confidence that they were not exposed to conditions exceeding their design or environmental specifications, a problem that has plagued nearly all satellite programs.
- An MPS system could assist NASA testing of a scaled version of a crew-return vehicle (CRV) to bring future Space-Station personnel back to Earth. The CRV is an unpowered but steerable reentry vehicle. The MPS system could monitor pressure and acceleration of the endo-atmospheric CRV flight tests being conducted at Edwards Air Force Base, California.
- A compact wireless version of an MPS system could have greatly facilitated data collection in ground testing of a newly fabricated Atlas payload transporter used to carry the payload for an Atlas rocket from its fairing encapsulation building to the launch pad. Test goals were to identify mode shapes and frequencies and to log peak acceleration events during a road test.

The number of space-related applications is growing, and the opportunities for the use of an environmental, data-logging MPS system are challenges awaiting system developers.

10.3 Macro System-Level Architecture

The macro-level architecture of an MPS is based on its requirements, availability of hardware and software, and the trade-offs facing the designer. These are discussed in the following sections.

10.3.1 Requirements for a Launch-Vehicle Multiparameter System

Discussion of the requirements that form the basis of a design for a multiparameter system will focus on the needs of an instrumentation system for the Titan IV, but may be extended to the space applications previously cited. Measurement parameters are taken from documentation of the Wideband Instrumentation System (WIS)[15] and Lift-off Instrumentation System (LOIS)[16] used on a Titan IV. These systems are based on hardwired sensors that are polled by a controller; the resulting digital data are formed into packets that are transmitted to ground using RF with a pulse-code modulated carrier at 2.2–2.4 GHz. Two unity-gain antennas, one on each side of the launch vehicle, are used in conjunction with a high-gain ground antenna. The vehicle is tracked to within a few degrees above the horizon, necessitating the use of an RF spectrum region that suffers little

atmospheric attenuation. The line-of-sight range is approximately 100 km. For the RF spectral region of 10 GHz and below, little attenuation is observed. Since maximum data rates are proportional to the carrier bandwidth, a sufficiently wide band near 10 GHz appears to be ideal.

A preliminary survey of the types of measurements employed by the WIS system is listed in Table 10.1. They are vibration, acoustic levels, acceleration, pressure, strain, and shock resulting from pyrotechnic firing events. Vibration and acoustics generally refer to the shaking that occurs due to the higher-frequency phenomena, such as turbulent flows across certain regions of the skin of the vehicle; acceleration refers to the lower frequency shaking and acceleration of the vehicle in the direction of motion. The bandwidths overlap as indicated in the table. The total launch event is 10 min, in contrast to the time spent to install the sensors (more than a week) and to await other vehicle preparation procedures (which can easily require a few months). The sensor system should be robust enough to accommodate the maximum amplitude and bandwidth level that a sensor might experience and to anticipate changes in the location of sensors without regard to the system's environmental qualification level. The measurement requirements and qualification levels may vary from location to location on the vehicle. The minimum and maximum temperature extremes to be encountered are −44 to +75°C,[17] a range readily met by most MEMS sensors.

Total raw data rates for the vehicle MPS system can be easily estimated:

$$\text{Measurement rate} = \sum_i \text{sensor rate}(i) \times \text{bandwidth}(i) \times \text{oversample factor}, \qquad (10.1)$$

where the summation is performed over all sensor types, i. For a given sensor it is unlikely that more than a single shock event will occur for the duration of the launch at a single location. Shock data may be acquired with a large bandwidth of 10 to 20 kHz, but its duration may be only 20 ms.

Table 10.1. Raw Characteristics and Requirements of an Environmental MPS System[18]

Measurement parameters (based on LOIS and WIS)[a] and sample rates at a single location (maximum amplitude values) • 3-axis vibration: 10–2000 Hz, ± 300-g max amplitude; 12 kHz • Acoustics: 10–4000 Hz, ± 185-dB max range; 8 kHz • Acceleration: 0–50 Hz, ± 10-g max amplitude; 100 Hz • Pressure: 0–50 Hz, 0–16-psia range; 100 Hz • Strain: 0–50 Hz, 900 min/in. or ± 2000 psi; 100 Hz • Number of sensor locations: 500 • Measurement time: 10 min • Shock event: 10 kHz for 0.02 s; 400 measurements/event • Elapsed time from installation to launch: several weeks to months
Communication: 2-way; range, ≥100 km; wireless 10-GHz telemetry band
Mountable with wire leads to specified attachment points as needed
Compact, low-weight, low-power form factor
Survivability and environmental parameters: vibration, shock, EMI, radiation, temperature, atmospheric to vacuum pressures, contamination resistant, humidity resistant
Self-configurable
Built-in test for installation and system health monitoring
Reliable data and system function

[a]LOIS: Lift-Off Instrumentation System, WIS: Wideband Instrumentation System

By using thresholding as a criterion to send data, the average data rate can easily be met by the communication system. Thresholding is a concept in which data may be measured but not acted upon unless a threshold value is exceeded. When this occurs, the data, for example, may be stored or reported as necessary. Thresholding acts as a filter to reduce the quantity of data while maintaining information and data quality. Data may need to be stored when aggregate peak data rates occur. The total raw data rate is calculated by summing over all sensor node rates:

$$\text{Total data rate} = \sum_{i} \text{sensor node rates } (i) \tag{10.2}$$

For the entire vehicle, total raw data rate acquired by all 500 nodes is 10 MHz. The bit rate is calculated from the total data rate. For a constant sample accuracy,

$$\text{Bit rate} = \text{total rate} \times (\text{bit accuracy} + \text{overhead}) \tag{10.3}$$

For data of 8-bit accuracy, we can assume 10 bits/sample (e.g., an additional 2 bits for parity checkbit and signbit) × 10 MHz, or 100 Mbits/s must be transmitted by the RF link to ground networking. Overhead for data packeting and error correction is neglected. The spectral efficiency of a communication system is often expressed by the ratio of the information carried per carrier-bandwidth frequency.

$$\text{Spectral efficiency} = \text{bit rate/carrier bandwidth} \tag{10.4}$$

Assuming a bandwidth of 200 MHz for the WIS, the spectral efficiency is 0.5 bits/s/Hz with no thresholding. This spectral efficiency is easily met by modern digital communication systems.[19] It is also assumed that some form of data compression will be used to reduce data transmission rates.

Sensor data below a certain threshold level will be of little value for anomaly resolution purposes unless there is assurance that the data reporting system is indeed functional. Otherwise the lack of data may indicate a nonfunctioning sensor or sensor node or other malfunctions. Thresholding while reducing data flow to "essential" data will introduce intermittent data reporting, which may be solved using time tagging. The nature of this effect may also depend on the flight profile, location of the sensor node, and the sensor type. In addition, data compression, such as that used by video compression schemes such as MPEG, JPEG, and Wavelet-based chip,[20] could also lead to intermittent data flow.

There may also be a latency reporting requirement. Latency can be defined as the time delay of the reporting system between time of measurement and the time these critical data are transmitted by the vehicle to the ground station. This time requirement can be set by considering the velocity at which a catastrophic phenomenon is able to propagate over a critical distance of the vehicle and to reach a component in the chain of the reporting system that will prevent critical data flow. One crude estimate of the maximum latency time may be derived from the distance of the edge of the solid-rocket booster to the center of the core vehicle, approximately 5 ft, and the speed of propagation of a blast wave at the speed of sound, 340 m/s, which yields a latency time of 4.5 ms.

A related issue concerns the nodes: unless there is confirmation that all nodes were functioning during the catastrophic event, no definitive conclusion can be made concerning the fidelity of the data. If data are not present, was the data amplitude below threshold or was the node not functional? Thus periodic confirmation of node integrity is needed if no data have recently been transmitted from that node. There will be significant value in the assurance that only a major vehicle anomaly could make the node nonfunctional.

Network self-configuration is an important feature for the launch-vehicle network and refers to the network's ability to organize itself. During installation, the configuration of the network architecture is changing. When the network is placed in active mode, the network configures itself by polling all nodes so that the master nodes will know which sensor nodes are within their compartments. Network self-configuration is also needed during the launch as each successive stage of the vehicle is ignited and then dropped off and the network undergoes a dynamic configuration change. Thus, self-configuration is an essential feature of the network architecture. For networking robustness, redundant network paths could be added to enable dynamic routing of messages and data around nodes that may have failed for whatever reasons.

Compartments where many sensors may be mounted are located throughout the launch vehicle and are shown in Fig. 10.1. The maximum compartment-to-compartment distance is between compartment 2A, where the transmitter for the telemetry system to the ground receiver is found, and compartment 1C at the base of the vehicle, a distance of 120 ft. The distance to the top of the payload fairing from compartment 2A is about 86 ft for the longest available payload fairing. These distances represent the maximum ranges for an RF transceiver on the vehicle. Distances within a compartment are typically on the order of 10 ft, where multiple-path reflections of RF signals may be anticipated because of extensive use of metal and the many vehicle parts. Outside the vehicle, relatively good lines of sight may be maintained, and little interference may be expected.

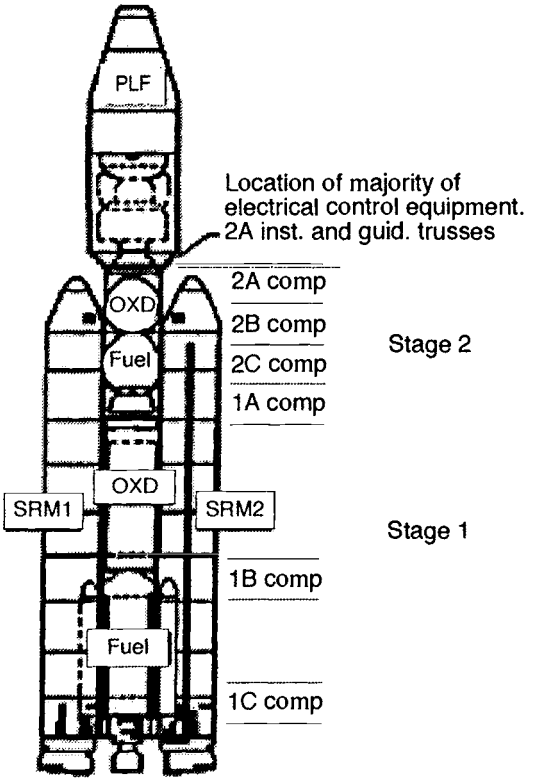


Fig. 10.1. Compartments of a Titan IV vehicle are labeled by stage numbers 1 or 2. Stage 1 uses the two outer solid-rocket motors and the lower liquid center-core stage. The second stage is liquid fueled. The length of the solid rockets is 112 ft, and the length of the payload fairing (PLF) is approximately 50–86 ft, depending on the upper stage and the payload configuration used.

10.3.2 Hardware/Protocols—Choices for MPS Implementation

Once requirements for a system are fixed, the designer must select specific implementations of hardware and software. Among various issues to be considered are performance specifications, amount of effort to integrate the many components into a system, networking protocols, the type of developmental tools, and whether the cost of the approach is within budget. The following surveys, primarily of commercial off-the-shelf (COTS) components, are provided to help the integrator developer with the many choices involved in this design. The components include sensors, microcontrollers, networking/device protocols, batteries, and communication devices.

10.3.2.1 Commercially Available Sensors

Surveys of available technologies provided a data base from which commercial sensors, accelerometers, and pressure sensors could be selected for the design of the MPS system. A large number of sensors are indeed available. Almost all sensors are in the form of sensing elements with analog output; very few are capable of providing direct digital output. Selection criteria required the sensors to be fabricated using MEMS methods. Examples of noncommercial sensors are discussed in Subsec. 10.6.3.

10.3.2.1.1 Commercial MEMS Accelerometers

A large number of accelerometers are commercially available, driven by the automobile airbag market. Results from a search of accelerometers and vendors are shown in Appendix 10. A at the end of this chapter. During the past year, accelerometer demand has outstripped production, resulting in temporary shortages. Competition from a large number of manufacturers has kept prices low. Individually packaged commercial COTS accelerometers in chip form may be purchased for as low as \$19–\$35 per sensor in small quantities (25 or below).

Accelerometer sensors are available in compact chip form or finished packaged assemblies with external mounting fixtures or environmental mounts and often with amplifier circuits. Some have customizable gain, offset settings, and printed circuit boards. The large weight and sizes of finished sensor packages relative to the internal die within are mostly associated with the final packaging, which addresses the interconnects and the physical integrity of the sensor package.[21] Thus repackaging of sensors with their supporting data-logging infrastructure, i.e., processors, memory, and communications, should lead to a significant reduction in weight and size of the completed package. These advantages are further discussed in Sec. 10.6.

10.3.2.1.2 Survey of MEMS Pressure Sensors

Pressure sensors produced by nano/micro fabrication techniques were surveyed. Many are available in 8-pin DIP packages. An abbreviated list of the sensors surveyed is shown in Table 10.2, which also lists only a few of the many configurations. In general, these sensors use the mechanism of the deformation of a membrane that is sensed by a strain gauge patterned on the membrane. Capacitive pressure sensors have been fabricated, as discussed in Subsec. 10.6.3.1. These gauges are incorporated into a Wheatstone bridge to produce an analog voltage change as a function of pressure. The table is intentionally short and does not reflect the wide variations in features that manufacturers offer. A summary of the findings follows.

- Some vendors offer versions that provide a buffer amplifier with adjustable gain.
- Some models incorporate a small tube to allow a connection to a pressure or vacuum tube.
- Chips from different vendors are often pin-for-pin identical, reflecting the level of competition.
- Pressure is usually sensed through a small hole on the chip.
- Power is typically under 150 mW.

Table 10.2. Typical MEMS Pressure Sensors for Low-Pressure Applications

Vendor	Model No.	Supply Voltage (V)	Input Range (psia)	Size / Weight (in., in., in./g)[a]	Power (mW)	Cost ($)
IC Sensors[b]	1431-015-A	3	0–15	0.3, 0.3, 0.14/0.30	2.0	20.0
Si Microstructures[c]	5310-015-A-P	5	0–15	0.3, 0.3, 0.14/0.30	7.5	12.5
Lucas Nova Sensors[d]	NPC-410-015-3L	3–15	0–15	0.3, 0.3, 0.14/0.25	7.5	20.0

[a]Dimensions in inches, mass in grams.
[b]EG&G IC Sensors, Milpitas, California
[c]Silicon Microstructures, Fremont, California
[d]Lucas Nova Sensors, Fremont, California

- Bandwidth can be as high as 1 kHz.
- Gauges are available with full-scale pressure ranges from 0.15 to 4000 psi.
- Pressure sensors are available in many models reflecting the attachment interface, compatibility with fluids under measurement, and electronic mounting options such as surface, DIP, and transistor cans.
- Prices (as of 1997) are often under $20 each in low quantities.

10.3.2.2 Survey of Microcontrollers and Data-Logging Systems

Output from most sensors is analog voltage. These signals must be converted to digital form using analog-to-digital converters (ADCs) and ported into a processor for calibration, time stamping, thresholding, and data compression, using the application software. The processor must also perform networking services, such as communication-media access, time synchronization across the networked sensors, formation of data packets, network configuration control, autonomous configuration control of network nodes, and packet routing. These functions are often implemented in hardware called a microcontroller, generally a combination of an ADC and a processor. Some of the microcontroller's more significant attributes for a designer include power consumption during active and sleep modes, processing speed, flash memory (EEPROM) or program memory, number of ADC channels, form factor, number and type of communication ports, communication rates, and hardware/software support by the vendor.

All the controllers listed in Appendix 10.B (at the end of this chapter) are relatively small and have a correspondingly small power appetite. The table is not a complete listing, but rather includes representative devices available to system developers. Some entries are also stand-alone data-logging systems, such as, the PC 104, the TattleTale 8, the Honeywell TSMD, the University of Michigan Microcluster, and the Adcon Telemetry m-T device. Of these, the smallest COTS system is the m-T, useful for low-data-rate, low-power applications. The others are microcontrollers with many of the attributes desired in a data-logging system. Using packaged chips or compact microcontroller cards considerably reduces the effort to develop an instrumentation node. Some noncommercial devices are also included, such as the Honeywell Time Stamp Measurement Device (TSMD), the University of Michigan Microcluster Watch, and the Air Force Research Laboratory Advanced Instrumentation Controller (AIC), as examples of the direction and the state of the art in advanced microcontroller/data-logging designs. The University of Michigan Microcluster is discussed in detail in Secs. 10.6 and 10.7.

Device characteristics are highly variable. Physically, the largest data-logger system is the PC104 with a maximum dimension of 90 mm. The smallest is the University of Michigan Microcluster with a volume of 5 cc for the complete system of processor, sensors, and communication. Peak power consumption is also quite large for the PC 104 processor at 10 W compared to 50 mW for the AIC. The number of ADC inputs varies from 4 to 10, with sampling rates from 8 to 100 kHz. The number of I/O ports for communication and memory varies from 1 to 6. Finally, environmental specifications for each device in meeting a space application also vary. One will be used for deep space, and others have been flown in Shuttle missions. All space missions require meeting launch vibration specifications and a minimum level of radiation hardness.

The Apple Newton using the ARM processor is included because of the processor's high figure of merit for mobile computing, expressed as a ratio of processing capacity to power consumption. With a high figure of merit, a smaller battery size can meet mission processing requirements. In principle, despite the processor's high clock rate and processing power, optimizing clock speed can optimize processor power for each application. In general, electrical power scales with the clock speed for each processor. However, the corollary is not necessarily true, meaning processors of high maximum processing power do not automatically use higher electrical power. Future low-power networked processors will very likely share similar characteristics with the ARM processor where networking overhead must be supported.

10.3.2.3 Software/Networking Protocols

Software/networking selection is often determined by a number of criteria:

- Desired computational and networking tasks
- Language and quality of the library that supports the available processor
- Availability of such fundamental software as device drivers
- Level of implementation of networking and networking services
- Availability of technical support and services by vendors
- Ease of use determined by diagnostic tools and development time and tools.

Another motivating selection criterion is the extreme reluctance of a user to redevelop the software for networking control, especially if it is already implemented for a processor. Standardized software that drives networks is needed for open networking systems.[22] Hardware is explicitly configured for networking,[23] and includes LonWorks[24] and Controller Area Network (CAN).* Each is briefly discussed with few details. Most are intended to be used with wired buses.

The LonWorks consortium offers networking hardware and software. The neuron-node hardware provides all layers of networking but one, as implemented by a media-access CPU and a network CPU, which allows the user to concentrate primarily on the application layer in an application CPU. All CPUs are combined in a single Neuron chip. The protocol is an accepted standard for ASHRAE BACnet North American building automation standard, adopted by the European Forecourt Standards Forum for European fuel stations and accepted by the American Association of Railroads for pneumatic braking.

The Neuron chip requires 85 mW of power, and the application layer is programmed in the Neuron-C language. Compatible communications media, such as wired or wireless transceivers, are available from vendors, as are developmental tools for nodes or systems and network service tools. Available gateways, routers, repeaters, and network interfaces allow for multiple media use and connecting hosts. In addition, a portable LonTalk protocol is available royalty free for imple-

*See Internet sites <http://www.nrtt.demon.co.uk>, <http://www.dgtech.com>, and <http://www.kvaser.se> for CAN products.

mentation on any microprocessor. Typical messages or packets use 10 bytes/message; additional bytes may be added for more data. The LonWorks networking system is designed for low data rates, robust reliability, and control of network. High data rates can be accommodated using a faster communications medium or a separate communication applications layer. LonMark is a standard that has been adopted by a large industrial consortium that maintains interoperability between the hardware products, offers a means by which products can be interchanged, and maintains a plug-and-play integration even with custom-engineered components.

CAN is a shared broadcast bus with speeds up to 1 Mbits/s. The protocol comes in several flavors of differing levels of capability. It is based on sending messages or frames that can be varied in lengths of 0–8 bytes. The frame has an identifier that must be unique. The CPU for the more powerful FullCAN architecture can store several frames, updates the buffered frames, and marks them for transmission. If the identifier matches, the shared variable can then be examined. The object is to create a set of shared variables in the network. The CAN controllers available include Intel 82527, Phillips 82C200, NEC uPD72005, Siemens 81C90, and Motorola TOUCAN on 683XX devices. Developmental hardware is offered by numerous vendors.[24] Data rates on CAN are limited by the physical bus length and transit time necessary for error recovery. A bus length of up to 50 m is supported at 1 Mbit/s. A gate array implementation has been reported for a FullCAN Controller.*

10.3.2.4 Smart Transducers

Since 1993, industry and government have recognized the need to select a single, open, network-independent communication interface standard for smart transducers, where the transducer is defined as a sensor or an actuator. This ongoing effort in the United States is led by the Manufacturing Engineering Laboratory of the National Institute of Technology (NIST) and the Instrumentation and Measurement Society's Technical Committee on Sensor Technology TC-9 of the Institute of Electrical and Electronics Engineers (IEEE). The goal is to define a uniform approach to support multiple bus standards and to enable transition to most of the existing popular networks. The interface is digital, defines a standard transducer electronic data sheet (TEDS) and its data format, and defines an architecture with application software independent of protocol and technology. Smart transducers would be able to convert and process signals, compensate for environment, convert analog-to-digital or digital-to-analog signals, control logic, provide local memory (EEPROM), and identify themselves. Each transducer will be physically associated with TEDS.

Two draft standards for the Smart Transducer Interface† have been proposed, the Network Capable Application Processor Information Model, IEEE P1451.1, and the Transducer to Microprocessor Communication Protocol and Transducer Electronic Data Sheet, IEEE P1451.2. The benefit of the first standard is to provide a network-independent, neutral interface for the application processor. Once the application is developed by the vendor, the user would link the transducer application software with a driver library supplied by the vendor to enable a plug-and-play environment. To use the transducer with another network, the user would simply recompile and link with the library provided by the vendor. This concept of abstracting the application from hardware has been used successfully for Windows® and Postscript printers.

---

*Initec AG, Bielst. 10, CH-4104 Oberweil, Switzerland, ph +41-61-7169616.

†Both proposed smart transducer interface standards, P1451.1 and P1451.2, are available from the IEEE Standard Department, 455 HOES Lane, Piscataway, NJ 08855; telephone: 800.678.4333.

The second standard, IEEE P1451.2, describes the contents of TEDS, a digital hardware interface to access TEDS, read sensors, and set actuators. Measurement aspects are captured in a smart transducer interface module (STIM) and application-related aspects in a network-capable application. The interface allows for triggering of one or more transducers and a variable-date clock rate. A generic STIM consists of the transducer, signal conditioning, TEDS, and a microcontroller that implements the P1451.2 interface. TEDS, which describes the type of sensor, the operation, and the attributes of the transducer, is structured into five parts: meta, channel, calibration, application-specific, and extension. In meta-TEDS is the description of the data structure, worst-case timing, and channel-grouping data. In channel TEDS is the upper/lower range limits, physical units, warm-up time, existence of self-test capability, calibration mode, and trigger parameters. The calibration TEDS contains the calibration date, the calibration interval, and parameters for a multisegment model. The application-specific TEDS is self-evident, and the extension TEDS is used for future extensions.

STIM is hosted by a network through a network capable application processor (NCAP). The P1451.2 standard describes both the hardware and firmware interface residing in the NCAP side of the NCAP/STIM interface. This firmware includes the network protocol, the application firmware, and the STIM driver. Three network providers have developed IEEE 1451.2 drivers for the NCAPs, Allen-Bradley (DeviceNet), Echelon(LonWorks), and Honeywell Micro Switch (Smart Distributed System).

10.3.2.5 Power Supply

The primary energy source for many MPS systems will be batteries. "Primary" and "secondary" refer to low-cost commercial, single-use batteries and to rechargeable batteries, respectively. In the absence of permanent wires to provide power, other sources such as solar cells or RF power may be used for microsystems.

Battery attributes that designers must consider include voltage, capacity, energy density, size, cost, storage life, discharge current, degradation mode, and rechargeability. Recharging of secondary batteries must also consider current limits or temperature rise during charging and the source of this power. Other considerations include package reliability, lifetime limited by recharge cycles, operating temperatures, cell mortality, cell leakage current, etc. Batteries are available in a number of sizes and voltages. Cells must be stacked in series to achieve higher voltages. In addition, the discharge-voltage-versus-time curve must be considered, and appropriate power conditioning used to optimize stored-power utilization. If the cells show a continuous voltage drop during the discharge profile of the battery, then a switching regulator must be employed to preserve efficient conversion of stored energy to delivered energy at fixed voltages.[25] Up to 90% efficiencies are possible. For batteries that exhibit a flat discharge profile, a voltage regulator may be sufficient.

A brief summary of battery characteristics is shown in Table 10.3. Also, the 100% capacity rating for rechargeable batteries must be derated to allow for additional mass needed for the battery package surrounding the cells. These factors and the practice of not charging to full capacity to increase life cycle may reduce the energy/mass ratio by as much as a factor of five.[26] Since battery casing becomes a larger fraction of the mass as batteries become smaller, the derating factors likewise become higher. In general, lithium batteries have excellent energy densities but are only able to deliver power at low current. These will likely be the choice of a low-power MPS system. Higher current is possible for NiCd, Ni metal hydride, lead acid, alkaline, and thin-film batteries, all of which have significantly lower energy per mass. A comprehensive discussion of batteries is given in Chapter 6.

Table 10.3. Battery Types and Characteristics

Battery	Type	Cell Voltage (v)	Capacity (Wh/kg)	Discharge Current/Cell (mA)[a]
Alkaline[b]	Primary	1.50	336	NA
Lithium	Primary	2.80	260	NA
Lead Acid[b]	Secondary	2.10	252	200–5000
NiCd[b]	Secondary	1.35	244	10–1200
Ni metal hydride[b]	Secondary	1.35	278	120–3800
Lithium Thionyl Chloride[c]	Secondary	3.60	700	4–400
Thin film[d]	Secondary	3.80	100	1400–1700

[a]Typical discharge currents achievable. Values are dependent on size and operating temperature of battery.
[b]D. Linden, ed., *Handbook of Batteries*, 2nd ed. (N.Y.: McGraw Hill, 1995).
[c]Lithium Thionyl Chloride 3.6-V Batteries Technical Information, Tadiran, Ltd., Los Gatos, CA, June 1993.
[d]Bell Core polymeric mesh battery. (See Chapter 6.)

10.3.2.6 Communications

A large number of integrated circuits and circuit designs have been driven by the explosion in personal communication systems, such as cordless and cellular phones, handheld roaming inventory units, RF identification tags, and wireless LANs (local area networks). Many devices use digital methods of communication, and many commercially available chips and circuitry are driven by these applications. Some are listed in Table 10.4, together with derived communications requirements for the sensor master and ground transceiver. The reader is referred to a more detailed review and survey of spread-spectrum devices by Schweber.[27]

The user must be aware of potential problems in the design of a communication system. These problems include the effect of multipath signal channels due to reflections, Doppler frequency shifts, latency, bit-error rates, synchronization time, data-packet format, security, the pseudorandom noise (PN) code pattern to identify users, the modulation method used for spread-spectrum devices that define the "channel," the spectral efficiency of sending data per unit RF bandwidth, and finally, power consumption.

Elaborating further:

- Multipath reflections will often degrade the signal to noise at the receiver.
- Doppler shifts could reduce the coherency of the transmitted signals and cause a lowering of the signal to noise.
- Latency is the time taken to send and receive information in a propagated signal.
- The bit-error rates determine the required signal-to-noise ratio (S/N ratio) necessary at the receiver for each modulation method used. Often a bit-error rate of 10^{-6} is thought to be adequate for data, but for controls and commands a lower bit-error rate such as 10^{-8} is necessary. As one attempts to reduce transmitted power to conserve battery power for a mobile application, additional error correction may be necessary for commands and controls to achieve the necessary bit-error rate.
- Packet size is determined by the data content and the networking overhead information necessary to send and receive data.

Table 10.4. Derived Communication Requirements and Representative Implementations

Attribute	Sensor/ Master Req.	Master/ Master Req.	Master/ Ground Req.	Rf Serial Implementation	Rf Ethernet	PRISM DSS PC Wireless LAN
Peak data rate (Mbyte/s)	0.020	1[a]	10	0.0115 (UART limit)	0.3	0.1-0.2
IEEE standard	NA	NA	NA	RS-232	NA	802.11
Modulation	NA	NA	NA	NA	GFSK	QPSK, BPSK, or GFSK
Range indoor/outdoor (m)	10 indoor	100 outdoor	10^5 outdoor	90/900	161//910	120/1130
Rf power (W)	0.010	0.010	100	0.010	0.05	0.063
Antennas/gain	1	1	1	1-6	1-6	1
Bit-error rate	$<10^{-6}$	$<10^{-5}$	$<10^{-4}$	$<10^{-6}$	NA	NA
Data carrier frequency	NA	NA	<10 GHz	2401–2482	2401–2482	2401–2482
Channel access	NA	NA	NA	FHSS-CSMA or TDMA	CSMA/CA	CSMA or TDMA
FCC license requirement	NA	NA	NA	none (FCC 15.247)	none	none
No. of channels	17	30	1	82	15	82
All weather	yes	yes	yes	yes	yes	no
Vendor	NA	NA	NA	Digital Wireless	BreezeCom	Harris Semiconductor

[a]Assumes 10 submasters.

- Synchronization time is the time necessary for the receiver to synchronize its frequency-hopping pattern to the transmitter.
- The data-packet format that must be used includes all the data to synchronize, address, and quantify data length, and correct errors of the packet.
- Security may be necessary to prevent tampering with the packet or commanding of the system.
- The PN codes determine a predetermined frequency-hopping pattern for a given channel.
- The spectral density is a measure of how effectively a communication system is able to use its allocated frequency band.
- Power consumption is important as a parameter since it affects battery sizing.

Spread spectrum, in general, is a method by which a modulated carrier waveform is spread once by an intermediate frequency and a second time by a means independent of the information. For the direct-sequence spread spectrum (DSSS), a sinusoidal waveform with a frequency corresponding to the intermediate frequency, the ω_{if}, is modulated or multiplied by using a phase shift key (PSK) with an amplitude of +1 or −1 at a frequency corresponding to the digital baseband frequency, f_b. This baseband waveform consists of the binary sequence of information bits that will be transmitted. Lastly, the resultant signal is modulated once more by a spreading signal

function or PN signal, which has a modulation frequency, f_c. The final product of modulated signals is then up-converted to the carrier frequency. A number of simultaneous users are assumed to be broadcasting. At the receiver, the signal consists of the sum of all direct-sequence modulated signals plus additive white Gaussian noise and interference terms. The signal is demodulated, using the same PN sequence, with the result that only the desired modulated waveform remains. All other waveforms are spread over a much higher bandwidth. The despread signal is then PSK-demodulated to form the desired baseband signal, which contains the information that was transmitted in the original bit sequence. The probability of error in the received signal is a function of the number of users, the chipping rate (or the intermediate channel-hopping rate), and the bit rate. For code-division multiple access (CDMA), each user is provided with an individual PN code. If the codes are not correlated, then all independent users can transmit at the same time in the same bandwidth. The despreading in the receivers ensures that only the data sequence with the corresponding PN code is regenerated. CDMA systems can suffer from a near-far interference problem in which a near transmitter will mask the signal from a far transmitter. Adaptive power schemes have been proposed to solve this problem. The synchronization of a DSSS system is roughly equal to a multiple of the chipping period, $t = 1/f_c$, neglecting the additional time to perform the fine synchronization better than one chip period.

Spread-spectrum devices offer improved interference rejection, code-division multiple access, graceful degradation of performance as the number of users are increased, and lower implementation cost.[19] Commonly employed modulation schemes for spread-spectrum include, DSSS including CDMA, slow and fast frequency hopping (SFH or FFH), and carrier-sense multiple access (CSMA).[19]

For the SFH-SS and the FFH-SS, the binary PN-code generator causes the frequency synthesizer to hop from one of the many possible frequency bands selected by the generator. The RF signal is spread in a random fashion over the many available frequency channels. If the channel-hopping rate is slow compared to the bit rate, the system is referred to as SFH-SS. If the channel-hopping frequency is fast compared to the bit rate, the system is referred to as an FFH-SS. For FFH systems the probability of error is the ratio of the number of interferers with power greater than the carrier power divided by the number of channels available. As this number is often too high, forward error correction is used to reduce the probability of error to acceptable levels of 10^{-3} to 10^{-8}. For the purposes of synchronization, the FFH receiver can wait at a fixed-channel frequency and then advance in synchronism with the transmit FFH generator. This will correspond to about one hopping period at worst. Additional time must be allowed to fine synchronize, which may be a few multiples of the hopping period. In general, the FFH system will be able to synchronize faster than a DSSS system.

Error correction can give a decided advantage to the communication system by reducing the signal to noise needed at the receiver (transmitter) to achieve a given system probability of error but at the expense of transmitting additional bits and performing the additional computations necessary to correct or detect errors. Both functions result in some power penalty. With the properly designed code, a net gain in system performance is possible at the same transmit power.[19]

A block-error correction code is formed when a message or code word of k bits length is added to $(n\text{-}k)$ bits of redundant symbols. The total length is n bits for the code word that will be transmitted. The resultant code word is designated a (n,k) block code. The received code word is decoded by deciding that the most likely transmitted code word is closest in Hamming distance to the received code word. (The Hamming distance is the modulo-2 sum of the transmitted and received code words on a component-by-component level, with the assumption that the code words are represented as vectors.) Several codes are important. The (23,12) or Golay code is able

to correct for all patterns of three errors and offers an improvement in the bit-error rate by reducing it by more than 2 orders of magnitude at an energy of bit-to-noise-density ratio in slight excess of 8 dB. The Hamming (7,4) code can correct a single error with moderate improvements with an energy of bit-to-noise-density ratio of 8 dB. Comparison with other codes has been discussed by Feher.[28]

The allowed FCC bands for the carrier frequencies, the Instrumentation, Scientific, and Medical (ISM) RF bands, are located at 902-928 MHz, the 2.402-2.4835 GHz, and the 5.725-5.850 GHz. No licensing is required if certain restrictions are followed. A large number of devices for the two lower bands are available, driven by the growing widespread use of cellular phone technologies.

One of the most important standards for the implementation of LANs refers to the devices that follow the IEEE 802.11 standard for the 2.4-GHz band:

- Up to 1 W of power is allowed in the antenna.
- Data packets are restricted to 2048 bytes.
- The PN code length is 11 bits.
- Maximum data rates can be 1 Mbit/s for FHSS (frequency-hopping spread spectrum).
- The PN codes are used to define a frequency-hopping pattern. For FHSS both the transmitter and the receiver must periodically switch to a new carrier frequency.
- For the 2.4 GHz band, the hop rate must be at least 2.5 hops/s over the 79 sub-bands of 1-MHz-wide channels. This slow hopping rate should ensure complete transmission of a data packet before the next hop occurs.
- For the DSSS, the chipping rate or channel-hopping rate can be as high as 11 Mchips/s with 11 designated frequency channels, each with a 22-MHz bandwidth. The chipping rate is faster than the data rate, enabling processing gain.
- Maximum data rates for the DSSS can be 2 Mbits/s.

10.3.3 The Aerospace Launch Vehicle MPS System: Trade-Offs in Architecture, Requirements, and Design

The design of an MPS system is not always straightforward because of a variety of factors, including the availability of development tools, schedule, and development goals, and the availability of reliable chips, such as, sensors, processors, communication chips, and software. A designer of an MPS system rarely develops all the sensors, processors, software, networking protocols, power supplies, and communication chip sets. Instead, those subsystems for development and those for purchase will have to be adroitly selected. When funds are constrained, selecting COTS components would seem to minimize costs; however, this can be deceptive because consequent software development could be costly. Many components that are offered with software drivers, for example, will be suited only to personal computers, namely, PCs with a Windows® 95 operating system. Access to the source codes is often not allowed, given the competition of the commercial marketplace. Software availability or development cost for the driver software will be a major selection criterion for selecting COTS components in a plug-and-play environment.

A top-level design of a full-up networked MPS system was carried out for the full-up launch-vehicle application and is presented here together with critical issues that the designer would face. Based on the sensor survey, many MEMS accelerometers and pressure transducers would be suitable for the launch-vehicle application. For measurement of acoustic levels, small compact microphones are available in non-MEMS form. Two groups have developed MEMS microphones that may be suitable.[29] Strain gauges have been traditionally fabricated from resistive metallization on deformable thin, polymeric substrates that are readily bonded to a surface undergoing strain

deformations. Numerous temperature sensors are also available such as thermocouples and platinum resistance devices. For temperature and strain, the sensor must be physically isolated from the MPS node to prevent erroneous readings caused by effects of the thermal mass or the elastic modulus of the node, which includes the processors, batteries, transceivers, external case, etc. Many sensors include some form of signal conditioning; calibration of the sensor and signal conditioner is needed. Some sensors provide for a built-in functional test; detailed calibration must still be performed by the user.

A hierarchical networking architecture was selected (Fig. 10.2) and sensors were assigned to numerous engine compartments of the launch vehicle. An average of 50 sensor locations are assigned to each compartment for a total complement of 10 compartments. The compartment metal walls are not conducive to RF transmission, and a port or gateway is necessary to communicate between compartments. This may be accomplished using a small wall penetration, for example, with an antenna on one side and the processor on the other. The distance within a compartment is quite small, approximately 10 ft across, often with no direct line of sight from node to node. In each compartment a local network could be configured with one or more masters that would enable communication to other masters outside. These masters can then communicate and direct data or commands upward or downward in the hierarchy. This scheme allows an advantageous reuse of the RF components and spectra inside each compartment.

Data sent up the hierarchy through the masters will then reach a transmitter capable of communicating to a ground unit. The transmitter on the ground will have a high-gain antenna to receive the RF signals with a processor to perform network control, collect data, unpack the packets, archive data, and perform data analysis. There will be an asymmetry in the upward data flow

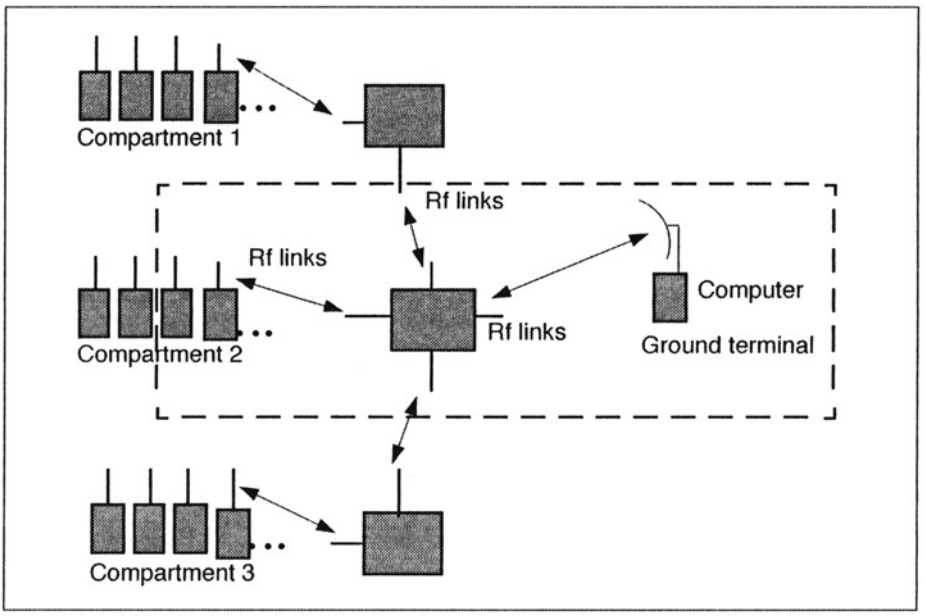


Fig. 10.2. Schematic architecture for the data flow of an MPS consisting of a large number of sensor nodes, at least one master node for each compartment, and a single ground node. The full system may consist of up to 500 sensor nodes organized in this hierarchical structure. The "command" communication network is not shown. The dashed lines outline the portion of the full system that was selected for a demonstration; it consists of two sensor nodes, one master node, and one ground node.

from the sensor nodes, as discussed earlier, versus the downward flow of data and commands to the sensor nodes.

An alternate scheme of direct communication between each sensor node and a ground terminal was rejected for the following reasons. Each sensor node must have access to an outer wall for an antenna to link with the ground, requiring too many compartment wall penetrations and a more complex installation. The transmitter power and electrical power for each terminal must be high enough to complete the link to the ground and would require either an adaptive power transmission strategy using vehicle location to enable a small physical size for each sensor node or a large battery size.

The network is designed for several modes of operation: installation, test, launch standby, operation, and sleep. The system will be installed at the factory because of the large number of nodes (up to 500) that must be employed. During the installation mode, the network system is installed and integrated. Once the master nodes are in place, individual links can be tested. A separate installation master is preferred when installing a single sensor node. The test mode is invoked to verify that the installed system is capable of the mission function. Prior to the launch, the system will be placed in a sleep mode to conserve power, for which there are several strategies. For example, with a single low-frequency watchdog timer on each processor, the sensor node could wake periodically, perform a self-test, move into a receive mode, wait for commands from the master, transmit its status, and return to sleep. Alternatively, the ground terminal could request a self-check until the masters awaken to test their sensor nodes and return them to sleep. Total active data-logging time is 10 min, which is the flight duration for a Titan IV. Tens of minutes before launch, the system will be placed in a standby mode. Since launch countdowns are often interrupted, the system must decide whether to stand down and store power, and it must be able to transition from one mode to another. Precautions may be necessary to avoid inadvertent mode transitions or tampering with the system configuration.

The functions of the sensor microcontroller are to convert analog signals to digital levels, sample all sensors at the requisite sample rate, form a data packet that is time tagged, send it to the appropriate network node, receive and execute commands from the master, and place itself in the appropriate network mode. It may also perform thresholding or data analysis as a means to reduce data. In addition, it must inform the master periodically of its health status, especially during active data-gathering if no data are sent because of thresholding. The raw data from each compartment is 1/10 of the total, or 10 Mbits/s. This rate is within a factor of 3–5 of present Ethernet data rates. For inexpensive hardware, wireless LAN data rates are at the 1–2 Mbits/s rate and will soon be increased to 10 Mbits/s.[30] While a combination of thresholding, data compression, and peak off-loading could reduce net data being transmitted, peak off-loading may only be useful for a normal successful launch. The one caveat of peak off-loading is that during an unsuccessful launch, the system may have programmed assumptions about the integrity of the system throughout its entire mission life, which may become invalidated and lead to loss of critical data describing the failure.

Certain high frequencies at low power that may be used for the small range within the launch vehicle cannot be transmitted at long range in the atmosphere. For transmission from space to ground stations at the typically long-slant ranges in excess of 100 km, the atmosphere is transparent only below 15 GHz.[31] Since the data rates scale with carrier frequencies, there is an upper limit on total peak bandwidth for data transmission. But for the 2.4-GHz ISM band with modest data compression, this is not expected to be a major problem.

Selection of processor hardware is often driven by the availability of supporting software, chips, or dies that meet environmental specs and processing needs, and development tools for

software and diagnostics that can work with brass-board hardware to solve implementation problems, such as, timing, connections, and proper instruction sets.

Once software and data storage is sized based on the algorithms and protocols required, the processing speed can be approximated from how frequently the cycle of data taking, data storage, packet formation, and communications occurs along with an estimate of the networking servicing overhead. For this project, a 32-bit processor of the Intel-486 class was selected as adequate for the sizing of the processors used for the sensors and master nodes. Many equivalent processors/boards able to support the necessary digitizing rates are listed in Appendixes 10.B and 10.C. This demonstration system stressed function over form; thus a palmtop PC for which drivers are available for the PCMCIA cards was deemed acceptable.

Several networking issues identified included latency of networking services such as time distribution or synchronization, fault tolerance, command and control, and differing protocols and data rates for sensor data and network commands. At high bandwidths, error correction and scaling to support high data rates are additional concerns.

For communications, the advantage of frequency reuse is inherent in the hierarchical architecture defined by the compartments of the vehicle (which form natural cells) and the metal compartment walls. This allows implementation of identical architecture from cell to cell. Moreover, if compartments are indeed isolated from the exterior, the identical frequencies may be reused for the master-to-master communications. Within a cell, a simple token passing scheme of the sensors by the appropriate master, a form of time division multiple access (TDMA), could be used to meet required communication rates. This may be easier to implement using a single transceiver on each sensor node and a single transceiver on the master. Multiple transceivers may also be used on the master to increase the total data rate to the master, but at the expense of greater hardware complexity and power.

The explosion of personal communication devices such as cellphones and pagers during the 1990s was realized using digital techniques. Digital communication has the following advantages.

- Allows improved RF spectral utilization or capacity
- Enables digital data compression methods
- Reduces overhead for signaling over analog methods
- Enables a robust source and channel coding methods
- Improves performance during interference due to cochannel and adjacent channel interference
- Allows flexible bandwidth allocations to meet demand
- Expands the services over analog systems (such as, data services, encryption, authentication)
- Improves access and hand-off control.[19]

In the near future, a large number of spread-spectrum devices will become available with the necessary bandwidth. In 1996, many fast serial RS-232 digital RF transceivers were available. For the purpose of real-time control, these serial implementations could be chosen despite their slower data rate. For the purposes of a demonstration, the current data rates are satisfactory but would need to be increased for a larger, high-data-rate system.

Power for the entire system will probably come from stored battery energy. The necessity for power management when using battery power is demonstrated by the following example to calculate battery capacity. A PIC 17C756 microcontroller and a Harris PRISM chip set were used, and calculations for a sensor node and a master node were performed. The sensor node for this exercise has three analog-device accelerometers, a single PRISM transceiver, and a single PIC microcontroller. The master node has one microcontroller and two transceivers. Calculations for power consumption are done for the network in the sleep mode and the active mode. The sleep

mode duration is assumed to be 60 days, and the active mode, 10 min. During the sleep mode the sensor node must occasionally transition the microcontroller from sleep to active mode to turn on the transceiver for a period of 5 s every 30 min. During this brief "on" mode, the node must go through a warm-up/stabilization period and discern whether the master node is attempting to command it to remain awake. If no wake-up call is received, the transceiver is turned off, and the microcontroller goes back to sleep until the next cycle. The power of each device is summarized in Table 10.5.

The sensor and master nodes sleep-mode consumes the bulk of the power because of the "on" status of the microcontroller and transceiver. If the on/off duty cycle time can be reduced further, a more favorable sleep-to-active mode power utilization is possible. For the master node during sleep-mode, the transceiver outside the compartment must listen and react to messages destined for each compartment. The master then transmits the message to the sensors assigned to it. To simplify matters, we assume that the transceivers are on for 5 s and cycle through this routine once every 30 min. Obviously we neglect the energy to perform the transmit portion of the wake-up call, which will occur periodically over a maximum time span of approximately 30 min. During the active mode the master node has the transmitter and microcontroller "on." For this design, greater than 95% of the total stored energy is used during the network's sleep mode. If we use lithium thionyl chloride batteries, the weight of the batteries for the sensor and master nodes will be 3.8 g and 6.6 g, respectively, not allowing for any margin or battery package. This calculation has not properly accounted for the additional time that the controller and transceiver must be on to precede the event to be monitored by the system. The system will be awakened to the nearest period preceding the time of the active event. Prior to the 10-min active period, the system could be awakened as early as 30 min before the actual event, and the maximum time of the active mode will stretch to 40 min.

For many systems requiring a significant period of sleep time or shelf storage time, the designer must implement a low-duty-cycle sleep time to prevent sleep power requirements from dominating the battery sizing. Other strategies to reduce power loads include placing the transceivers temporarily into other energy conservation modes as allowed by the chip designers, using low current leakage, miniature mechanical or MEMS relays to turn power on or off, capacitive sensors, and additional techniques available to the designer of a μCluster as discussed in Subsec. 10.6.4.4.

Table 10.5. Device Power Consumption for Sensor and Master Nodes in Sleep and Active Modes

Power	Sensor Node Sleep	Sensor Node Active	Master Node Sleep	Master Node Active
Accelerometers (mW)	0	150	NA	NA
Microcontroller (mW)	0.055	150	0.055	150
Transceiver (mW)	0	460	0	920
Sleep-mode duty cycle (on/off in s)	5/1795	NA	5/1795	NA
Subtotal (W-s)	9.1×10^3	4.6×10^2	1.6×10^4	6.4×10^2

10.4 The Plug-and-Play COTS Approach

A test bed was constructed using COTS components for the networked MPS system. The approach was to demonstrate function rather than meet form and fit functions. Available components were chosen primarily for rapid implementation rather than for the comprehensive requirements for the full-up launch-vehicle networked MPD system. Any shortcomings in meeting full-system requirements are noted.

Two sensor nodes, a single master node, and a single ground-terminal node were designed and integrated. See Fig. 10.3. (The relationship of the selected demonstration to the full system is shown in Fig. 10.2.) For the sensor node, a set of three accelerometers for three-axis sensing and a single pressure sensor were chosen as representative of high and low data-rate sensors. For the microprocessor, a palm-sized 486PC and an ADC PCMCIA card were selected. The communication subsystem was implemented using an RS-232-based spread-spectrum transceiver, one unit per sensor node.

The master node used a 486PC for its processor and identical transceivers, one for each sensor node and another for the communications to the ground terminal, for a total of three transceivers. The ground terminal used a single transceiver and a laptop using Labview, a commercial instrumentation software with a graphical user interface. The assembled system is shown in Fig. 10.4 and is described in the following section.

10.4.1 Selected Sensors

Three different accelerometer models were chosen for the COTS demonstration.

- Silicon Design, Issaquah, Washington: 1210-50-J single-axis and 2412 three-axis
- EG&G IC Sensors, Milpitas, California: 3255-050 and 3255-500 single-axis
- Motorola, Phoenix, Arizona: MMAS40G10D single-axis

The first two sensors had adequate bandwidths of either 1.6 kHz or 2.0 kHz, but the third had a reduced 400-Hz bandwidth, inadequate for the launch vehicle application. It was selected because it was one of the cheapest accelerometers available at the time, but many MEMS accelerometers of suitable bandwidth and range for the launch-vehicle application are now available.

The Analog Devices 5-g accelerometers were selected for further calibration studies. In the dc mode, calibration can be accomplished by orienting the sensors in three directions to the Earth's

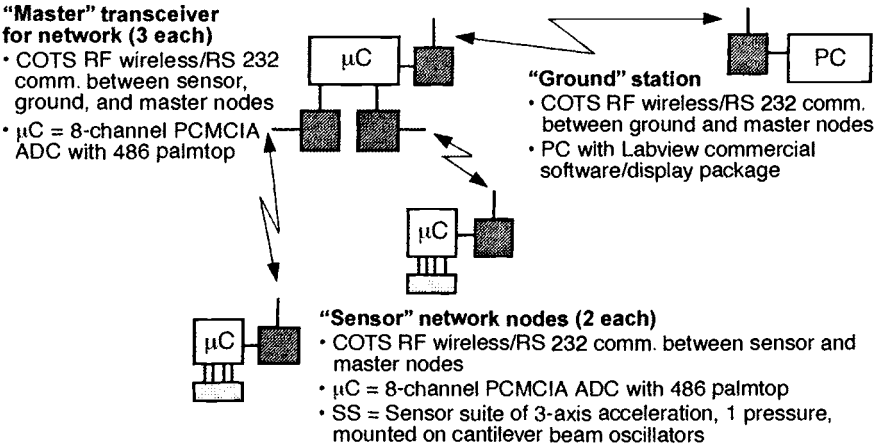


Fig. 10.3. Schematic of the functional networked MPS system. The four nodes depicted are two sensor nodes, a master node, and a ground terminal.

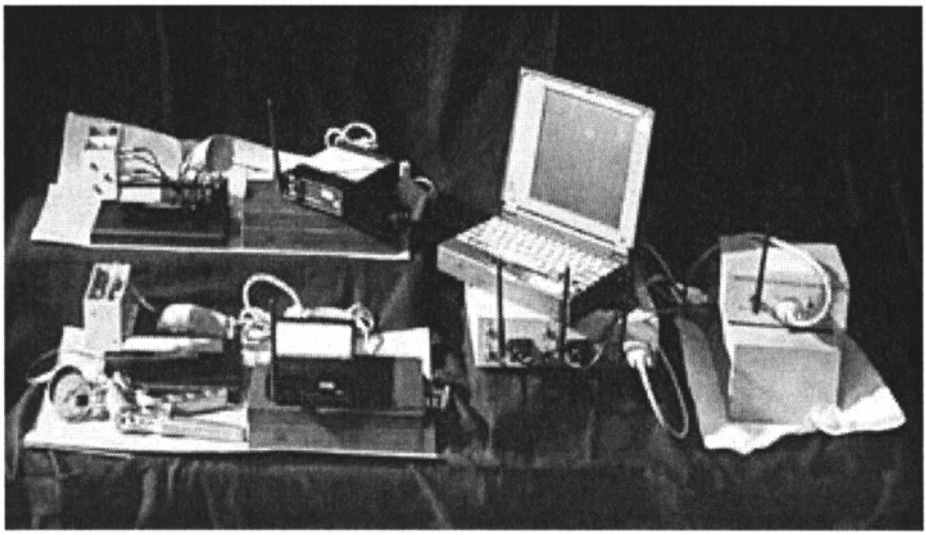


Fig. 10.4. COTS implementation of a partial MPS system (as shown in Fig. 10.2). The two sensor nodes are located to the left above one another. To their right is a single master consisting of three transceivers and an open palmtop computer. The ground terminal, consisting of a single transceiver, is shown to the right of the master. The laptop was not available at the time of the photo.

gravitational field, +1 (with), 0 (perpendicular), and −1 g (against). These sensors have a built-in amplifier to enable ac or dc coupling, variable gain, and zero-offset. The accelerometers are very linear over their range, and the slope responses can be reproduced to within 5% of one another.

10.4.2 Processor and ADC

The processor hardware includes three 75-MHz, 486 palmtop computers (model iLuFA 350 from Chaplet, Sunnyvale, California) for the sensor and master nodes. The palmtops support PCMCIA plug-ins, which include a 100-kilosamples-per-second A/D converter (model DAQ Card-1200 from National Instruments, Austin, Texas) used in the sensor node and additional serial ports (model MP540301 dual high-speed serial card from Mobile Planet) used in the master node. The palmtop computers were selected for their small physical profile and the readily available PCMCIA plug-ins and drivers for DOS or Windows® 95. The ground node is simulated using a Pentium, 133-MHz laptop PC (Hitachi model E-133TC).

The ground terminal was developed using a Labview software package,* a commercial, graphics-based software/hardware interface to accept the time-tagged data packets and to display the resultant waveforms in real time for a demonstration. The ground terminal displays the digital output of the pressure sensors and an operator-selectable time duration for the display of the 3-axis accelerometer data of amplitude versus time. Data from both sensor nodes are displayed simultaneously to present visual feedback to the operator.

10.4.3 Communications

Three pairs of wireless, 2.4-GHz RF spread-spectrum, RS-232 transceivers were selected for serial communication between the sensor and master, and for the master to ground links. The transceiver (Digital Wireless, model WIT 2400) is implemented on a printed circuit board with

*Labview software is available from National Instruments.

2.4 × 3.0-in. dimensions. Additional space is needed for an antenna and a small circuit to convert RS-232 serial voltage levels to TTL voltage levels. A robust sensor-network communications protocol was designed that emphasizes flexibility, simplicity, and bandwidth conservation. An amplitude thresholding concept was used to reduce the data being reported by the system. Since data below a predetermined threshold are not reported, the data packet protocol employs time stamping for each reported sensor amplitude. One novel feature of the protocol is that it supports a wide variety of error correction techniques that can be selected by the network user to balance data integrity with bandwidth consumption. The protocol, in its current form, will support fewer than 10 sensors.

10.4.4 Assembly and Testing

All major components of the software were developed in parallel to hasten development time, including the complete network stack, the software libraries for the data acquisition hardware, and the application layers for the multiparameter sensor demonstration. The "spiral model" of software development was selected for its characteristic rapid, repeated cycle of incremental design, implementation, test, and integration. This process ensures that at each stage in the development cycle, successive versions of the system are integrated, bug-free, and meet performance requirements.

High-speed serial drivers were developed, which are capable of providing the maximum performance with a PC-compatible architecture and use the maximum transfer rates provided by the spread-spectrum RF transceivers. Initial timing studies of the data acquisition rates were performed on the hardware selected for the prototype. The goal to acquire 2,000 samples per second for each axis of the three-axis accelerometers was easily met. A total acquisition rate of 12,000 samples per second has been achieved, which is twice that required for the prototype demonstration. This limitation, consistent with model prediction, is dominated by the fundamental transmission rate of the RS232 hardware and the UART hardware (both are specified at 115 kbits/s) and not by the data acquisition hardware speed or the software execution speed. The observed data rate is adequate for many space applications using a small number of sensor nodes. Larger numbers of sensor nodes can also be supported, but the master-to-master link must support a rate slightly larger than the sum of sensor-to-master data rates. For the master-to-master link, the present implementation of RS-232 would be inadequate. Faster transceivers at 2 Mbits/s using other protocols such as 10 Based T Ethernet or the IEEE 802.11 standard could be used to improve throughput either at the sensor-to-master or master-to-master links. For the full-up system using data compression techniques, load averaging methods, or upgrading to a faster IEEE 802.11 device, the master-to-master link requirements would be satisfied.

10.4.5 Second Generation Multiparameter Sensor System

The development of a second-generation MPS system sensor concept[32] is under way to achieve a miniaturized sensor node using the PIC 17C756 microcontroller and an RF-communication, PCMCIA card that uses the IEEE 802.11 standard for the 2.4 GHz band. The master node consists of identical RF-communications cards, a laptop computer, and Labview software that displays data and provides rudimentary system commanding, system-mode control, and data archiving. The goal is to achieve the first milestone of developing compact hardware and corresponding software capable of high-speed (1–2 Mbits/s) communications that could be used in applications similar to that of the launch vehicle previously discussed. The architecture of the sensor nodes is organized into custom sensor and processor boards, a networkable RF-PCMCIA card from Harris, and a custom battery pack. The sensor board includes 3-axes accelerometers, pressure transducer, humidity sensor, and temperature sensor, all with appropriate signal conditioning. The processor

board includes the PIC microcontroller, 4 Mbytes of memory, an ACTEL-programmable gate array to interface the processor to the memory, and the RF card. The system modal diagram supports the following modes: sensor-node installation, calibration and status check, data acquisition and temporary data storage, power savings (sleep), and data reporting/command response to the master. The first milestone consists of the completed sensor node sending data to the master. Network self-configuration and other network controls will be implemented as a second milestone, with field testing implemented as the third milestone.

Development is 80–90% complete for the first milestone. Nearly all software modules have been designed, implemented, and tested piecewise, except for the RF-communications module. The microcontroller and software for data acquisition have been tested, and a maximum of 12 Ksamples/s cumulative measurement rate is projected using 6 ADC channels. However, additional time allocations for RF transmission and overhead were not included in this estimate. The mechanical design of the enclosure and edge connectors were completed and will undergo shaker testing to validate the robustness of the design. The custom sensor board has been designed and is undergoing testing. Complete circuit designs are awaiting the details of the RF-programmable gate array and are estimated to be 90% complete. Once the design is completed, breadboard hardware will be assembled to integrate and test the entire design, with anticipated completion in early 1999. Final sensor nodes will be available shortly thereafter.

10.4.6 Summary of the Aerospace COTS Development

A high-data-rate, MEMS-based instrumentation system concept to demonstrate function was designed, based on the wireless, high spatial-density-measurement requirements for a launch vehicle environmental monitoring application. Commercially available components, including sensors, processors, A/D converters, and wireless RF transceivers, were selected for the development of a demonstration based on a subset of the full-sensor network. The demonstration required the development of protocol design and implementation, including network stack, software libraries, timing tests, and network throughput. Scaling issues to the full-system performance were identified. Many of the concepts and sensor hardware could be easily reused for a microsystem of small form factors.

A second-generation sensor-node development was initiated as a direct design of an MPS system. Its emphasis is on high-speed data acquisition and data reporting similar to the launch-vehicle application. The sensors include commercially available accelerometers, temperature, pressure, and humidity sensors. The processor is a PIC 17C756 microcontroller and a networkable RF transceiver. About 80–90% of the hardware and software design is completed. Piecewise testing/simulation of hardware and software was completed, but final integration and test will occur when the software module is completed. The RF software module will lead completion of the final design late in early 1999.

10.5 Wireless Integrated Network Sensors

A concept developed at the University of California, Los Angeles, to monitor and control a system capable of networking, signal-processing, sensing, and decision-making at low power with RF communications shows great promise for a variety of applications.[33] These applications include transportation, manufacturing, health care, environmental, and safety and security monitoring. The system is referred to as the Wireless Integrated Network Sensors (WINS). Size of a sensor node board is approximately 1×2 in.

The architecture of the sensor node includes components for sensors, an ADC, a spectrum analyzer with a set of analog filters on the input, memory buffer, a microcontroller, and RF transceiver. The WINS system, a collection of sensor nodes with a gateway to a more conventional

network, is capable of acting as a distributed sensor network. The sensor suite consists of a dual thermopile and horizontally and vertically sensing accelerometers. Each thermopile consists of 32 elements of a bismuth-antimony junction capable of 1.8 nW/(Hz)$^{1/2}$ sensitivities. The accelerometers have sensitivities near 2.4 mg for frequencies near 11 Hz. For a band centered at 5 Hz, the sensitivity is 10-7 g/(Hz)$^{1/2}$. The spectrum analyzer and sensors are on continuously. The spectrum analyzer output triggers a wake call to the microcontroller that determines the next course of action of either additional processing or notification of a user or neighboring node. This design is tailored toward detection of vibrations, especially those caused by rotating machinery or repetitive phenomena and temporal changes in infrared (IR) signatures detectable with poor spatial resolution, presumably large objects in the field.

Power utilization is a concern in the design of the WINS system. Low power of less than 30 μA is needed with peak power less than 3 mW; compact Li-coin cell batteries are used. The sensor node RF range is limited by the low-power transmitter of 1–3 mW and by the power and sensitivity to noise of the receiver. To reduce receiver noise, a high Q, LC circuit design was implemented for the voltage controlled oscillator (VCO) in the receiver. A low-power mixer was also designed, using 0.8 HPCMOS (high-performance complementary metal oxide semiconductor). The mixer modulates the carrier frequency at 900 MHz. Power dissipation for the receiver oscillator is 300 μW at 500 MHz and for the mixer, 70 μW at 3 V. These are some of the lowest power CMOS RF oscillator and mixer reported. Overall system communication power can be optimized by relaying information in multiple hops, compared with the use of a single hop with a higher power transmitter in a multipath RF signal propagation environment. A second strategy to reduce power is to increase the delay time for event recognition, which allows for reduced power in processing and computation. This is ideally matched to low-bandwidth sensors of less than 10 kHz.

In summary, a very compact, low-power, low-cost data reporting system was built with the potential for applications in environmental monitoring, safety, security, manufacturing, biomedicine, and condition-based maintenance.

10.6 Design of a Microinstrumentation Cluster

The University of Michigan Microinstrumentation Cluster, or μCluster,[34] takes a different approach to microsystem design. The μCluster is a multiparameter sensing system that supports a variety of MEMS sensors within each sensing node of a macrosystem. Work on the μCluster project focused on the development of individual microsensors and the definition of a generic microsystem architecture with flexibility to meet the requirements of many different sensing applications. Throughout the development, a number of interesting trade-offs have been analyzed. Since many of these will be common to any similar microsystem design, they are discussed in this section.

10.6.1 μCluster Requirements

When designing a microsystem from the ground up, the first step is to determine the goals of the intended system. For the μCluster, the goals were the following.

- Create a generic microsystem that would support a wide variety of applications simply by changing the system sensors (and possibly modifying the control software). The intended applications range from distributed environmental monitoring to industrial process control.
- Provide a standard design for the individual sensor nodes (microsystems) that could be used within each macrosystem.
- Select a specific application for which a prototype system could be designed, in order to establish specific design requirements of the μCluster. Environmental monitoring was selected.

- Create a system capable of monitoring its environment and its position within that environment while operating at micropower levels from its own battery supply, communicating to a host system via a wireless link, and maintaining a small physical size.
- Develop capability for the microsystem to enhance sensor performance and functionality with system-level features, respond to event-triggered interrupts, and utilize in-module digital data compensation.

To accomplish these goals, the microsystem must have the following features.

- It must utilize some sort of internal sensor bus to which (almost) any sensor can be attached provided it has the proper interface hardware. This sensor bus makes the microsystem architecture independent of any specific sensor or sensor technology and makes it easy to add or remove sensors from the system.
- To operate at micropower levels, the μCluster uses only low-power components (i.e., capacitive sensors) and employs a power-management scheme involving both hardware and intelligent control.
- Wireless communication requires the microsystem to have an on-board transmitter for data output and a standard output format so that the data can be received remotely.
- Minimizing the physical size of the microsystem requires limiting the number and size of components (including batteries) and identifying an appropriate packaging technology.
- Some built-in intelligent control is needed to allow in-module decision-making and data-manipulation.

Thus, the μCluster must bring together low-power electronics, state-of-the-art microsensors (and perhaps microactuators and microinstruments), and wireless communication to form a standalone intelligent microsystem capable of delivering processed sensor data/information to a remote host as part of a network of similar microsystems.

10.6.2 Generic Microsystem Architecture

Once the overall system features are identified, a block diagram of the microsystem can be constructed. From a macro-system view, it is assumed that there will be a single host system that receives the data from the individual microsystems. In some applications there may be many microsystems for each host; while in others there may be only one. The μClusters would connect to the host through a networking scheme similar to that discussed in Sec. 10.3. In defining the microsystem architecture, a great advantage is to provide a generic structure that can support a variety of sensing applications. A generic open architecture allows individual aspects of the microsystem to be altered without affecting the entire system; for example, with generic open architecture, a system that measures temperature and humidity could be converted to a two-axis accelerometer simply by changing the sensors (and control software) but retaining all other system features. The alternative would be to design application-specific microsystems that have no common structure and require redesigning from the beginning each time a new application is to be supported. A generic open architecture also provides a set of standards for the individual components of the microsystem. Once in place, these standards simplify the design of both sensors and new microsystems by establishing a fixed protocol for communication between components.

The controller to provide intelligence and programmability will, in essence, be the heart of the microsystem and is a good starting point to define the system. Next, the generic microsystem modular approach to sensor population will allow each front-end sensor chip, which may contain a variety of sensors and actuators (addressable elements), to be considered as an individual module that can be replaced without altering the rest of the system. Each such sensing node must have all the necessary electronics for making measurements, converting data to a standardized output

format, and communicating this information to the microsystem controller. To accommodate multiple sensing nodes within the microsystem, a standard sensor-bus format (and related protocol) must be established to allow communication between the controller and the local network of sensor nodes. The internal sensor bus must be defined in a way that is generic and yet covers the needs of a variety of sensors. The sensor bus must allow for sensors to be removed, added, or exchanged without altering the bus itself. An appropriate sensor bus capable of these features is discussed in Subsec. 10.6.4.1. Ideally, the sensors could be added/exchanged without external modification of the system software (e.g., the sensor node would automatically upload its control/personality information to the microcontroller on system power-up).

A wireless communication link and some power-management features complete the system block diagram. Because of size and power limits, the wireless link will be an RF transmitter capable of transmitting an appropriate (e.g., amplitude-shift-keyed) output format. This format will also simplify the system by using the same protocol for both wireless and hardwired output connections. A variety of concerns regarding power management must be analyzed in order to provide a generic approach to this complex issue. This will be discussed in detail in Subsec. 10.6.4.4, but for now it can be assumed that power management is another block that interacts intimately with the controller. This is shown in the block diagram of Fig. 10.5, which illustrates the necessary building blocks for the generic microsystem and how they interact. The illustration also provides additional information as to how the individual sensor nodes can be viewed, either as a single block or as several interconnected blocks. The significance of this will be addressed in subsequent sections on sensors and packaging. Each microsystem building block is discussed individually in the following section. The trade-offs in choosing each component will be discussed as well as how these choices affect each other and the final microsystem definition.

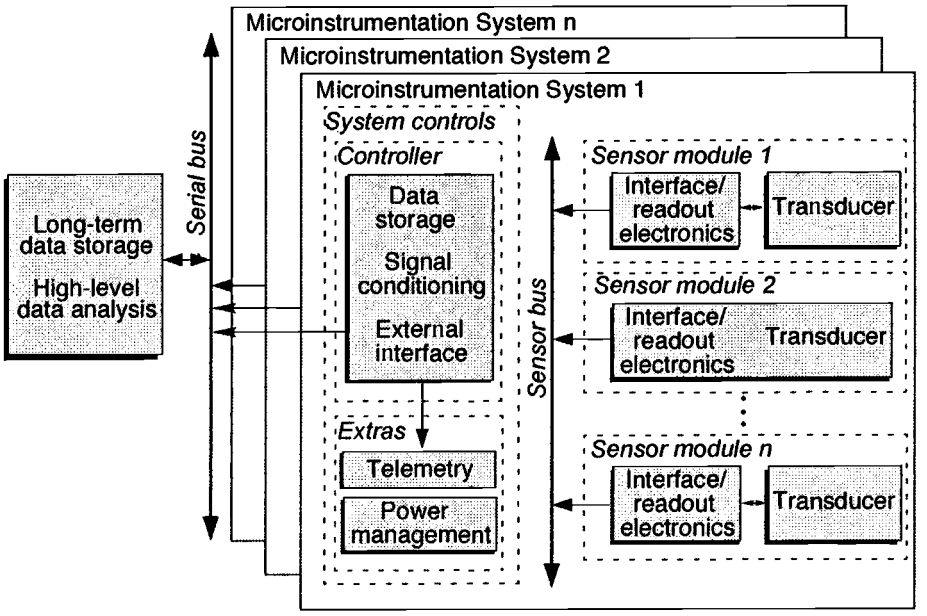


Fig. 10.5. System architecture block diagram for a generic microsystem. This architecture has been adopted by the University of Michigan Microinstrumentation Cluster, called the μCluster. This diagram shows the microsystem-level architecture as well as the position of each microsystem in the overall macrosystem that joins multiple microsystem sensor nodes to form a distributed sensing system.

10.6.3 Integrated Sensors and Microactuators

Highly integrated microsystems such as the μCluster depend on low-cost microprocessors and memory and on the ability to form the front-end transducers.[35,36] The first integrated silicon sensors emerged in the mid-1960s in the form of visible imagers. These devices required no process technology other than that used for integrated circuits, and the only packaging modification was use of a transparent window in the package lid. Visible imagers have been continually improved over the last 30 years and are now approaching the resolution of photographic film. Video cameras have replaced movie film with magnetic tape, and digital still cameras are now entering the mass market. By the early 1970s, the first selectively etched pressure sensors were reported,[37] but it was not until the early 1980s that high-volume products began to appear using micromachining (i.e., the precision etching of three-dimensional microstructures, usually in or on silicon). The first of these products were pressure sensors,[38] followed by accelerometers,[39] flowmeters,[40] and other devices. Most of these devices have been driven by automotive applications or by applications in health care. Monolithic gyros[41,42] are now in development, along with increasingly complex microinstruments. Of the different parameters to be measured, pressure, force, and acceleration have yielded well to transduction by silicon microstructures, and the high accuracy rate of these devices rival the capabilities of monolithic data converters. Near-inertial-grade accelerometers and gyros will likely emerge during the coming decade and, combined with the global positioning system (GPS), should come to be widely applied for precise position sensing, tracking, and navigation. Thermal devices have been readily integrated in silicon for monitoring temperature, and micromachined devices that utilize dielectric beams and diaphragms are creating a paradigm shift in IR imaging.[43,44] Magnetic devices have not yet made extensive use of micromachining, but magnetometers and Hall devices have been realized in silicon for many years. Chemical devices (e.g., for exhaust gas analysis, process control, and pollution monitoring) remain among the most needed of all sensors, and yet their problems are among the most complex. Nevertheless, approaches based on micromachining are among the most promising for these devices as well.

Ability to form microactuators on or in silicon added a new dimension to microsystems in the mid-'80s. Lateral comb resonators[45] have been widely employed in accelerometers,[46] tunable micromechanical filters,[47] and elsewhere. Microactuation is more difficult than microsensing, and many near-term microactuators will likely continue to be embedded devices used for self-testing sensors. Exceptions are found in projection displays based on micromirrors[48] and in microvalves.[49] Micromachined silicon-based fuel injectors have also been realized with some success.

10.6.3.1 Technology Options

In the fabrication of integrated sensors and microactuators, all of the technologies developed for integrated circuits are used. In addition, special technologies are employed to create microstructures (diaphragms, beams) for transducing the variables of interest. These special technologies are micromachining, wafer bonding, and electroforming. See Chapter 1 for more details of microfabrication methods and options.

10.6.3.1.1 Micromachining

As originally applied, the term "micromachining" referred to selective etching of the silicon wafer ("bulk micromachining"). Early efforts on beam-led integrated circuits* at Bell Telephone

*Beam-lead integrated circuits (electroplated lead-tabs that extend beyond the chip) were developed in the early 1960s at Bell Telephone Laboratories. They were a flip-chip technology compatible with hybrid thin-file passive components on ceramic. Widely used until the mid-1970s, beam leads are still used in some sensor and display approaches.

Laboratories led to the development of isotropic and anisotropic silicon etchants. Micromachining switched to the anisotropic etchants in the early 1970s to reduce undercutting, reduce agitation sensitivity, and allow the use of impurity-based etch-stops (Fig. 10.6). Silicon etchants such as ethylene diamine pyrocatechol (EDP) attack the <100> directions in silicon much faster than the <111> directions, and the existence of a highly doped (p++) boron layer in the silicon results in a rapid falloff of etch rate,[50] creating an etch-stop. Thus, a simple boron diffusion can be used to retain bulk silicon layers from 1.5 to 15 μm thick with no critical timing of the etch. Since the doping levels in p++ silicon are too high to permit electronic device fabrication, the etch-stop is sometimes used as a buried layer [Fig. 10.6(b)]. An alternative is to use an electrochemical etch-stop [Fig. 10.6(c)], which requires a bias distribution network on chip and more elaborate etching procedures, but is also coming to be widely used. Recently, there has been the increased use of devices in which front-side patterns oriented along the <100> lateral directions in silicon are undercut. These structures are more quickly formed than the back-etched structures and are more compatible with standard foundry process flows. Dry etching[51] is also being increasingly used both for shallow etches and for deep high-aspect-ratio devices. Aspect ratios exceeding 30:1 can be achieved.

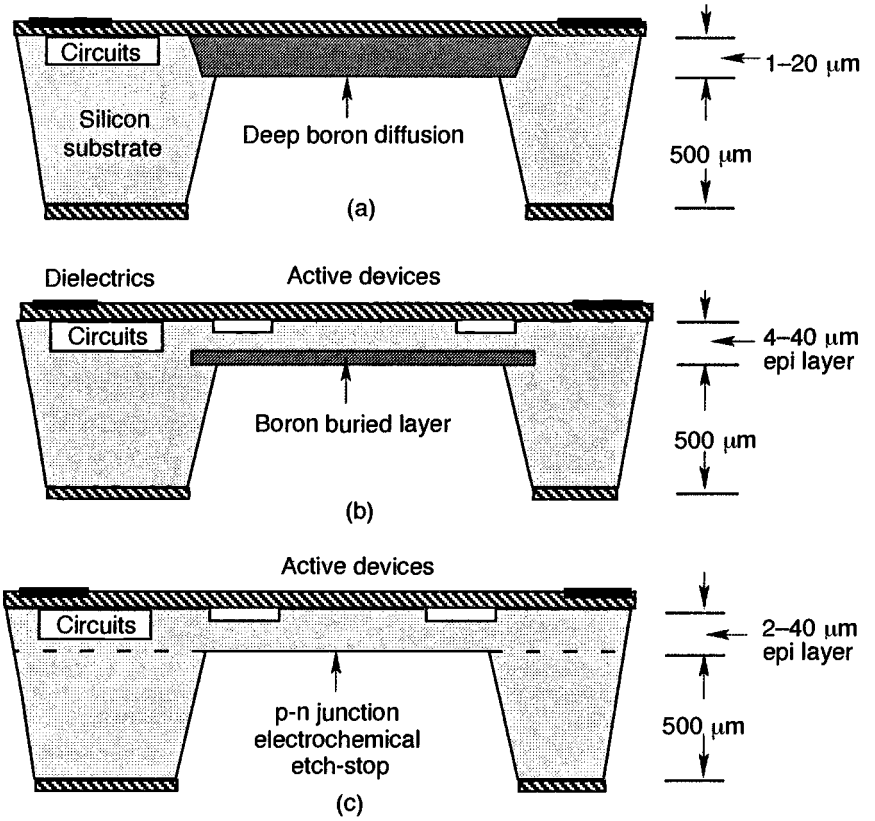


Fig. 10.6. Impurity-based etch-stops in bulk micromachining. Use of (a) simple boron diffusion, (b) diffused boron buried-layer, and (c) p-n junction epi-substrate etch-stops are illustrated.

10.6.3.1.2 Wafer-to-Wafer Bonding

In electrostatic (anodic) bonding, the silicon wafer is placed against a glass substrate and heated to a temperature of 400°C–500°C. The glass is chosen to closely match the thermal expansion coefficient of silicon, with Corning 7740 a common choice. From 400 to 1000 V are applied across the silicon-glass interface, which pulls the two materials into intimate contact and fuses the materials in a seal that is stronger than either of the materials separately. The advantages of these structures include relatively simple wafer alignment (the glass is transparent) and low parasitics. However, the glass is more difficult to cut than silicon, and the thermal expansion coefficient is not a perfect match to silicon, giving rise to temperature sensitivity. Where better temperature matching or on-chip circuitry is desirable, direct silicon-silicon fusion bonding is possible.[52] Two silicon wafers are cleaned, placed in contact, and heated to temperatures typically exceeding 1000°C. The materials bond to effectively form one piece of material. Where silicon-glass anodic bonds will form across surface steps of several hundred angstroms, silicon-silicon bonds require surfaces that are flat to within angstroms and very clean. The high annealing temperatures can also compromise some processes. Nevertheless, silicon-silicon bonding can create unique microstructures, and its use is increasing.

10.6.3.1.3 Electroforming Processes

A third approach to achieving three-dimensionality involves electroforming. An old technology used in the early days of integrated circuits, it has been recently rediscovered for integrated sensors. Using photoresist, or polyimide defined using dry etching, as a plating mold, it permits the batch formation of rather thick metal structures. In the widely publicized LIGA process,[53] X-ray lithography permits the formation of very thick structures (>1 mm) using electroplated nickel. Very high aspect ratios are possible in such devices at the cost of a synchrotron source. The process offers unique capabilities for devices that can be produced in no other way.

10.6.3.2 Device Options

Bulk micromachining has been discussed above for the formation of beams and diaphragms from the wafer bulk. Front-side bulk processes that undercut the microstructure [Fig. 10.7(b)] are especially compatible with foundry fabrication, typically require no modification of the process flow except for a final etch just prior to die separation, and are also becoming increasingly used. Surface micromachining[54] emerged in the mid-1980s as an alternative to bulk processes. In surface micromachining [Fig. 10.7(c)] a sacrificial layer (usually phosphosilicate glass [PSG]) is deposited on the wafer and patterned. The intended microstructure material (polysilicon or metal) is deposited over it and patterned so that it anchors to the wafer over the ends of the sacrificial material, which is subsequently removed to leave a beam, cantilever, or diaphragm. Such devices have been used for accelerometers,[46,52] for pressure sensors (using seals provided by CVD dielectrics),[55] and for other microstructures. Control of stress in the deposited layer is an important challenge with polysilicon high-temperature (>1000°C) anneals generally required. Stiction[56] is also a significant problem in some structures because of forces involved in drying the structures after release. This has led to special release procedures.

Both hybrid and monolithic mixtures of transducers and readout circuitry are now being used, with a trend toward monolithic implementations as processes become better understood. Certainly for high-volume applications, the monolithic approach is attractive since added process development costs can be recovered, final cost is somewhat lower, and reliability may be improved in the absence of internal wire bonds. Even here, however, it seems likely that the integration levels on transducer chips will remain modest, with more complex digital signal processing and/or microprocessing done as a separate in-module IC. This reflects the current approach in

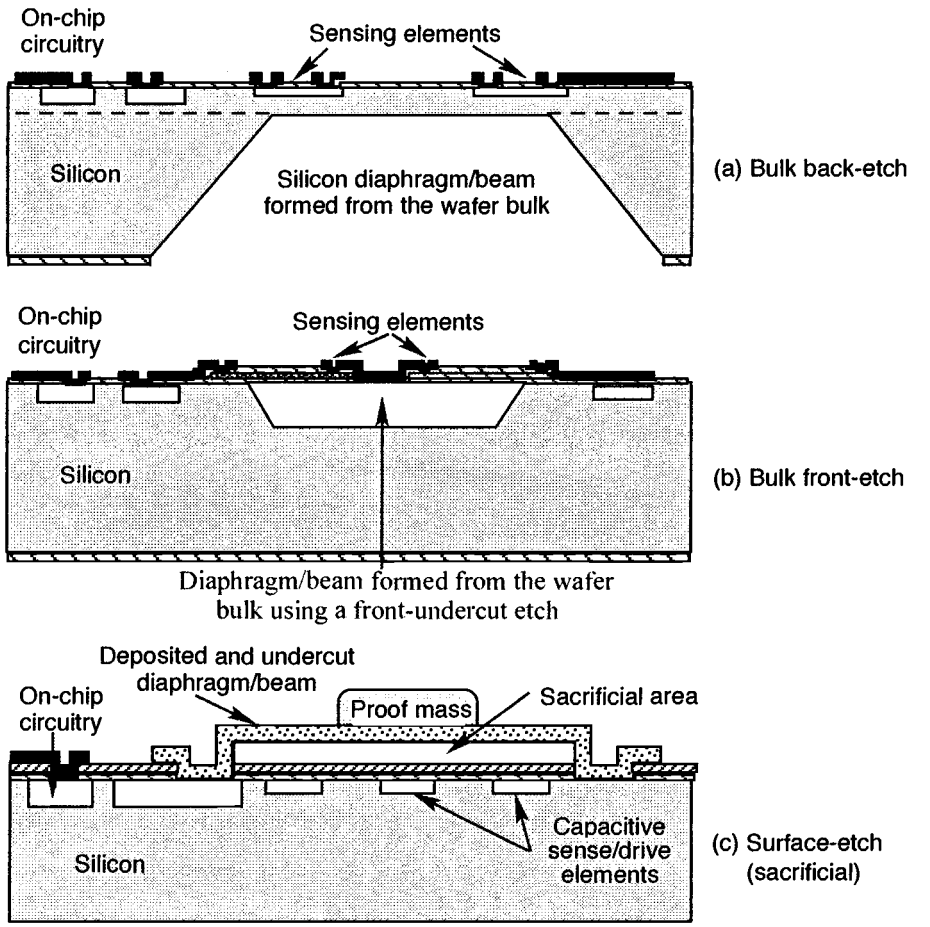


Fig. 10.7. Comparison of bulk, front-undercut, and surface micromachined-device structures.[4]

the μCluster. Bulk micromachining typically requires the formation of etch-stops prior to the circuit fabrication sequence, whereas surface micromachining requires the addition of the microstructures after the normal circuit flow. In either case, well-designed circuitry can go a long way toward making a marginal transducer look good. Surface-micromachined structures based on lateral capacitance, for example, may have full-scale capacitance ranges of no more than 100fF, making it a real challenge to achieve a broad analog output range. Bulk microstructures can do significantly better but are somewhat more difficult to merge on-chip with circuitry.

10.6.3.3 Representative Devices

A wide variety of sensors and actuators are in development, and during the next 5 years many of these are expected to move to commercial production. Indeed, the worldwide annual market for such devices is expected to increase several fold during the coming decade, reaching into the tens of billions of dollars. The present μCluster contains sensors for pressure, temperature, humidity, and acceleration (both threshold [impact] and continuous [tracking]), with additional sensors for acoustic and chemical analysis in development. In this section, we will look at the barometric pressure sensor as an example of a device currently in the μCluster and will then mention briefly an IR (thermal) sensor and a gas detector as other examples of emerging devices.

10.6.3.3.1 High-Resolution Barometric Pressure Sensor

Micromachined, silicon pressure sensors employing dielectric or silicon diaphragms can be grouped according to their transduction mechanism as piezoresistive or capacitive and according to their readout structures as static, resonant, or tunneling devices. In spite of impressive recent progress in sensitivity and accuracy, emerging applications in environmental monitoring offer substantial challenges in that they demand both high accuracy and a wide operating pressure range in order to precisely resolve the local pressure differences needed for global weather forecasting. With a desired resolution of about 25 mtorr (equivalent to about one foot of altitude shift at sea level) over a dynamic pressure range from 500 torr to 800 torr, this amounts to resolving nearly 1 part in 10^5 over a temperature range from perhaps –25°C to +60°C. In this section, we illustrate an approach to meeting these needs.

Figure 10.8 shows the structure of the barometric pressure sensor.[57] This device is composed of four capacitive pressure-sensing elements, each formed with a bossed diaphragm and a readout electrode on an opposing glass substrate. The diaphragms and bosses are both circular to avoid high stress concentrations. Since the bosses are much thicker (12 μm) than the diaphragms (2 μm), most bending and stretching when exposed to differential pressure occurs in the thin diaphragm. The bosses act as parallel plates with respect to electrodes on the glass substrate. These electrodes serve as reference plates, and the diaphragms suspended over them act as movable plates, forming pressure-variable capacitors. Since this sensor is intended for use in the μCluster, it is important that the device be physically small and dissipate very little power.

The device is vacuum sealed, so barometric pressure deflects the diaphragm downward toward the reference electrode on the glass. The capacitance is inversely proportional to the gap distance between the boss and the reference electrode so that the capacitance and pressure sensitivity increase rapidly as the plates approach each other. However, the operating range of the device becomes very limited since the plates soon touch as the pressure increases. A series of diaphragms having slightly different diameters will deflect different amounts due to barometric pressure, however, and will hence cover different portions of the overall measurement range. As the plates touch, ending one measurement segment, the glass provides a built-in strain relief for that device, and the next smaller diaphragm will move down into the working gap range, continuing the measurement into the next higher subrange in pressure. Figure 10.9 summarizes this device operation. The 9.5-μm zero-pressure gap is reduced to a working distance of 0.4–0.7 μm over the pressure range of interest.

Atmospheric pressure is first measured using a pressure sensing element (S1) that is designed to cover a broad operating range (500–1000 torr ± 1 torr). Depending on this reading, one of the

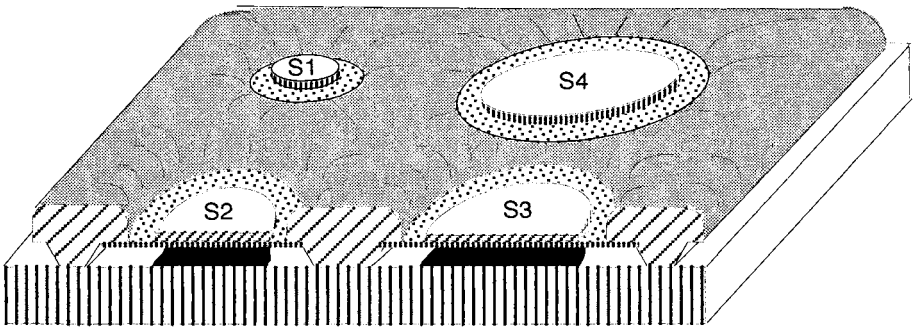


Fig. 10.8. The overall structure of the multi-element barometric pressure sensor.[4]

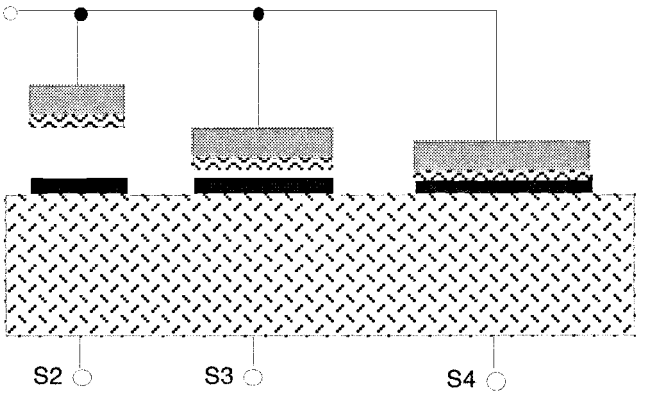


Fig. 10.9. Sensor operation. The multiple sensing elements of a device under a typical working condition are shown. The center device would be used for measurements here.[57]

other sensing elements is then selected to provide a higher sensitivity look at that particular subrange (e.g., 600–650, 650–700, or 700–760torr). The other pressure sensing elements will either have a larger gap separation or will be touching and hence strain relieved, as shown in Fig. 10.9. The capacitance of the selected element is read out using a switched-capacitor integrator[58] (Fig. 10.10) whose output is digitally compensated for temperature and nonlinearity effects. The switched-capacitor interface circuit is integrated in a 3-μm CMOS process on a 2.2 × 2.2-mm chip that also provides command decoding, control, temperature sensing, and interfacing through a sensor bus to the embedded microcontroller.[59] The pressure sensors are self-testing, using the applied electrostatic voltage derived by lengthening the readout pulse width.[58]

The pressure sensor is fabricated using a five-mask, silicon-bulk-micromachined, dissolved-wafer process[60] with a die size of 5 × 6 mm. The silicon processing starts with a p-type (100) oriented silicon wafer of normal thickness (550 μm). A KOH etch is first performed to define a recess that will later provide the capacitor gap, connecting channels among the different elements and the tunnels for the output leads. The KOH etch time and temperature determine the depth of

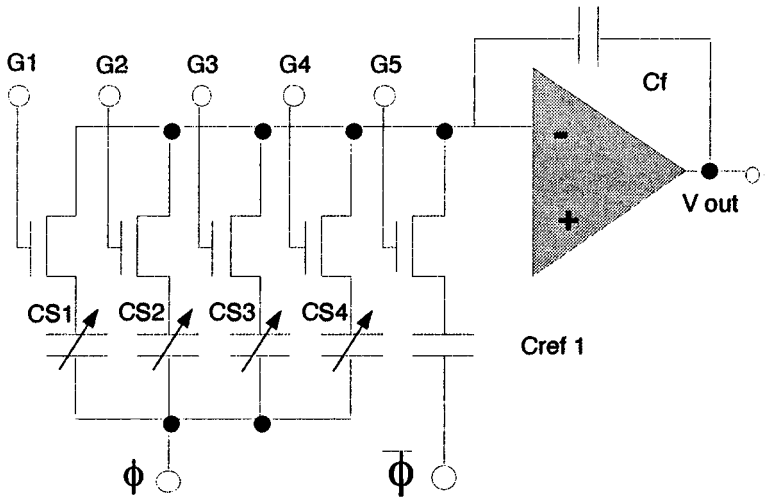


Fig. 10.10. The switched-capacitor integrator used for measuring the transducer output capacitance as a function of applied pressure.

the recess (nominally 9.5 μm). Note that this depth is not particularly critical, since by adding segments on either end of the measurement range, gap variations from device to device can be accommodated. Photolithography is next used to define the supporting rim and the center bosses, which are diffused to an etch-stop depth of 12 μm. The thin (2-μm) portion of the diaphragm is then formed using a shallow boron diffusion. A layer of LPCVD SiO_2 is then deposited and patterned on the diaphragm to prevent electrical shorts when the capacitor plates touch as well as to compensate the diaphragm internal stress associated with the boron diffusion. The glass processing consists of metallization to form the reference electrodes and the drilling of an optional hole in the glass substrate for use in vacuum sealing. Once the silicon and glass processing are completed, the silicon and glass wafers are electrostatically bonded together. The silicon wafer is then etched away in EDP, leaving only the heavily boron-doped areas on the glass. While a batch vacuum-sealing process involving deposited materials at wafer level has recently been developed,[60] the present devices used in the μCluster were sealed by first sealing the lateral lead tunnels and then placing the devices in a vacuum chamber subsequently evacuated to a pressure of less than 1 torr. A pass-through arm was then used to apply a vacuum sealant to the vent hole in the substrate. The sealant was then cured in vacuum.

Figure 10.11 shows the measured (uncompensated) pressure response for the "global"-sensing diaphragm, which spans the measurement range from 500–800 torr. It has a nearly linear, capacitance-pressure output characteristic with a pressure sensitivity of about 2 fF/torr (420 ppm/torr). Figure 10.12 shows measured responses from three of the other device elements. The sensing element having the largest diaphragm size gives output readings over a pressure range of 600–650 torr. The sensing element with the next smaller diaphragm size provides output in a pressure range of 650–700 torr, and the smallest diaphragm responds from 700–750 torr. Pressure sensitivities for these "segment" diaphragms are about 25 fF/torr, corresponding to about 4200 ppm/torr or equivalent to a resolution of about 1-ft altitude near sea level.

The barometric pressure sensor illustrates the utility of an embedded microprocessor along with software control, forming an integrated microsystem. The approximate pressure can be read

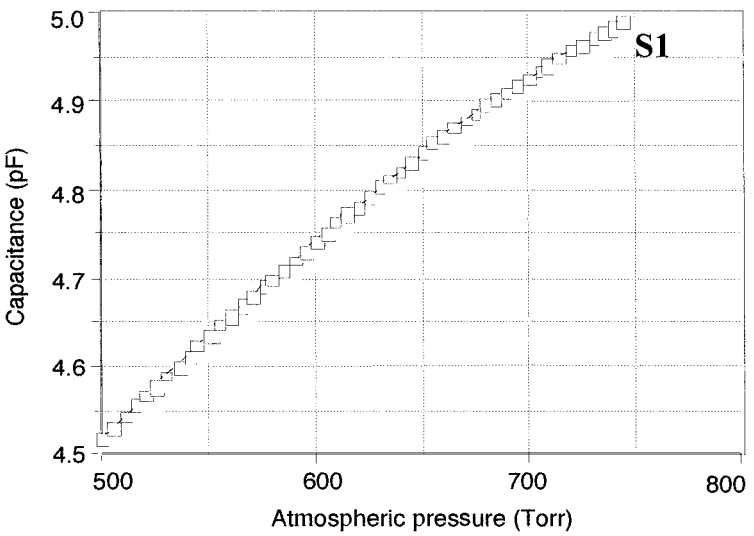


Fig. 10.11. Response of the global transducer spanning the overall barometric pressure range of the device. The pressure sensitivity is about 2 fF/torr

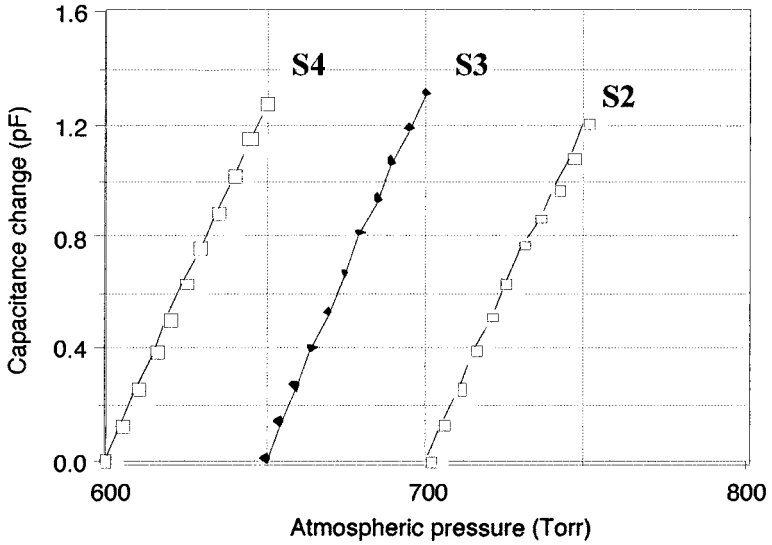


Fig. 10.12. Measured responses from the higher-resolution (segment) sensing elements. The pressure sensitivities are about 25 fF/torr.

using a global sensor, and a more exact reading can be obtained using the appropriate segment device. Digital compensation of the measured output allows an order-of-magnitude more resolution and accuracy than would laser trimming, and self-testing is possible to a considerable degree. Similar strategies can be employed for a broad range of other devices in such systems, including the following examples.

- Broad-range temperature sensor arrays
- Programmable frequency-range vibration accelerometers
- Broad-spectrum acoustic sensors
- Cross-correlated chemical identification devices and biological agent detectors

10.6.3.3.2 Micromachined Infrared Sensors

Throughout the 1970s and most of the 1980s, most work on IR imaging devices was done using compound semiconductors or silicon Schottky diode arrays operated at liquid nitrogen temperatures. While adequate for some applications, the cooling requirement made such devices inappropriate for most commercial applications. By the mid-1980s, however, silicon micromachined imagers capable of operating at ambient temperature had been demonstrated for process control applications,[43] and by the early 1990s higher-density imagers were emerging targeted at night-vision applications.[44] Figure 10.13 shows the cross section of one of the early imagers. This device uses a dielectric window for thermal isolation between an array of series-connected hot junctions and an array of cold junctions located on the chip rim. When incident radiation falls on the array, the hot junctions are warmed and the thermopile converts the temperature rise into an electrical output. Such devices allow a remote temperature resolution of better than 1°C with a thermal time constant of a few milliseconds and typical responsivities of 50–100 V/W using windows measuring 400 μm on a side. Such pixel sizes are adequate for process control, but do not permit the numbers of windows required for night-vision applications. More recently, surface micromachining has been used to realize much denser arrays having pixel sizes of 50 μm on a side with responsivities as high as 70,000 V/W and a typical noise-equivalent temperature

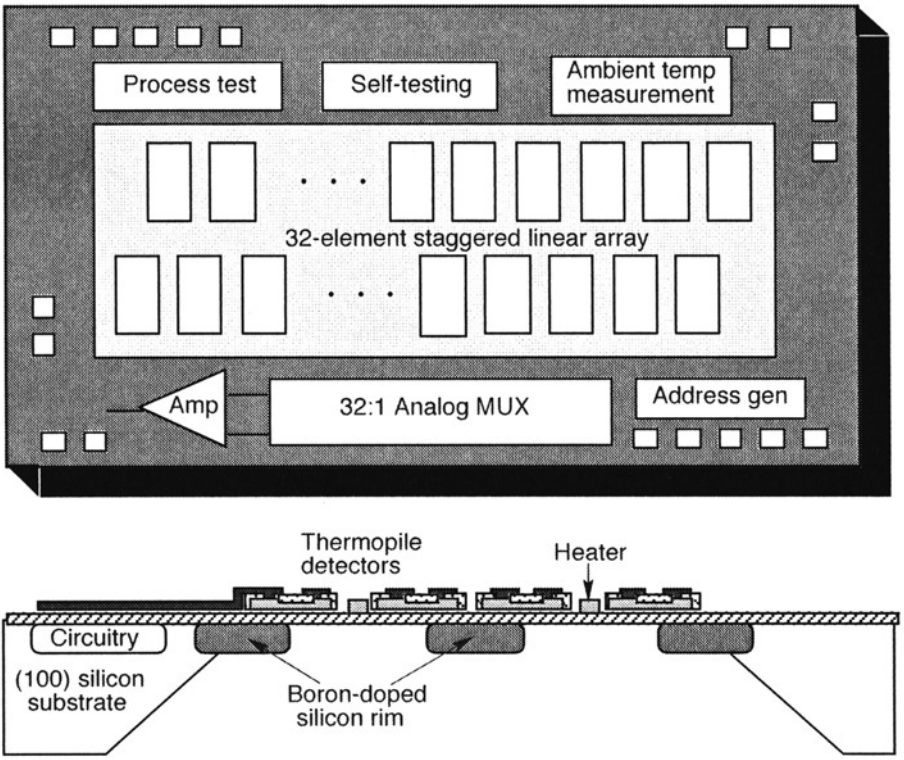


Fig. 10.13. Top view and cross-section of a micromachined thermal line imager. Thermocouples are supported on dielectric windows where the hot junctions are heated by incident radiation, converting input power into an electrical output.[61]

difference (NETD) of about 0.05°C.[43] Imagers composed of 80,000 pixels have shown excellent night-vision capabilities while operating at ambient temperature. Finally, active-pixel micromachined arrays using heated arrays with local feedback within the pixels have reported responsivities of more than 10^6 V/W.[62] Such arrays should find important applications in both ground- and space-based applications.

10.6.3.3.3 Micromachined Gas Sensors

One of the more promising approaches to chemical (gas) sensors is based on the structure shown in Fig. 10.14.[63] A dielectric window is again used, this time with a bulk micromachined silicon heater under it. On the top side of the structure is a thin deposited film (e.g., 35 Å-Pt on 50 Å-TiO_x) along with four electrodes to allow accurate measurement of its resistance. With an appropriately prepared film, gases such as oxygen and hydrogen can be detected with ppm accuracy. By accurately controlling the film temperature and ramping it over a predetermined range, temperature-programmed desorption (TPD) can be used as an aid in determining the gas or gases present. Microcalorimetry effects can also be employed for gas analysis in such structures. While selectivity remains a problem in these devices, a common approach is to use an array of dielectric window detectors with each one coated with a different conducting film so that the array produces a unique signature for a given gaseous mixture. Embedded microprocessors and/or neural networks can then in principle be used to deconvolve the signature to specify a unique gas

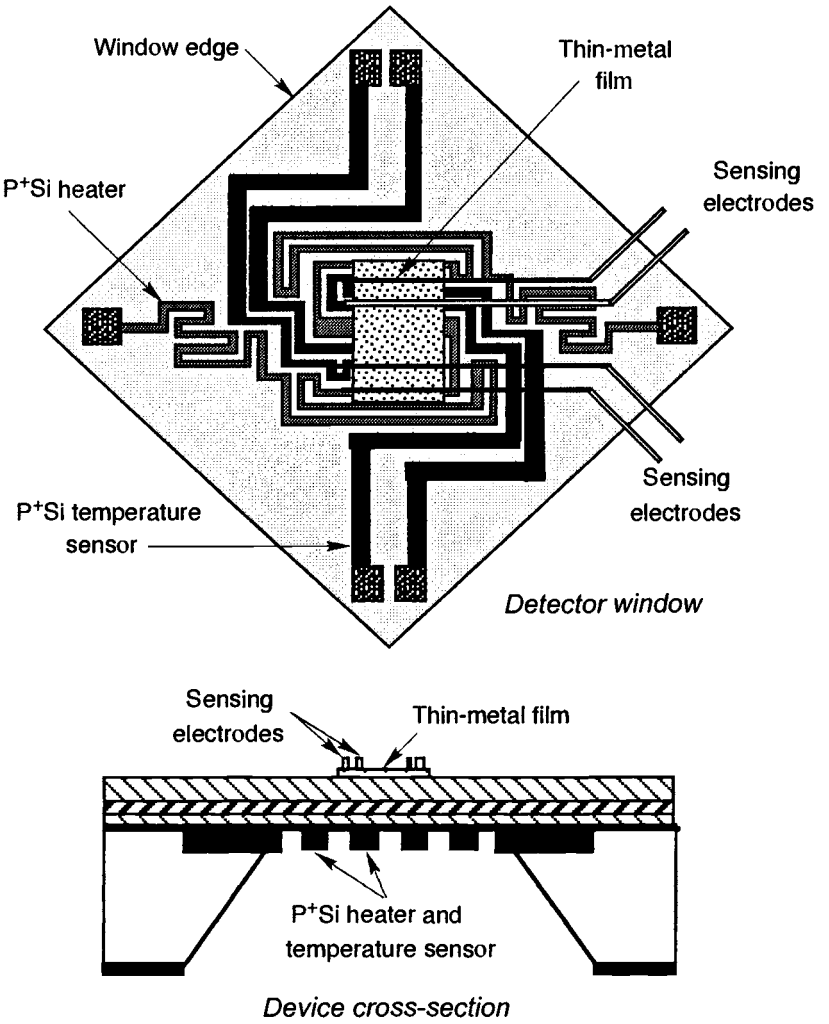


Fig. 10.14. Top view and cross-section of a micromachined gas detector. The device employs a heater suspended under a dielectric window. A thin film on top of the window changes its conductivity in response to gas adsorption from the surrounding environment.[63]

composition. Thus, such arrays are promising candidates for use in microsystems. Power must be conserved as much as possible in battery-powered applications by pulse-heating the windows, which have typical thermal efficiencies of 6°C /mW in air. It is significant that detector films can be deposited using chemical vapor deposition (CVD) by placing completed arrays in an appropriate gas stream and selectively heating the desired window to catalyze film growth.[64] The electrodes, which are already in place, can be used to terminate growth at the appropriate film resistance. This technique allows an array to be effectively programmed for a given application.

A still more versatile approach to analyzing gaseous mixtures can be based on gas chromatography. A number of research efforts focus on miniaturizing such devices using micromachining. As shown in Fig. 10.15, a series of microvalves forms a gas-sampling system that allows a sample of unknown gas to be injected into a carrier gas stream. As this sample passes through the column,

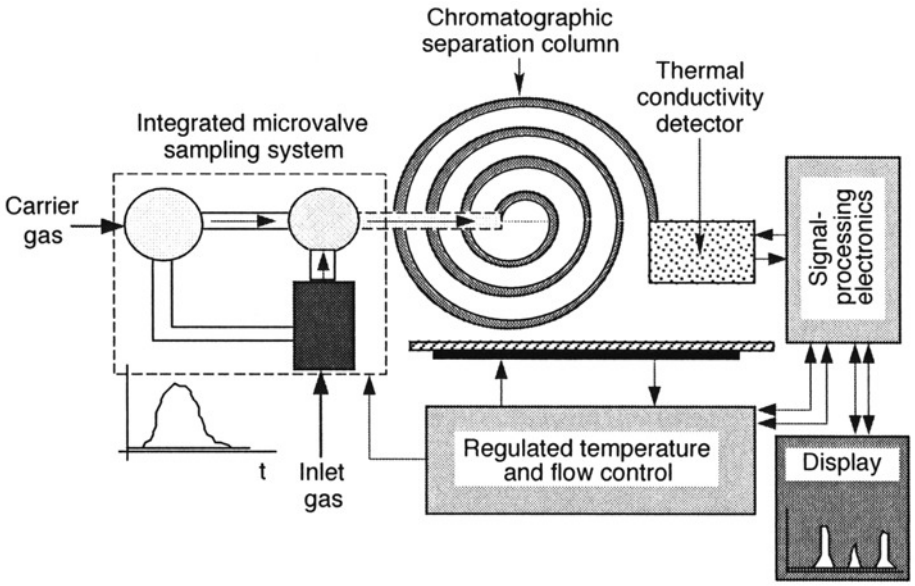


Fig. 10.15. Block diagram of an integrated gas chromatograph. A sample of unknown gas is injected into a carrier and passed through a long capillary tube, where the different molecular species are separated in time. Emerging gases are detected using a thermal conductivity detector and identified as to type by their characteristic delays. The amount of gas present can be determined from the integral of the detector response.[65]

different molecules spend different amounts of time stuck to the walls of the column, so that at the far end of the tube they emerge separated in time. They can then be detected using thermal conductivity sensors or by other means. While the first silicon gas chromatograph was reported almost 20 years ago,[66] microvalve technology at that time did not allow the realization of a fully integrated microsystem. Over the next few years, such microinstruments should emerge, and if incorporated within a microsystem unit such as the μCluster, should find wide application in military and civilian applications.

10.6.4 Other Components and Features of the μCluster

10.6.4.1 Sensor Bus

An important aspect of the μCluster's generic architecture is the standard sensor bus, which establishes a fixed mechanism by which the controller can communicate with the sensor network within the microsystem. Although a variety of network standards is available for use as the sensor bus, the one chosen for the μCluster is based on the Michigan Serial Standard[67] and was selected because it provides the necessary features while minimizing the number of signals and complexity of the bus. This, in turn, minimizes both the electronics required for interfacing to the bus as well as the trace count to reduce package size. Because it plays a significant role in the microsystem and affects many aspects of system operation, the sensor bus is discussed in detail in this section.

As shown in Fig. 10.16, the μCluster's sensor bus consists of three power lines, four signals for bidirectional serial communication, a shared data line for sensor output, and a data valid/interrupt signal. The three power lines are the ground reference (GND), the main system power (VDD), and a switched 5-V reference supply (VREF). VDD is defined as 6 V to be compatible with common battery voltages. This supply is on constantly while the μCluster is active. VREF,

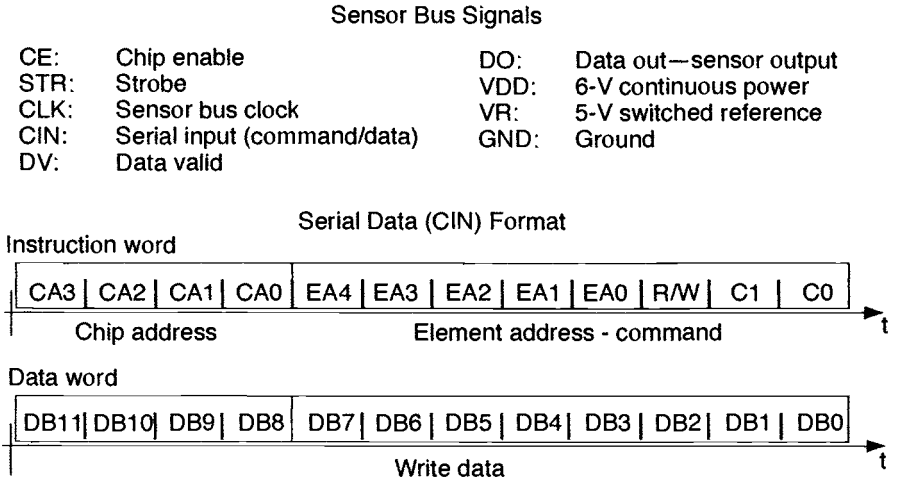


Fig. 10.16. The μCluster sensor bus provides a standard for communication between the controller and the front-end sensor network. Shown are the signal lines that make up the sensor bus as well as the format for the serial data used to issue instructions to sensors on the μCluster.

on the other hand, is switched on and off as part of a power-management scheme discussed in Subsec. 6.4.4.4. VREF is part of the sensor bus so that electronics on the sensor front end can be turned off when not in use. VREF is a 5-V reference that will remain constant even as the battery voltage, VDD, decreases over time. A constant reference voltage is necessary for the analog readout circuitry on the sensor front end. The sensor data output signal (DO) is shared by all sensor modules in the microsystem and multiplexed by the modules themselves. That is, each sensor module will have its output disabled from the DO line unless commanded by the controller to put data on the line. Data on the DO line can be in the form of an analog voltage, a serial bit stream, or frequency-encoded data (provided the controller can handle each of these formats). The data-valid (DV) signal is used to let the controller know when valid data is available on the DO line. DV also doubles as an interrupt that can be used to request communication with the controller as part of, for example, an event-triggered response. Use of the DV interrupt provides the μCluster with the capability of on-demand sensor readings rather than being limited to simple preset sensor scans.

The four serial communication signals on the sensor bus are a chip-enable signal (CE), a data strobe (STR), a serial data line (CIN), and a clock (CLK) signal. A timing diagram for these serial communication lines is shown in Fig. 10.17. Because of the nature of the sensor bus, communication can only take place with one front-end sensor module at a time. When active (high), the CE signal indicates that the sensor bus is in use by a specific sensor module so that no other module can interrupt the system until this line goes inactive (low). The STR signal defines a window around the commands being sent from the controller to a specific sensing node so that bus interface hardware will know when a message begins and ends. Used in combination, the CE and STR signals tell the sensor modules when they should be reading or ignoring incoming data. Command and data bits sent from the controller to the sensor modules are placed on the CIN line, and the bits are synchronized to the CLK signal so that they can be easily read by bus interface hardware. The data format for information on the CIN line is shown in Fig. 10.16.

At the beginning of a communication from the controller, the CE bit goes high, and a 4-bit chip address is put out on the serial data line, CIN. Each sensor module has a predefined chip address

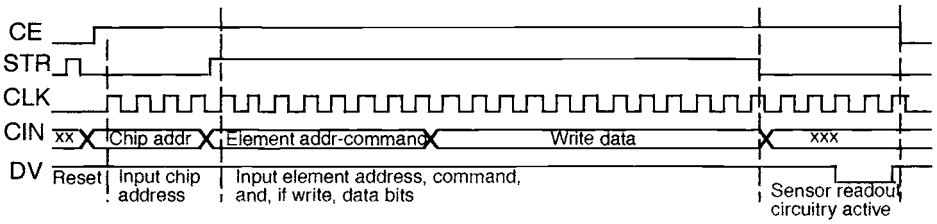
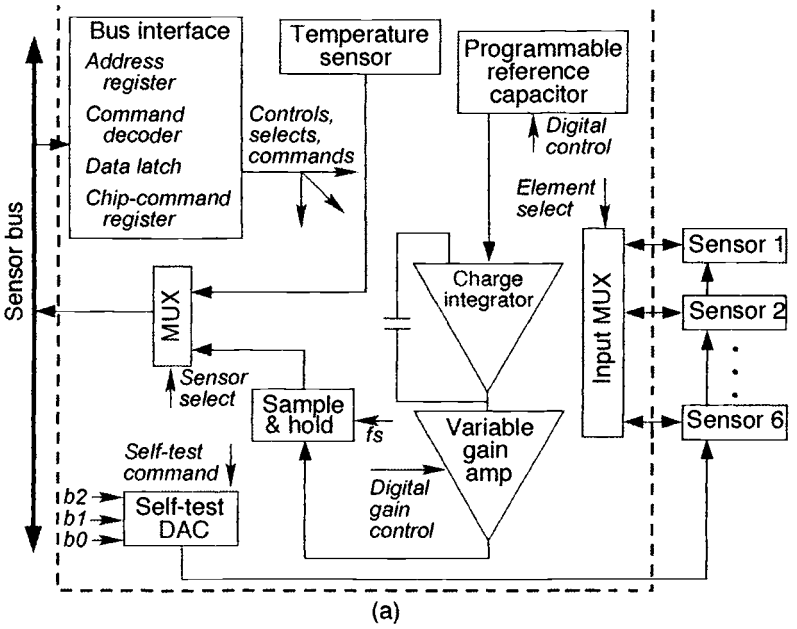


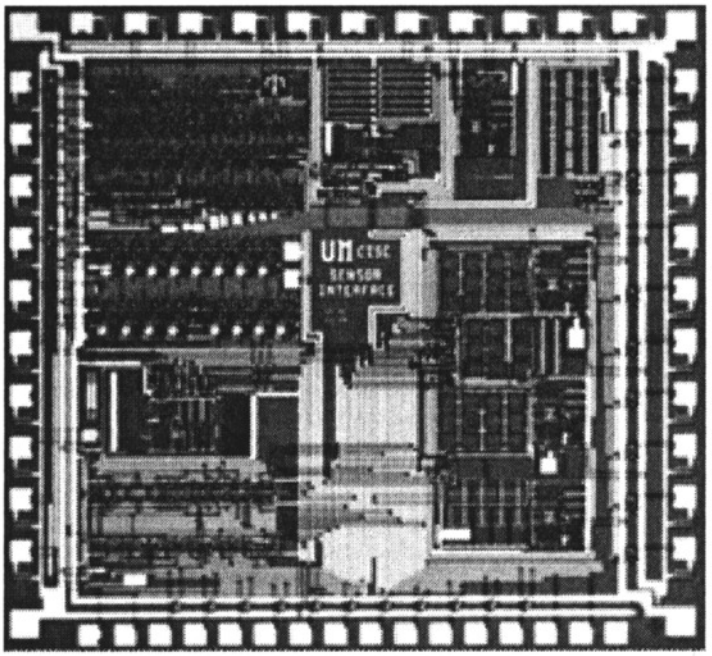
Fig. 10.17. Timing diagram of the serial communication signals on the sensor bus.

and will only listen to a message if it begins with a matching chip address. At the end of the chip address, the STR line goes high to mark the beginning of a sensor-specific message. This message begins with a 3-bit command code followed by a 5-bit element address. The 5-bit element address can be used to access up to 32 readable elements (e.g., sensors) and 32 writable elements (e.g., actuators, digital registers) per sensor module. This format allows multiple sensors to share the same bus interface hardware, or allows each element in a multielement sensor array to be accessed individually. If the command is a write instruction, data bits follow the element address to be stored or used directly by the sensor interface circuit. At the end of this message the STR signal will go low. If the command is a read instruction, the sensor readout circuitry must convert the sensor data into one of the possible output formats. During this time, the CLK line is available to the readout circuitry and can be programmed to any frequency and duration (limited only by the controller generating the clock signal) as required by the specific readout circuitry. Once the readout is complete, the bus interface hardware places the data on the DO line and indicates that valid data is available by pulling down the active low DV line. Once the data has been read by the controller, the CE line goes inactive to reset the bus interface hardware and disable the output on the DO line. At this time the controller can either monitor the DV line for an interrupt or initiate another communication to a sensor module.

The sensor bus described is a generic form. For use on the μCluster, a more specific implementation was developed, for which appropriate command codes and element address bits were defined and a corresponding bus interface circuit was designed. The bus interface circuitry was implemented as part of the Capacitive Sensor Interface Chip (CSIC) that includes additional circuitry for the readout of capacitive sensors and an on-board temperature sensor.[59] A block diagram of the CSIC is shown in Fig. 10.18. For application with the μCluster and the CSIC, the 3-bit command code is used to define one of five command options implemented by the CSIC. Of these commands, one is a read instruction and the other four are write instructions that access different data registers, including a 4-bit Chip Command Register (CCR). The CCR is a 4-bit, on-chip control register that can be accessed without writing to a specific element address. This reduces hardware overhead and simplifies setting the four primary control bits used by the sensor readout circuitry. The sensor readout circuitry on the CSIC is a three-stage switched-capacitor circuit[68] that can convert capacitive sensor data into an analog voltage read by the microsystem controller. This circuit uses the five-bit element address to determine which of six possible sensor inputs will be read and which of four possible reference capacitors will be used during the data conversion. The on-chip reference capacitors provide programmable offset control and allow the CSIC to be used with sensors that have nominal capacitance in the range of 1 pF to 12 pF. The multiplexed sensor inputs allow one CSIC to be used with up to six sensing elements. The CSIC also has programmable gain settings that use the bits of the CCR to control the gain of the sensor readout circuit. The programmable gain allows the CSIC to accommodate sensors with a wide range of sensitivities. The final block of the CSIC is a simple temperature sensor formed by a ring



(a)



(b)

Fig. 10.18. (a) Block diagram of the Capacitive Sensor Interface Chip (CSIC). CSIC includes circuitry that will interface to the sensor bus, readout multiple capacitive sensors, and measure the temperature of the chip and its surroundings. (b) Photograph of a fabricated CSIC die, which has been laid out approximately as shown in the diagram.

oscillator with an output frequency that varies with temperature. The temperature sensor on the interface chip provides an accurate measurement of the temperature in close physical proximity to other sensors on the μCluster. This makes the temperature sensor very useful for digital compensation of temperature sensitivities in other sensors as discussed in Subsec. 10.7.4.

10.6.4.2 Microcontroller

As discussed earlier, a large variety of commercially available controllers may be used in microsystem applications. For use with the μCluster, the Motorola 68HC711E9 Microcontroller Unit (MCU) was chosen as the best of the available alternatives because it offers the following features on a single 8 × 10-mm die:

- An 8-bit microprocessor
- An 8-bit analog-to-digital converter
- Timing hardware
- A synchronous serial peripheral interface (SPI)
- An asynchronous serial communications interface (SCI)
- Built-in memory: 512 bytes RAM, 512 bytes EEPROM, and 12 kbytes EPROM
- Low-power mode

Each of these features is necessary in order to be compatible with the goals and requirements of the μCluster:

- The microprocessor is needed to execute control software that can be stored in the on-chip memory.
- The A/D and timing hardware are necessary to read out sensor data in the analog and frequency-encoded formats provided in the sensor bus definition.
- The SPI allows for high-speed serial data transfer across the sensor bus with minimal software overhead.
- The SCI is a universal asynchronous receiver transmitter (UART) type interface that provides a well-known standard for communication between the microsystem and external host system.
- In addition to storing control software, the on-chip memory includes EEPROM, which is very useful for storing sensor-specific data within each microsystem
- The low-power mode available on the 68HC11 MCU is vital in minimizing the power consumption of the μCluster.
- The 68HC11 is available in die form necessary for the multichip module packaging used for the μCluster.

Many other available controllers offer some, but not all, of these required features. However, as the popularity of microsystems grows in the coming years, manufacturers are sure to answer the demand with other controllers that offer these features and more. There already is evidence that new controllers are being specifically designed for sensing systems. The limiting aspects of the 68HC11 are its 8-bit architecture, relatively slow speed (a maximum bus speed of 2 MHz), and the 8-bit A/D. There is a need for a wider processor word (e.g., 16 bits) and higher A/D accuracy (≥12 bits), but these advances need to be achieved while reducing the overall power dissipation. This performance may be possible if the generality of the processor is reduced to eliminate unneeded functions, emphasizing the sensing/control functions discussed here.

10.6.4.3 External Communication

Although a primary goal of the μCluster is to utilize wireless communication so that the host system can receive data from a remote site, it is also important to have a hardwired interface to the external world. The hardwired interface is useful for developmental purposes since it allows the

user to program the controller and examine and debug system operations easily. It is also useful in many applications where a hardwired link is preferred because it eliminates the need for a matching receiver to the wireless transmitter. In the design of the μCluster, it was only necessary to provide wireless data output, so all inputs utilize a hardwired interface. Since both a wireless and hardwired link will normally exist, the two links should be as similar as possible. Using a well-known standard for external communication will minimize the interface hardware necessary between the μCluster and the host system.

Given these guidelines and the fact that the MCU chosen for the μCluster has a built-in RS-232-compatible UART type interface, the asynchronous serial RS-232 format was chosen for external I/O. For direct connection to the serial port of a common PC, only a widely available level-shifter buffer is needed to convert the 6-V signals of the μCluster to the 12-V levels of the PC. This standard was chosen more for its simplicity than its functionality. Indeed, there are many other network standards available that may be more appropriate for a given macrosystem. Note that the RS-232 does not limit the μCluster to a macrosystem based on this network scheme; an external network conversion interface could be used to convert the RS-232 format to an alternate network protocol without affecting the inner workings of the μCluster.

For the wireless transmitter, the HX1005 by RFM, Inc., was selected. This component has typically been used for keyless entry systems in automotive applications. The advantages of this component were its small size and low-power consumption, which were found to be more compatible with the demanding goals of the μCluster than other commercial components. The HX1005 is an RF transmitter with a carrier frequency of 315 MHz and is compatible with the UART communication interface chosen for data output from the μCluster. By using amplitude-shift-keyed modulation, the same data sent to the UART could be directed to the transmitter for wireless data output. Data at the receiver end can be easily converted to an RS-232 format. This makes the method of data transfer, whether hardwired or wireless, transparent to the host system. The range of transmission from the μCluster using the HX1005 is largely dependent on the receiver and the type of antenna used. A range of more that 100 ft has been observed using a low-power receiver that runs off of four AA batteries. Greater range can be obtained at the cost of increased power, both on the μCluster and at the receiver.

10.6.4.4 Power Management

A primary goal of the μCluster was to provide all desired sensing functions while minimizing power consumption. Options for managing the power consumption can be reduced to two approaches: minimizing the power consumption of each component in the microsystem and keeping all unnecessary components turned off until needed. The first of these is relatively simple from a system point of view: choose components for each functional block that have the lowest possible power consumption. Of course, this may mean facing difficult trade-offs in performance that have to be analyzed thoroughly. Most of the difficult issues in this area are faced in the design of the individual components and, as such, are beyond the scope of this chapter. It is worth noting, however, that capacitive sensors provide a tremendous advantage over piezoresistive devices since the only power they consume is that of the readout circuitry, which can be minimized by careful design. For this reason, the pressure, humidity, and acceleration sensors used on the μCluster are all capacitive.

In the second approach—keeping all unnecessary components turned off until needed—power management becomes more of a system issue. Power management is addressed in the sensor bus, which includes a voltage reference that can be switched off when the sensor front end is not being used. If there are sensors or subsets of sensor nodes that need constant power, there is also a

supply voltage on the sensor bus that is always on. The μCluster has an advantage here over some other sensing systems in that most of the parameters being measured have very low bandwidth and as such do not need to be monitored often. Measuring barometric pressure, relative humidity, or temperature can be done, for instance, once a second without significant loss of relevant data. However, in the case of acceleration, which can generate signals with significantly broader bandwidth, an alternate approach is necessary. Again the sensor bus provides for this by allowing for event-triggered interrupts between normal sensor scans. Using this feature, a threshold accelerometer, which sets a mechanical switch when an acceleration beyond a certain threshold is experienced, can be used to interrupt the controller and request an immediate reading of a more accurate on-board continuous accelerometer.[69] Thus, only a low-power digital circuit that monitors the threshold device and generates the interrupt needs to be on continuously. A similar approach can be used by many other sensors that might normally be too power hungry for such low-power systems as the μCluster.

In addition to switching off the sensor front end between sensor scans, it is also necessary to minimize the power consumption of the MCU, which in the case of the μCluster is, by far, the largest power consumer in the system. Since the sensing activities of the μCluster are of very low frequency, there is a large block of time between sensor scans when it is not necessary to have the MCU running. As long as some circuitry monitors the sensor-bus interrupt signal, the MCU can be shut down just as the sensor front end is. To accomplish this, the built-in MCU low-power mode is utilized. In this mode, the clocks on the MCU are shut off, all outputs are latched to their current values, and data in RAM is maintained, while the power levels are reduced by three orders of magnitude. A similar mode is available on many controllers. The disadvantages of this mode are that the MCU is nonfunctional, and it must be restarted by an external signal. It is therefore necessary to include some control circuitry external to the MCU that will monitor microsystem activities while the MCU is in low-power mode and will wake the MCU after the appropriate delay.

A power management chip (PMC) was developed for use in the μCluster to accomplish this task. The PMC has on-chip clocking circuitry that is activated by the MCU before it enters its low-power mode. The MCU sends a code to the PMC that defines the desired delay, which can be one of eight discrete values between 15 s and 5 min. After this, the MCU goes to sleep and the PMC is in control until the delay time is over. During this delay it is necessary for the PMC to monitor the sensor bus in case an event-triggered interrupt is generated. When either an interrupt is received or the delay time expires, the PMC wakes the MCU, which takes back control of system operations.

A block diagram of the PMC is shown in Fig. 10.19. As shown in this illustration, the PMC also contains switching circuitry that controls the delivery of power to other areas of the system. These switches are set by inputs from the MCU, and they control whether the transmitter gets power and whether the 5-V reference of the sensor bus is on or off. The 5-V reference is implemented with an Analog Devices REF-195. This component is compatible with the goals of the μCluster and is available in die form. The PMC is, in a sense, an extension of the MCU, under its direct control, and used to control and minimize the power consumption of the μCluster.

10.6.5 Summary of Microinstrumentation Cluster Design

Having defined a generic microsystem architecture and discussed many aspects of the components to be used in one specific realization (the μCluster), we will review some of the trade-offs encountered and the decisions affected by some choices. Designing the μCluster began with determining a generic microsystem architecture and defining a standard sensor bus. Both aspects

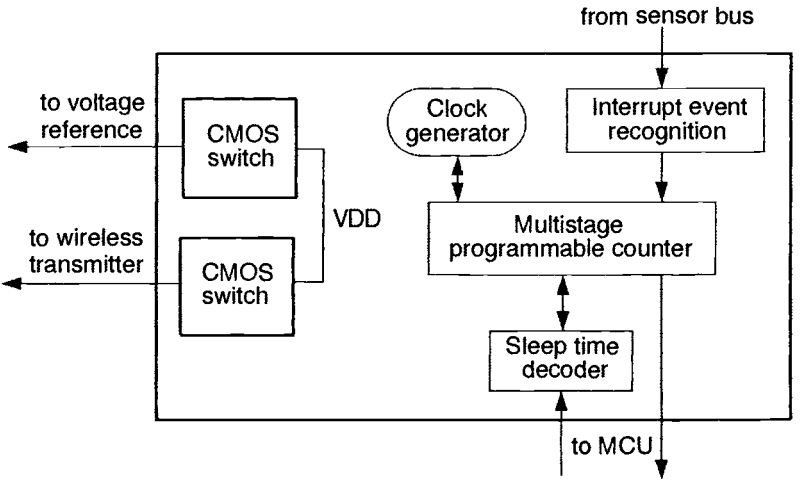


Fig. 10.19. Block diagram of the power-management chip used to control power flow across the μCluster and monitor system activities while the MCU is in low-power mode.

placed specific requirements on the microsystem controller, which is important to select early in the design because many other design choices relate to this component.

The controller had to communicate across the defined sensor bus for intramodule operations and across an external network bus for interaction with a host system; it had to read analog, digital, and frequency-encoded data from the sensor front end; and it required a nonvolatile memory device in which to store control software and sensor variables. Although these functions could be obtained by separate electronic components, it is much better if they are built into a single controller, especially when size is an issue. Once a controller that met these requirements was selected, the sensor interface circuitry could be designed.

The sensor readout circuitry on the interface chip was designed to meet the requirements of several MEMS transducers already under development for measuring the targeted environmental parameters. After designing the sensor-interface circuitry, the transducers could be modified, if necessary, to meet the interface circuit requirements. For example, the nominal capacitance of the transducer had to fall into the range supported by the sensor readout circuitry on the interface chip.

Meanwhile, as iterations between interface circuitry and transducers were being made, other system-level issues such as power management and external I/O could be addressed. Determining the appropriate power-management schemes revealed that hardware in addition to the controller would be necessary. The power-management chip was then designed, and a commercial 5-V reference selected. These two components contain the necessary control circuitry outside of the controller itself.

At the same time that the power-management chip, the sensor interface chip, and the transducers were being designed and fabricated, the wireless link was investigated. After an initial attempt to design a custom transmitter failed to meet the size and power budget of the μCluster, a commercially available alternative was sought. The transmitter had to be compatible with both the power levels available on the μCluster and the inputs possible from the chosen controller. With the identification of an appropriate component, the design phase was complete. Making a complete system out of the various components is the subject of the next section.

10.7 Fabrication and Testing of the μCluster

10.7.1 Fabrication

Fabrication of a complete microsystem such as the μCluster involves many different efforts. Organizing these efforts so that work can be performed in parallel provides greater efficiency. While some of the specific tasks may vary with the project, the plan for fabrication of the μCluster provides a good example of how to structure the necessary efforts. Figure 10.20 shows a time-based production map of the activities in constructing the prototype microsystem. The activities are divided into three primary functional efforts dealing with the sensor front end (transducers and interface electronics), the system hardware (MCU, I/O, etc.), and software. They cover the project from basic transducer and electronics design to overall system operation, packaging, and testing. Because different people were responsible for different aspects of the project, the map is useful in illustrating the interrelation of activities and the order in which tasks must be completed.

At the beginning of the map in Fig. 10.20 are activities such as design, fabrication, and selection of the components that form the μCluster. Selection of the MCU directly relates to many of the other tasks, including, in this case, the design of floating-point math software routines to provide more accurate calculations than normally allowed by an 8-bit microprocessor. Near the middle of the map, all the components are available, including a set of "known good die" for the custom devices used. The later part of the map shows the final development of a prototype unit with packaging and testing. After the prototype is checked, and assuming that no redesign is

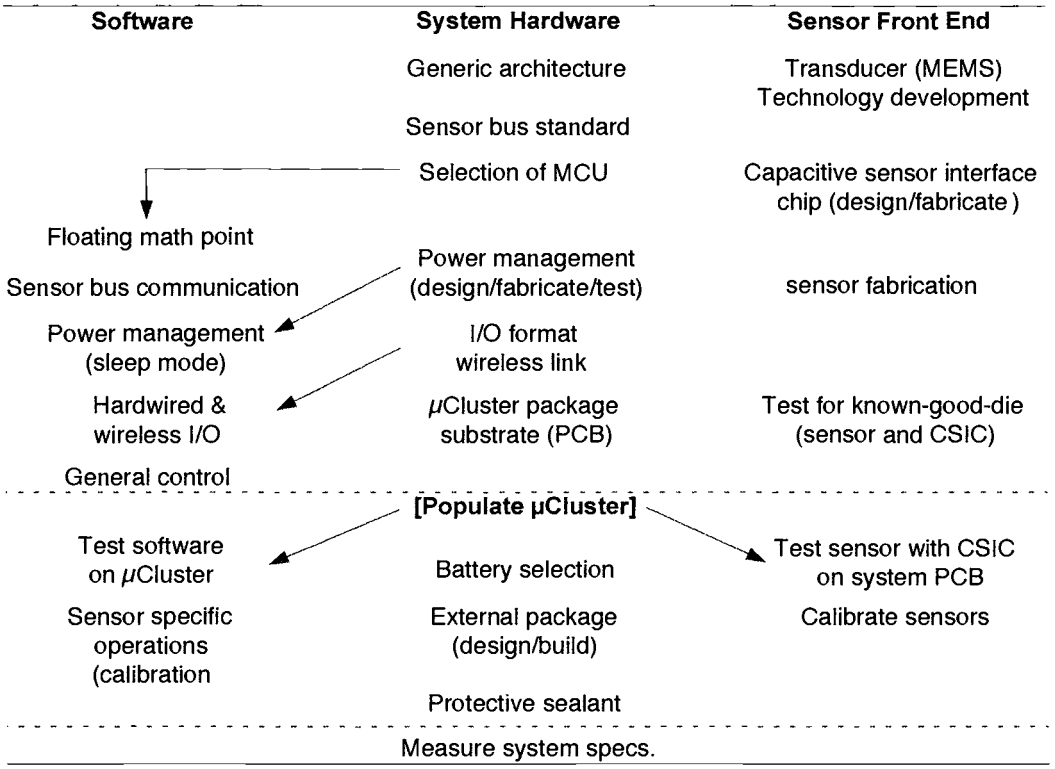


Fig. 10.20. Map of tasks to be performed during the fabrication of the μCluster. The tasks are mapped from top to bottom in the order they must be completed. The interrelation of parallel activities are indicated by arrows connecting the tasks.

necessary, multi-unit fabrication can begin. This involves keeping a supply of known good components and package elements while further units are produced and tested. During this production stage, additional modifications or enhancements to the system software can take place. Although the map is specific to the development of the μCluster, it is consistent with the work that will be necessary in similar microsystem projects.

10.7.2 Packaging

Three areas of packaging were investigated: sensor packaging, multichip packaging, and external system packaging. In sensor packaging, one fundamental question is whether to fabricate the sensors as hybrid devices with separate die for the transducer and the readout/interface electronics or whether to make them monolithic, all on the same integrated circuit die. Advantages of a hybrid design are higher yield (two chips can be fabricated and tested individually), independent iteration of the design of each chip, and reduced process complexity. Advantages of a monolithic design are reduced interconnect requirements and improved performance by having the readout circuit on the same chip as the transducer. In general, hybrid designs make system-level development more difficult; while monolithic designs make component-level development more difficult. In most cases, performance will be better with monolithic devices. However, for the initial μCluster, a hybrid design allowed designers of the state-of-the-art transducers the freedom to work without the encumbrance of interface circuitry.

A most important area in sensors today is monolithic packaging technology used to build the first-level package directly on the chip with the transducers. These packages use deposited thin films such as silicon dioxide and silicon nitride, possibly with metal barriers against moisture and chemical contaminants from the environment. In some implantable sensors for biomedical applications, these films are the only package possible.[70] As the technology is further developed, silicon carbide and diamond films may be used for chip-level packaging as well. In addition, chemical vapor deposition (CVD),[71] silicon-glass anodic bonding,[72] and silicon-silicon fusion bonding[52] are used to form wafer-level vacuum cavities for devices such as accelerometers, where the variable can be coupled through a hermetically sealed microstructure.

For hybrid devices, sensor packaging remains an important issue. At this level, the sensor package is a surface on which the hybrid sensors and interface components are mounted and connected using flip-chip or wire bond techniques. Packaging also consists of sealing the chips and placing them in an outer shell to protect against the environment (moisture, dust, radiation, etc.). This multichip-module (MCM) approach minimizes the size of the overall system by eliminating intermediate chip packaging and brings the components together as closely as possible on a common substrate. The approach also allows different technologies to be used as needed (e.g., monolithic sensors, silicon-on-glass hybrids, or silicon-silicon fusion-bonded structures). The package technology chosen for a given microsystem depends on the application and may also depend on the number of units to be produced. In the case of the μCluster, the packaging options were limited by the available in-house equipment. A multilayer printed circuit board (PCB) was chosen as the substrate for the μCluster since it allowed components to be easily attached and connected either by wire bonding or soldering. This approach was also good for prototype development because the board could be altered much more easily than in some other MCM technologies that bury components deep within the package.

The PCB for the μCluster was designed such that all the components, including the sensors, the wireless transmitter, and a 10-pin hardwired I/O connector, are attached on the top of the board. This gave easy access to test points on the board and simplified the replacement of components. Figure 10.21 shows a populated μCluster PCB with the components labeled. Contact

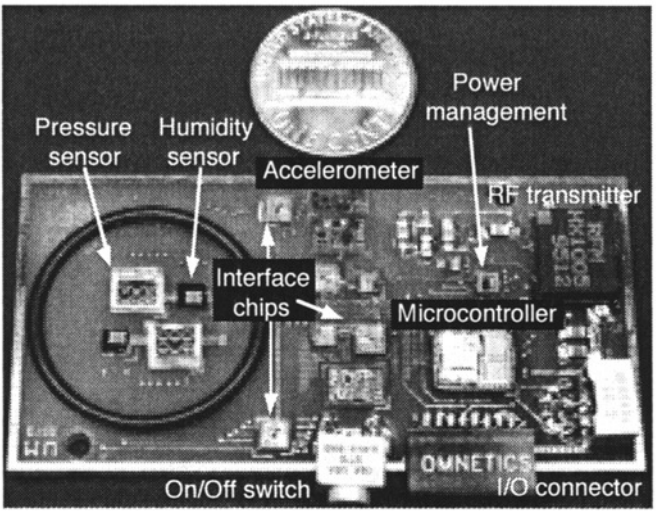


Fig. 10.21. A populated μCluster PCB showing the layout of the components on the system.

points for two 3-V coin-cell batteries are on the back of the board. Because some of the sensors on the μCluster require access to the environment that they monitor, the entire μCluster could not be sealed in an external package, and yet most of the electronics should be sealed from environmental effects. To accommodate these conflicting needs, the package designed for the μCluster includes an O-ring that is sealed to the PCB. The O-ring works with the external package to isolate the sensing elements that need environmental access from the rest of the system. A port on the external package has an open path between the area within the O-ring and the outside environment. The rest of the system is sealed from the outside environment by the external package, which for the μCluster is an anodized aluminum case. This material can easily be machined to the desired form and provides good electrical isolation. The antenna for the wireless transmitter passes through and wraps around the external package. A wrist strap can be attached to the package so that the μCluster can we worn in wrist-watch fashion as shown in Fig. 10.22.

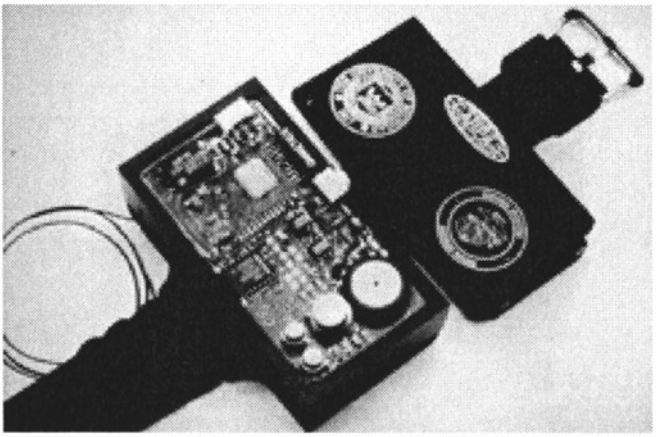


Fig. 10.22. A packaged μCluster in an anodized aluminum case with a wrist strap, which makes the system a wearable unit.

10.7.3 Testing

During development of the μCluster a number of testing steps were identified to provide final sensor calibration and system specification. The first-level tests were to find known good die for the components used on the μCluster. Commercial components have usually already undergone this type of test, but all custom components had to be tested after fabrication so that only components known to work within specifications would populate the system. After these tests, the system could be populated and checked for basic functionality. The following series of preliminary tests was run to ensure all the components were operating and communicating as expected.

- External communication with the MCU
- Read and write of the MCU memory
- Communication between the MCU and the power management chip
- Power management functions (power switching, sleep mode, etc.)
- Communication between the MCU and the sensor interface chip(s)
- Wireless transmission of test data

After these preliminary tests, the μCluster must undergo a variety of tests designed to calibrate and compensate each of the sensors on the system. Here we will define calibration as setting the electrical output of the sensor to the appropriate level over the desired range of measurement, that is, adjusting gains and offsets, and then compensating to remove cross-parameter (temperature) sensitivities. To calibrate each sensor, tests were first run to determine the appropriate gain and offset adjustments in the capacitive sensor interface chip (CSIC). To tune the response more finely, the CSIC could then be laser-trimmed to provide the maximum sensitivity for each sensor. If any trouble with either the transducer or the CSIC was found at this time, the component could be removed and replaced. After these initial sensor tests, a protective coating had to be placed over the wire bonds connecting the sensor to its readout chip in order to protect it from environmental effects such as humidity. Unfortunately, this coating would often cause a shift in sensor output by altering the parasitic capacitance associated with the transducer inputs, necessitating additional final adjustments that must be made electronically. Similar system-level tests were then performed to ensure no problems occurred during the packaging process. After this, the system was tested to measure parameters such as power consumption (and battery life), wireless transmission range, and sensor performance over an environmental temperature range (−20°C to +50°C). The results of such tests are given below.

10.7.4 Calibration and Compensation

One of the primary goals of this particular microsystem was to demonstrate a fully calibrated sensing system capable of delivering data to a host system that required no further data processing. This feature frees the host system to do higher-level operations. To accomplish this goal the μCluster uses a combination of hardware and software techniques. The hardware techniques have been discussed and involve electronic and laser trimming of the sensor readout circuitry to set the gain and offset of the switched capacitor circuit. However, these techniques are of limited use since they only provide coarse adjustments, do not sufficiently address nonlinear effects, and do not correct for undesired temperature sensitivities. Digital compensation, where the data corrections are performed in software, offers the ability to perform all these adjustments and can result in an order-of-magnitude improvement in device accuracy.[73]

Three traditional methods of implementing digital compensation are look-up tables, polynomial computation, and some combination of the two. Choice of approaches is largely determined by the degree of linearity displayed by the sensor, the required accuracy, and the time it takes to

complete the digital calibration task. (This last issue is of special importance in low-power systems such as the μCluster.) There are highly developed design-of-experiments procedures that allow the number of test points to be minimized and positioned for a given degree of accuracy.[74]

Except for the temperature sensor, which uses a combination method, the sensors on the μCluster employ polynomial compensation techniques. The accelerometer uses the most basic form of polynomial because it has a highly linear response. Therefore, a simple $y = mx + b$ equation can be applied, where x is the measured sensor output, m and b are coefficients determined by testing each accelerometer, and y is the resulting acceleration. By storing values for m and b in the memory of the MCU, the proper acceleration level can be obtained with a simple two-step mathematical operation on the measured data. Calibration of the accelerometer is further simplified because it does not display a significant temperature dependence over the desired operating range and can, therefore, be calibrated by a one-variable equation.

The other sensor on the μCluster that can be calibrated by a one-variable equation is the temperature sensor; however, this sensor is highly nonlinear, with an output response that fits a logarithmic curve, making calibration much more complex. If polynomial compensation is used, a 5th-order equation is necessary to obtain the desired accuracy from this sensor. At a minimum, evaluating the 5th-order equation requires nine multiplications and five additions and uses six coefficients. This can require significant MCU processing time, reducing system bandwidth and increasing power consumption. An alternative approach, taken with the temperature sensor, is to break the sensor output into multiple segments and then fit a lower-order polynomial to each segment. With this method the MCU need only compare the measured data to a set of segment values to determine which segment equation to use. For this sensor a 2nd-order polynomial in each of four 20°C segments was as accurate as a 5th-order polynomial over the entire range, yet required only three multiplications, two additions, and 12 coefficients. Another scheme, shown in Fig. 10.23, uses linear equations over each of eight 10°C segments requiring 1 multiplication, 1 addition, and 16 coefficients for the same accuracy. Note in Fig. 10.22 that the lines form a logarithmic curve as should be expected. Similar methods could be used by many transducers, but the chosen method must match the needs of the system and the specific sensor.

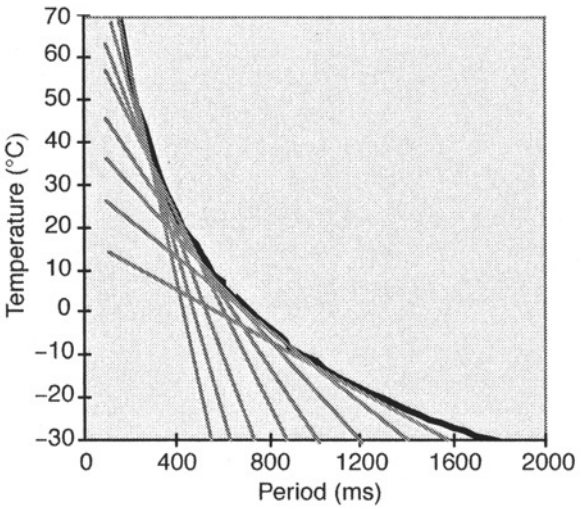


Fig. 10.23. Temperature sensor output as it is fit by linear equations each covering a 10°C segment. The actual response of the sensor is shown by the darker line.[65]

The calibration method used on the μCluster pressure and humidity sensors is a two-variable 4th-order polynomial that has been fit to a set of data taken over temperature and the sensed parameter (pressure, humidity). This equation has 25 terms to be evaluated and summed in calculating the true output. However, many of these terms have negligible effect on the calculation and can be ignored, reducing the equation to about 10 to 15 significant terms. The advantage of this digital calibration approach is that even a very nonlinear sensor response can be accurately tracked across the measurement range. Figure 10.24 shows an example of a pressure sensor output as a function of pressure and temperature that is tracked very well by the 4th-order compensation polynomial.

10.7.5 Final Results

Table 10.6 summarizes the characteristics of the University of Michigan Microinstrumentation Cluster. This multiparameter sensing microsystem can be seen in Figs. 10.21 and 10.22. To provide feedback regarding the design and operation of the μCluster, several units have been assembled, calibrated, and packaged. These units are either in use or available for use by other research agencies where field trials and experiments are being conducted to evaluate the performance of the μCluster. μCluster units at the Naval Research Laboratory (Washington, D.C.) are being flown on unmanned air vehicles. Units have also been delivered to researchers at the Naval Command, Control, and Ocean Surveillance Center, where they will be deployed on ocean buoys to measure environmental parameters. Previously, μCluster units have been taken on field exercises with the U.S. Marine Corps (Fig. 10.25). Here field trials were conducted to evaluate the μCluster's ability to provide meteorological data used for calculations on the trajectories of artillery fire.

Ongoing research efforts seek to improve the performance of both the microsystem and the integrated sensors employed on the system. Additionally, a variety of new sensors are being investigated for use in future microsystems. These new microsensors will be compatible with the low-power and small-size goals of the μCluster while adding the ability to monitor acoustic and

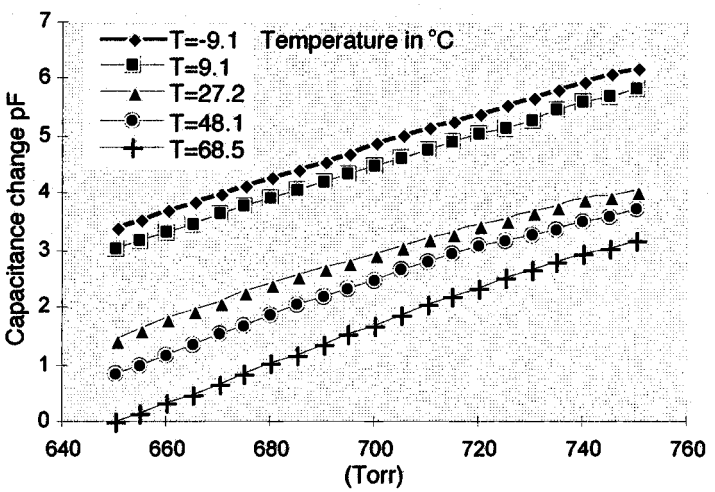


Fig. 10.24. Barometric pressure sensor output at various temperatures.[60] Note that the nonlinearity of the response that can be accurately tracked with a two-variable 4th-order calibration polynomial that also compensates the sensor's sensitivity to temperature.

chemical phenomena, as well as providing additional navigational data (acceleration and rate of turn). The future of microsystems in the world of electronics is indeed bright, and the intelligent operation of the μCluster is just the beginning of what the MEMS industry will provide in the years to come.

Table 10.6. Specifications for the Microinstrumentation Cluster

Configuration	Wristwatch/Business Card
Internal system volume	5/15 cc
Power supply	External 6-V/275-mA-hr coin cell batteries Standard 9-V batteries
Telemetry range	>100 ft
Telemetry frequency / modulation	315 MHz/ASK
Average power dissipation[a]	<500 μW
Portable operating life[a]	90 days
Measurement aperture time	10 μsec
Minimum scan interval	60 μsec
Sensor scan rate	1/minute (typical) adaptive and event-triggered

[a]Operating mode dependent

Fig. 10.25. U.S. marine wearing a μCluster during Operation Steel Knight V. Notice that the unit can be strapped to the soldier's wrist while it transmits data to a remote receiver and laptop PC.

10.8 Appendixes

Appendix 10.A COTS Micro/Nano Accelerometers

Model	FS Input (+-G)	FS Output (V)	Band-width (Hz)	Max Rating (g)	Noise (g)	PS (V)	Current (ma)	Size/Wt (in.,in.,in./gm)
				Analog Devices				
ADXL05	0.05	0.8–2.8	4000	500–1000	0.032	4.75–5.25	8	10 pin TO-100/5
ADXL50	50	0.8–2.8	4000	500–2000	0.4	4.75–5.25	10	10pin TO-100/5
ADXL151	50	3.8	1000	500–1000	0.03	4–6	1.8	0.49,0.10, 0.21/5.0
ADXL250 two axis	50	3.8	1000	500–1000	0.03	4–6	3.5	0.49,0.10, 0.21/5.0
				IC Sensors				
3255-050	50	0.5–4.5	0–2000	2000	0.25	4.5–7	10	0.53,0.30,0.16/1.5
3255-250	250	0.5–4.5	0–3000	2000	1.3	4.5–7	10	0.53,0.30,0.16/1.5
3255-500	500	0.5–4.5	0-3000	200	2.5	4.5–7	10	0.53,0.30,0.16/1.5
3140-002	2	0.5–4.5	0–250	40	0.005	8–30	5	0.9,0.6,0.21/6.5
3140-005	5	0.5–4.5	0–500	100	0.0013	8–30	5	0.9,0.6,0.21/6.5
3140-010	10	0.5–4.5	0–700	2000	0.0025	8–30	5	0.9,0.6,0.21/6.5
3140-020	20	0.5–4.5	0–1050	4000	0.005	8–30	5	0.9,0.6,0.21/6.5
3140-050	50	0.5–4.5	0–1600	10000	0.013	8–30	5	0.9,0.6,0.21/6.5
3140-100	100	0.5–4.5	0–2300	20000	0.025	8–30	5	0.9,0.6,0.21/6.5
3140-200	200	0.5–4.5	0–2500	4000	0.05	8–30	5	0.9,0.6,0.21/6.5
3022-002-p	2	-0.03–0.03	0–250	400	0.00007	5	1.5	0.9,0.6,0.21/6.5
3022-005-p	5	-0.55–0.055	0–300	400	8.00E-05	5	1.5	0.9,0.6,0.21/6.5
3022-010-p	10	-0.45–0.45	0–400	400	0.00022	5	1.5	0.9,0.6,0.21/6.5
3022-020-p	20	-0.045–0.045	0–600	400	0.00044	5	1.5	0.9,0.6,0.21/6.5
3022-050-p	50	-0.55–0.055	0–1000	1000	0.0008	5	1.5	0.9,0.6,0.21/6.5
3022-100-p	100	-0.45–0.45	0–1500	2000	0.0022	5	1.5	0.9,0.6,0.21/6.5
3022-200-p	200	-0.045–0.045	0–2000	2000	0.0044	5	1.5	0.9,0.6,0.21/6.5
3022-500-p	500	-0.55–0.055	0–2400	2000	0.008	5	1.5	0.9,0.6,0.21/6.5
3355-025 Triaxial	25	0.5–4.5	0–1000	2000	0.13	4.5–7	10	1.25,1.25,0.575/35

Appendix 10.A COTS Micro/Nano Accelerometers—*Continued*

Model	FS Input (+-G)	FS Output (V)	Band-width (Hz)	Max Rating (g)	Noise (g)	PS (V)	Current (ma)	Size/Wt (in.,in.,in./gm)
3355-050 Triaxial	50	0.5–4.5	0–2000	2000	0.025	4.5–7	10	1.25,1.25,0.575/35
3355-100 Triaxial	100	0.5–4.5	0–3000	2000	0.5	4.5–7	10	1.25,1.25,0.575/35
3355-250 Triaxial	250	0.5–4.5	0–3000	2000	1.3	4.5–7	10	1.25,1.25,0.575/35
				Kistler				
8302B2S1	2	0–5	0–300	2000	2.5E-05	12–36	40	0.7,0.625,0.2/2.8
8302B10S1	10	0–5	0–180	2000	0.00013	12–36	40	0.7,0.625,0.2/2.8
8302B20S1	20	0–5	0–160	2000	0.00025	12–36	40	0.7,0.625,0.2/2.8
				Silicon Design				
1210x-010	10	0.5–4.5	0–800	2000	0.002	4.75–5.25	2	0.35,0.35,0.105/.75
1210x-025	25	0.5–4.5	0–1000	2000	0.005	4.75–5.25	2	0.35,0.35,0.105/.75
1210x-505	50	0.5–4.5	0–1600	2000	0.01	4.75–5.25	2	0.35,0.35,0.105/.75
1210x-100	100	0.5–4.5	0–2000	2000	0.02	4.75–5.25	2	0.35,0.35,0.105/.75
2210-10	10	0.5–4.5	0–800	2000	0.002	9–30	9	1.2,1.2,1/16
2210-25	25	0.5–4.5	0–1000	2000	0.005	9–30	9	1.2,1.2,1/16
2210-50	50	0.5–4.5	0–1600	2000	0.01	9–30	9	1.2,1.2,1/16
2210-100	100	0.5–4.5	0–2000	2000	0.02	9–30	9	1.2,1.2,1/16
2412-50 Triaxial	1210	0.5–4.5	see 1210	2000	see 1210	4.75–5.25	6	1.2,1.2,1/16
				Endevco				
7290A-10	10	-2 – +2	500	5000–40	0.01	9.5–18	7.5	1,0.83,0.30/10
7290A-30	30	-2 – +2	800	5000–40	0.01	9.5–18	7.5	1,0.83,0.30/10
7290A-100	100	-2 – +2	1000	5000–40	0.01	9.5–18	7.5	1,0.83,0.30/10
7264-200	200	-0.5 – +5	0–1000	1000	6	10–15	3.6	0.4,0.3,0.2/1
7264-2000	2000	-0.5 – +.5	0–5000	1000–10000	60	10–15	3.6	0.4,0.3,0.2/1
7265A	100	-0.5 – +5	800	1000	2	10–15	1.4	0.63,0.47,0.3/6
7265A/A-HS	20	-0.5 – +5	500	1000	2	10–15	1.4	0.63,0.47,0.3/6

Appendix 10.A COTS Micro/Nano Accelerometers—*Continued*

Model	FS Input (+-G)	FS Output (V)	Band-width (Hz)	Max Rating (g)	Noise (g)	PS (V)	Current (ma)	Size/Wt (in.,in.,in./gm)
7265AM3	2000	-0.5 – +.5	0–4000	1000–5000	40	10–15	1.4	0.63,0.47,0.3/6
7267A Triaxial	1500	-0.225–0.225	1200 – 2000	1000	3	10–15	0.1	0.9,0.75,0.75/50
				Silicon Microstructures				
SM-7130-010	10	1.5–3.5	0–500	2000	0.05	9–20	6	1.4,0.8 ,0.32/6
SM-7130-050	50	1.5–3.5	0–800	2000	0.2	9–20	6	1.4,0.8 ,0.32/6
SM-7130-100	100	1.5–3.5	0–2000	2000	0.5	9–20	6	1.4,0.8 ,0.32/6
SM-7130-300	300	1.5–3.5	0–2000	2000	1.5	9–20	6	1.4,0.8 ,0.32/6
				Entran Devices				
EGA-125F-10D	10	NA	250	50	0.1	15	NA	0.27,0.14,0.14/0.5
EGA-125F-25D	25	NA	500	125	0.25	15	NA	0.27,0.14,0.14/0.5
EGA-125F-100D	100	NA	750	500	1	15	NA	0.27,0.14,0.14/0.5
EGA-125F-250D	250	NA	1000	3000	2.5	15	NA	0.27,0.14,0.14/0.5
EGA3 Triax range	NA	NA	NA	NA	NA	NA	NA	0.5,0.5,0.5/3
EGE-73B2-200F	200	NA	500	2000	4	pos/neg 10	13	0.48,0.4,0.18/1
EGE-73B2-100D	100	NA	800	2000	1	pos/neg 10	13	0.48,0.4,0.18/1

Appendix 10.B Survey of Microcontrollers/Data Loggers

Attribute	PC 104[a]	Tattle-Tale 8 Card[b]	MicroChip PIC Controller[c]	Apple Newton[d]
CPU/memory	486 DX /16 Mbytes RAM; 4Mbytes Flash ROM	M68332 & PIC16C64/ 256kbytes RAM; 256 kbytes Flash /2k EEPROM	PIC17C756/32 kB Flash, 900 B RAM	StrongARM, 160 MHz/32 Mbytes
ADC at sample rate (kHz), 12-bit accuracy	8 inputs at 100[e]	8 inputs at 100	10 inputs at[f]	8 inputs at 100[e]
No. of DACs, 10-bit accuracy	NA[g]	NA[g]	NA[g]	none
No. digital I/O lines	[h]	14	50	NA[g]
Serial Ports at bps	2 ports[e]	2 ports at 500 kbps max	2	1 port at 150
Power (W)	10 at 66 MHz	0.30 at 20 MHz	0.20 at 33 MHz	0.45 at 160 MHz
Voltage levels (V)	5,12	7–15	2.5–6.0	3.3
Low-power capability, power (mW)	Yes, 1.3	Yes, 1	Yes, 0.05	Yes, 0.15
Internal clock rate (MHz)	100	Adj to20 max	Adj to 33	160
Environmental:				
Temp. range (°C)	0–50	0–70	Commercial/industrial	0–45
Acceleration	20 G	[f]	NA[g]	NA[g]
Radiation tolerant	no	no	no	no
Weight (g)/size (mm)	[f]/96×90×23	28/5×76×13	<1/12×12×1.2	985/250×114×64
Language supported	NA[g]	ANSI C or TxBasic	C or 58 RISC Instruction Set	ANSI C
Space experience	[f]	Data Loggers used in Shuttle space suits	MightlySat	None
Website	http://www.controller.com/pc104	http://www.onsetcomp.com	http://www.microchip2.com	NA[g]
Design standard	IEEE P996.1	NA[g]	NA[g]	NA[g]

[a]S-MOS System, Inc, San Jose, CA.
[b]Onset Computers, Pocasset, MA.
[c]Microchip Technology, Chandler, AZ.
[d]Digital Electronics produces ARM(Advanced RISC Machines) chips under license.
[e]With PCMCIA card.
[f]Unknown
[g]Not applicable
[h]Available as add-on modules from PC/104 consortium members such as WinSystems, Arlington, TX; Ampro Computers, Sunnyvale, CA; Micro/Sys, Inc, Glendale, CA; and similar sources.

Appendix 10.C Survey of Microcontrollers/Data Loggers

Attribute	Honeywell Time Stamp Measurement Device (TSMD)[a]	University of Michigan Microcluster Watch	Adcon Telemetry m-T	Phillips Lab Advanced Instrumentation Controller
CPU/Memory	87C51/128 kbytes RAM; 32 kbytes EEPROM; 256 bytes Ser. RAM	MC68HC11/768 bytes RAM,24 kbytes ROM or EPROM.640 bytes EEPROM	MC68HG11/ 12kbytes ROM, 512 bytes EEPROM.2-32Ser. EEPROM	8051/128 kbytes SRAM128 kbytes nonvolatile
ADC at sample rate (kHz)/ bit accuracy	7 at [b]/8	8 at 125/8	4 at 125/8	32 inputs MUXed at 25
No of DACs,10-bit accuracy	0	0	0	8
Serial Ports at kbps	1 port at 96	1 port at19 or 31	1 port at 19 or 31	6 ports
Power(W)	0.10	0.22	0.35	0.050 at 1 MHz
Voltage levels(v)	5	6	5-7	5, 3.3
Low-power capability, power, mW	Yes, [b]	Yes. 0.075	Yes, 0.5	Yes, 0.5
Internal clock rate(MHz)	0.3 & 11	Adj to 4	5	10
Environmental				
Temp range(oC)	-55 to 125	-40 to 125	[b]	Room to 120
Acceleration	15 kGs			30 kGs
Radiation tolerant	Yes			
Size(mm)	25 × 51 × 5	14 × 14 × [b]	70 × 29 × 8	3/25 × 40 × 2
Space Experience				Designed for Mission to Mars

[a]See Ref. 73.
[b]Unknown

10.9 References

1. J. Hosticka, "CMOS Sensor Systems," *Digest, Int. Conf. on Solid-State Sensors and Actuators (Transducers '97)*, (Chicago, June 1997), pp. 991–993.
2. J. Bryzek, "MEMS: A Closer Look," *Sensors Magazine*, July 1996, 4–9.
3. H. Huijsing, F. R. Riedijk, and G. van der Horn, "Developments in Integrated Smart Sensors," *Digest, Int. Conf. on Solid-State Sensors and Actuators (Transducers '93)*, (Yokohama, Japan, June 1993), pp. 320–326.
4. K. D. Wise, "Microelectromechanical Systems: Interfacing Electronics to a Non-Electronic World," *Digest, IEEE Int. Electron Device Meeting*, pp. 11–18, December 1996.

5. L. Spangler and C. J. Kemp, "A Smart Automotive Accelerometer with On-Chip Airbag Deployment Circuits," *Digest, Solid-State Sensor and Actuator Workshop* (Hilton Head Island, SC, June 1996), pp. 211–214.
6. H. Baltes, H. Haberli, P. Malcovati, F. Maloberti, "Smart Sensor Interfaces," *Digest, IEEE Int. Symposium on Circ. and Systems* (Atlanta GA, May 1996), vol. 4, pp. 380–383.
7. E. Yoon, K. D. Wise, "An Integrated Mass Flow Sensor with On-Chip CMOS Interface Circuitry," *IEEE Trans. Electron Devices* 39 (6), 1376–1386 (1992).
8. M. Clarkson, "Smart Sensors," *Sensors Magazine*, May 1997, 14–20.
9. K. D. Wise, "Integrated Microinstrumentation Systems: Smart Peripherals for Distributed Sensing and Control, Digest, *IEEE Int. Solid-State Circ. Conf.* (San Francisco, February 1993), pp. 126–127.
10. S. Middlehoek and S. A. Audet, *Silicon Sensors* (Academic Press, London, 1989).
11. E. S. Robinson, "ASIM Application in Current and Future Space Systems," in *Microengineering Technology for Space Systems*, edited by H. Helvajian, Monograph 97-02 (The Aerospace Press, El Segundo, CA, 1997). First published as The Aerospace Corp. Report no. ATR-95(8168)-2 (1995).
12. I. Chang, "Investigation of Space Related Mission Failures," The Aerospace Corp. Report no. TOR-94(3530)-04 (1994).
13. S. Amimoto, "Multiparameter Sensor for Launch Vehicle Application," Briefings to the Air Force Space and Missile Systems Center Chief Engineering Office on MEMS Application to Space Systems, El Segundo, CA, May 1995 (unpublished).
14. "Delta Explosion Halts $1 Billion in Launches," *Aviation Week and Space Technology*, 27 Jan. 1997.
15. E. E. Kalmi, WIS 44/K11 Configuration Document, Martin Marietta, 17 May 1993, and M. Becker, Titan IV PCM Measurement List, Lockheed-Martin, Contract no. F04701-96-C-001 (30 Aug. 1996).
16. E. Moss, "Engineering Test Order for LOIS," Martin Marietta Report no. TIV-89-34 (27 March 1990).
17. G. N. Smit, "Performance Threshold for Application of MEMS Inertial Sensors in Space," *Microengineering Technology for Space Systems*, edited by H. Helvajian, Monograph 97-02 (The Aerospace Press, El Segundo, CA, 1997). First published as The Aerospace Corp. Report no. ATR-95(8168)-2 (1995).
18. E. E. Kalmi, Martin Marietta WIS 44/K11 Configuration Document (17 May 1993); M. Becker, "Titan IV PCM Measurement List," Lockheed-Martin Contract no. F04701-96-C-001. (30 Aug. 1996); and E. Moss, "Engineering Test Order for Lift-Off Instrumentation System," Martin Marietta Report no. TIV-89-34 (27 March 1990).
19. K. Feher, *Wireless Digital Communication* (Prentice Hall PTR, Upper Saddle River, NJ, 1995).
20. A. Zatsman, *et al.*, "Industry's First Integrated Wavelet Video Codec Sets New Standards for Cost, Image Quality and Flexibility," *Analog Dialog* 30-2, 7–9 (1996).
21. J. C. Lykes, "Packaging Technologies for Space-Related Microsystems and Their Elements," in *Microengineering Technology for Space Systems*, edited by H. Helvajian, Monograph 97-02 (The Aerospace Press, El Segundo, CA, 1997). First published as The Aerospace Corp. Report no. ATR-95(8168)-2 (1995).
22. P. Madan, "Infranets, Intranets, and the Internet," *Sensors* 14 (3), 46–50 (1997).
23. J. Warrier, "Smart Sensor Networks of the Future," *Sensors* 14 (3), 40–45 (1997).
24. R. Bernstein, "The LONWORKS Standard," Echelon Product Seminar, Marina del Rey, 31 March 1997. See also Wide World Web sites for Echelon <http://www.echelon.com> and for LonMark: <http://www.lonmark.org>.
25. J. Kuhnel, "Voltage Regulators for Power Management," *Analog Dialog* 30-4, 13–15 (1996).
26. M. Robyn, L. Thaller, and D. Scott, "Nanosat Power System Considerations," in *Microengineering Technology for Space Systems*, edited by H. Helvajian, Monograph 97-02 (The Aerospace Press, El Segundo, CA, 1997). First published as The Aerospace Corp. Report no. ATR-95(8168)-2 (1995).
27. B. Schweber, "Choices and Confusion Spread Wider as Spread Spectrum Goes Mainstream," *EDN* (European edition) 41 (21), 79–87 (10 October 1996).
28. K. Feher, *Wireless Digital Communication*, p. 269.

29. W.H. Hsieh, T.Y. Hsu, and Y. C.Tai, "A Micromachined Thin-film Electret Microphone," paper 2B2b02, and M. Pedersen, W. Oulthuis, and P. Bergveld, "A Polymer Condenser Microphone on Silicon with On-ship CMOS Amplifier," paper 2B2.07, *1997 Int. Conf. on Solid-State Sensors and Actuators (Transducers'97)*, (Chicago, IL, June 1997) pp. 6–19.
30. Rubic Sarian, Harris Semiconductor, Nov. 1997, private communication.
31. W. L. Pritchard and J. A. Sciulli, *Satellite Communications Systems Engineering* (Prentice Hall, Inc., Inglewood Cliffs, NJ, 1986) p. 170.
32. S. T. Amimoto, R. Crespo, E. W. Fournier, J. V. Osborn, H. Ozisik, B. H. Weiller, E. M. Yohnsee, "Development of Launch Vehicle Nanotechnology Instrumentation at The Aerospace Corporation," *Proceedings of the 44th International Instrumentation Symposium* (Reno, NV, 3–7 May 1998), pp. 240–248.
33. K. Bult, A. Burstein, D. Chang, M. Dong, M. Fielding, E. Kruglick, J. Ho, F. Lin, W. J. Kaiser, H. Marcy, R. Mukai, P. Nelson, F. Newberg, K.S.J. Pister, G. Pottie, H. Sanchez, O. M. Stafsudd, K. B. Tan, C. M. Ward, S. Xue, and J. Yao, "Low Power Systems for Wireless Microsensor," *1996 International symposium on Low Power Electronics and Design, Digest of Technical Papers*, pp. 17–21; and W. J. Kaiser, "Low Power Wireless Integrated Microsensors LWIM)," *MEMS DARPA PI Meeting* (July 1998).
34. A. Mason, N. Yazdi, K. Najafi, K. D. Wise, "A Low-Power Wireless Microinstrumentation System for Environmental Monitoring," *Digest Int. Conf. on Solid-State Sensors and Actuators* (Stockholm, June 1995), pp. 107–110.
35. K. D. Wise, "Integrated Microinstrumentation Systems: Smart Peripherals for Distributed Sensing and Control," *Digest 1993 IEEE Int. Solid-State Circuits Conf.* (San Francisco, February 1993), pp. 126–127.
36. K. D. Wise, "Microelectromechanical Systems: Interfacing Electronics to a Non-Electronic World," (invited plenary), *Technical Digest, Int. Electron Devices Meeting* (San Francisco, December 1996), pp. 11–18.
37. Samaun, K. D. Wise, and J. B. Angell, "An IC Piezoresistive Pressure Sensor for Biomedical Instrumentation," *Int. Solid-St. Circuits Conf. Digest of Tech. Papers* (February 1971), pp. 104–105.
38. S. Sugiyama, M. Takigawa, and I. Igarashi, "Integrated Piezoresistive Pressure Sensor with both Voltage and Frequency Output," *Sensors and Actuators* 4, 113–120 (September 1983).
39. W. Yun, R. T. Howe, and P. R. Gray, "Surface Micromachined Digitally Force-Balanced Accelerometer with Integrated CMOS Detection Circuitry," *Digest IEEE Solid-State Sensor and Actuator Workshop* (Hilton Head, SC., June 1992), pp. 126–129.
40. T. Ohnstein, *et. al*, "Environmentally Rugged Wide Dynamic Range Microstructure Airflow Sensor," *Digest IEEE Solid-State Sensor and Actuator Workshop*, Hilton Head, SC. (1990), 158–160.
41. J. Bernstein, *et al.*, "A Micromachined Comb-Drive Tuning Fork Rate Gyroscope," *Proc., IEEE Microelectro-mechanical Systems Workshop* (February 1993), pp. 143–148.
42. M. W. Putty and K. Najafi, "A Micromachined Vibrating Ring Gyroscope," *Digest Solid-State Sensor and Actuator Workshop* (Hilton Head, June 1994), pp. 213–220.
43. I. H. Choi and K. D. Wise, "A Silicon Thermopile-Based Infrared Sensing Array for use in Automated Manufacturing," *IEEE Trans. Electron Devices* 33, 72–79 (January 1986).
44. R. A. Wood, C. J. Han, and P. W. Kruse, "Integrated Uncooled Infrared Detector Imaging Arrays," *Digest IEEE Solid-State Sensor and Actuator Workshop* (June 1992), pp. 132–135.
45. W. C. Tang, T.-C. H. Nguyen, and R.T. Howe, "Laterally Driven Polysilicon Resonant Microstructures," *Sensors and Actuators* 20, 25–32 (1989).
46. R. S. Payne and K. A. Dinsmore, "Surface Micromachined Accelerometer: A Technology Update," Digest SAE Meeting, Detroit, pp. 127–135, February 1991.
47. C. T.-C. Nguyen and R. T. Howe, "Quality-Factor Control for Micromechanical Resonators," *Digest Int. Electron Devices Meeting* (December 1992) pp. 505–508.

48. J. B. Sampsell, "The Digital Micromirror Device and Its Application to Projection Display," *Digest of Technical Papers 7th Int. Conf. on Solid-State Sensors and Actuators (Transducers '93)*, (IEE of Japan, 1993), pp. 24–27.
49. P. L. Bergstrom, J. Ji, Y. Liu, M. Kaviany, and K. D. Wise, "Thermally Driven Phase-Change Actuation," *IEEE J. of Microelectromechanical Systems* 4, 10–17 (March 1995).
50. H. Seidel, L. Csepregi, A. Heuberger, and H. Baumgartel, "Anisotropic Etching of Crystalline Silicon in Alkaline Solutions," *J. Electrochem. Soc.*, 3612–3632 (November 1990).
51. W.-H. Juan and S. W. Pang, "Released Si Microstructures Fabricated by Deep Etching and Shallow Diffusion," *IEEE J. Microelectromechanical Systems* 5, 18–23 (March 1996).
52. M. A. Schmidt, "Silicon Wafer Bonding for Micromechanical Devices," *Digest Solid-State Sensor and Actuator Workshop* (Hilton Head, SC., June 1994), pp.127–131.
53. H. Guckel, T.R. Christenson, K.J. Skrobis, J. Klein, and M. Karnowsky, "Design and Testing of Planar Magnetic Micromotors Fabricated by Deep X-Ray Lithography and Electroplating," *Digest 7th Int. Conf. on Solid-State Sensors and Actuators (Transducers '93)*, (IEE of Japan, 1993), pp. 76–79.
54. R. T. Howe, "Surface Micromachining for Microsensors and Microactuators," *J. of Vacuum Science and Technology, B* 6, 1809–1813 (December 1988).
55. H. Guckel and D. W. Burns, "Planar Processed Polysilicon Sealed Cavities for Pressure Transducer Arrays," *Digest IEEE IEDM* (December 1984).
56. C. H. Mastrangelo and C. H. Hsu, "Mechanical Stability and Adhesion of Microstructures under Capillary Forces," *IEEE J. Microelectromechanical Systems* 2, 33–62 (March 1993).
57. Y. Zhang and K. D. Wise, "A High-Accuracy Multi-Element Silicon Barometric Pressure Sensor," *Digest Int. Conf. on Solid-State Sensors and Actuators* (Stockholm, June 1995), pp. 608–611.
58. S. T. Cho and K. D. Wise, "A High-Performance Microflowmeter with Built-In Self-Test," *Sensors and Actuators, A*, 47–56 (March 1993).
59. N. Yazdi, A. Mason, K. Najafi, K. Wise, "A Low-Power Generic Interface Circuit for Capacitive Sensors", *Digest, Solid-State Sensor and Actuator Workshop* (Hilton Head Island, SC, June 1996), pp. 215–218.
60. A. Chavan and K. D. Wise, "A Batch-Processed Vacuum-Sealed Capacitive Pressure Sensor," *Digest IEEE Int. Conf. on Solid-State Sensors and Actuators* (Chicago, June 1997), pp. 1449–1452.
61. W. G. Baer, K. Naafi, K. D. Wise, and R. S. Toth, "A 32-Element Micromachined Thermal Imager with On-Chip Multiplexing," *Sensors and Actuators, A: Physical* 48 (1), 47–54 (1 May 1995).
62. C. C. Liu and C. H. Mastrangelo, "An Ultrasensitive Uncooled Heat-Balancing Infrared Detector," *Digest IEEE Int. Electron Devices Meeting* (December 1996), pp. 549–552.
63. N. Najafi, K. D. Wise, and J. W. Schwank, "A Micromachined Ultra-Thin-Film Gas Detector," *IEEE Trans. Electron Devices* 41,1770–1777 (October 1994).
64. S. Majoo, J. W. Schwank, J. L. Gland, and K. D. Wise, "A Selected-Area CVD Method for Deposition of Sensing Films on a Monolithic Integrated Gas Detectors," *IEEE Electron Device Letters*, pp. 217–219 (May 1994).
65. A. Mason, N. Yazdi, A. V. Chavan, K. Najafi, K. D. Wise, "A Generic Multielement Microsystem for Portable Wireless Applications," *Proceedings of the IEEE* 6 (8), 1733–1746 (August 1998).
66. S. C. Terry, J. H. Jerman, and J. B. Angell, "A Gas Chromatographic Air Analyzer Fabricated on a Silicon Wafer," *IEEE Trans. Electron Devices* 26,1880–1886 (December 1979).
67. N. Najafi, K. D. Wise, "An Organization and Interface for Sensor-Driven Semiconductor Process Control Systems," *IEEE Tran. on Semiconductor Manufacturing* 3 (4), 230–238 (November 1990).
68. Y. E. Park and K. D. Wise, "An MOS Switched-Capacitor Readout Amplifier for Capacitive Pressure Sensors," Proc., *IEEE Custom Integrated Circuits Conf.* (May 1983), pp. 380–384.
69. A. Selvakumar, N. Yazdi, K. Najafi, "A Low Power, Wide Range Threshold Acceleration Sensing System," Proc., *IEEE Microelectromechanical Systems Workshop* (San Diego, CA, February 1996), pp. 186–191.
70. J. L. Lund and K. D. Wise, "Chip-Level Encapsulation of Implantable CMOS Microelectrode Arrays," *Digest Solid-State Sensor and Actuator Workshop* (Hilton Head, June 1994 SC), pp. 29–32.

71. H. Guckel and D. W. Burns, "Planar Processed Polysilicon Sealed Cavities for Pressure Transducer Arrays," *Digest IEEE IEDM* (December 1984).
72. L. Spangler and C. Kemp, "ISAAC: Integrated Silicon Automotive Accelerometer," *Digest Int. Conf. on Solid-State Sensors and Actuators* (Stockholm, June 1995), pp. 585–588.
73. S. B. Crary, W. G. Baer, J.C. Cowles, and K. D. Wise, "Digital Compensation of High-performance Silicon Pressure Transducers," *Sensors and Actuators A* 21–23, 70–72 (1990).
74. S. B. Crary, L. Hoo, and M. Tennenhouse, "I-Optimality Algorithm and Implementation," *Proc., Symposium on Computational Statistics*, Vol. 2 (Neuchatel, Switzerland, August 1992), pp. 209–214.
75. G. Harvey, S. Louis, and S. Buska, "Micro-time stress measurement device development," Rome Laboratory technical report no. (ERSR), RL-TR-94-196 (November, 1994).

11

Chemical Microsensors for Gas Detection and Applications to Space Systems

B. H. Weiller*

11.1 Introduction

Chemical microsensors are small devices consisting of a physical transducer and a thin-film coating that selectively react with the chemical of interest. The transducer detects a change in a physical property of the coating and gives a signal that is correlated with the concentration of the target chemical. Typical physical properties measured are electrical resistance, mass, reflectivity, absorptivity, capacitance, and threshold voltage of semiconductor devices. The term "micro" denotes that the device features are in the micron size range. However, the overall size of the device can be in the millimeter to centimeter range.

The advantages of chemical microsensors for chemical detection are numerous when compared with the more traditional methods of chemical detection, which rely on large, complicated instrumentation best suited for the laboratory. The foremost benefits of microsensors are reduced cost, size, weight, and power consumption, and increased functionality. These benefits open up many new applications in support of space activities that would not be possible with traditional methods. Chemical microsensors are typically fabricated using the batch wafer-processing techniques developed in the silicon microelectronics industry. With this manufacturing approach, it is possible to make hundreds of small sensors on a single silicon wafer, resulting in improved uniformity and reduced cost. In addition, it is possible to integrate the sensors with analog and digital circuitry at the wafer level. Other integration approaches such as multichip modules allow integration with microprocessors, microcontrollers, and wireless transceiver components. This produces powerful, smart sensors that are very small and can measure chemical changes, control equipment, and transmit data from remote locations. These sensors do not require expensive, highly educated technicians for operation, unlike much laboratory chemical instrumentation. Furthermore, they can be used in hazardous locations and can be distributed around large areas to map out chemical concentrations as a function of time. This makes them ideal for plume-tracking applications on the ground or in remotely piloted vehicles (RPV).

Some of the devices or transducers that will be discussed include chemiresistors, metal-oxide semiconductor (MOS) devices, acoustic-wave devices, and fiber-optic sensors. Chemiresistors measure the change in resistance of a thin-film coating upon adsorption of a gas. Useful coatings range from metals, metal alloys, polymers, organic semiconductors, and conducting oxides. Metal-oxide semiconductor devices can become extremely sensitive chemical microsensors when a catalytically active metal such as palladium is used as the gate metal. This induces a chemical effect on device performance, and parameters such as threshold voltage of a transistor can be used to measure chemical concentration such as hydrogen. Acoustic-wave devices such as quartz crystal microbalances (QCM) and surface acoustic-wave (SAW) devices are very sensitive mass detectors that are also useful for trace detection. The frequencies of acoustic waves in piezoelectric crystals are highly sensitive to mass loading at the surface. The QCM relies on bulk acoustic

*Mechanics and Materials Technology Center, The Aerospace Corporation, El Segundo, California.

waves, while the SAW device uses higher frequency surface acoustic waves and is significantly more sensitive. Fiber-optic devices that can be used to detect chemically induced optical changes in thin-film coatings will also be discussed. Given the limited scope of this chapter, many devices cannot be discussed, including the use of optical devices for spectroscopic measurements of molecules in gases or liquids, chemiluminescent, electrochemical, tin oxide, and biochemical sensors.

11.1.1 Space-Related Applications

The first space-related application where chemical microsensors are useful occurs on the ground. Launch vehicles use or produce massive quantities of chemicals that are either highly toxic or explosive. Some examples are the hypergol propellants: hydrazine (N_2H_4) and its derivatives, unsymmetrical dimethyl hydrazine [UDMH, $(CH_3)_2N_2H_2$] and monomethyl hydrazine (MMH, $CH_3N_2H_3$) and nitrogen tetraoxide (NTO, N_2O_4) or nitrogen dioxide (NO_2),* all of which are quite toxic. Table 11.1 shows the relevant concentration limits for these species. The Threshold Limit Value (TLV)[1] is the limit workers can be exposed to during an 8-h shift and is set by the Occupational Safety and Health Administration (OSHA). The Short-Term Public Emergency Guidance Limit (SPEGL)[2] is the exposure limit for the general population during a short-term emergency event such as a launch anomaly. The TLVs for the hydrazines are very low (10 ppb), and they present quite a challenge for chemical detection strategies. Hydrogen gas is another propellant that is explosive at mixtures in air above 4%.[3] For the cryogens used in fueling, such as liquid hydrogen and nitrogen, displacement of oxygen in enclosed spaces is a concern around the launch pad. Therefore, detection of oxygen is important in areas where humans have access. Solid rocket motors produce large quantities of toxic exhaust in the form of hydrogen chloride (HCl), which is produced from the propellant (ammonium perchlorate) used in these motors. A typical Titan IV or Space Shuttle produces approximately 100 tons of HCl in the troposphere,[4] which is released into the environment in a plume that could reach population centers. Even for nominal launches, significant human resources have been used to detect and track plumes of exhaust gases to determine if population centers will be affected. In the event of fuel spills or aborted launches, it is critical that plumes of the more toxic propellants and oxidizers be tracked accurately.

There are significant problems with the instrumentation that is currently deployed in the field for chemical plume detection. Often it is not designed for field use, requires extensive technical training, does not provide real-time response, and is relatively insensitive and expensive.

Table 11.1. Concentration Limits for Toxic Gases Associated with Launch Vehicles

Species	TLV[a] (ppm)	SPEGL[b] (ppm)
Hydrazine	0.010	2.0
UDMH	0.010	24.0
MMH	0.010	0.24
HCl	5	1.0
NO_2	3	1.0

[a]Ref. 1.
[b]Ref. 2.

*NO_2 is in equilibrium with N_2O_4: $2NO_2 = N_2O_4$.

Improved technology for real-time, remote, *in-situ* trace detection of chemical plumes would address concerns about potential population exposure. Not only would the general population be better protected as a result, but the ability to meet current and future launch schedules would be greatly improved, and substantial cost savings from delay avoidance would be realized. Chemical microsensors provide an ideal way to do this: one can envision an array of sensors permanently arranged around a launch pad that would respond to gases in a short time frame and transmit their data through wireless communication links back to a base station in real time.

Chemical microsensors are also useful for chemical detection in space. For example, they are ideal for autonomous monitoring of chemical concentrations inside spacecraft environments such as the Space Shuttle or the International Space Station (ISS). Life-support equipment requires the concentrations of O_2 and CO_2, and relative humidity to be continuously monitored. Current missions do have equipment that performs these functions, but chemical microsensors would be a significant improvement in cost, weight, size, power consumption, and functionality. As a consequence, more sensors could be used and could provide increased coverage of large structures such as the ISS. In addition, many other potentially dangerous gases in spacecraft cabins are not currently monitored. Table 11.2 shows the Spacecraft Maximum Allowable Concentrations for Selected Airborne Contaminants (SMAC limits).[5] These limits are recommended by the National Academy of Sciences for certain chemicals for human environments in space. It should be noted that the SMAC list does not include all chemicals of concern, only the ones for which limits have been established. The limits are lower for longer exposure, and the detection of these trace species becomes much more important for long missions such as those planned for the ISS. Equipping the space station with instrumentation to detect even a fraction of these chemicals would be very expensive and difficult. Chemical microsensors would be ideal for this task and are currently being developed for this purpose.[6]

Contamination of spacecraft components is another area where chemical microsensors can play an important role. Contamination of sensitive optical and electronic components occurs on the ground or in space and can severely impair the functionality of spacecraft. Satellite materials outgas during thermal cycling in the vacuum of space, and material redeposits on other parts of the satellite. Optical components are most often affected and compromised by this problem. *In-situ* detection of contamination fluxes provides important information about the occurrence of the problem and potentially its source. Acoustic-wave mass sensors are most useful here, and QCM sensors are routinely flown on spacecraft for this purpose.[7] Chemical microsensors such as SAW devices would provide more sensitive detection ($\sim 10^3$ more sensitive) and could provide chemical identification as well with the appropriate coating. Monitoring clean-room environments for chemical and other contamination of exposed satellite components is another important role for chemical microsensors, and SAW devices are proving to be quite useful for this application.

Another application of chemical microsensors in space is planetary exploration. A major function of planetary probes is the chemical analysis of the soil and the atmosphere to determine the potential for life support. In the past, infrared (IR) spectroscopy and mass spectrometry have been used successfully for this purpose, providing a wealth of information; however, these instruments are expensive, complicated, and heavy. Chemical microsensors can and will play an important role in the push for "better, cheaper, and faster" ways to perform this task. One probe, the Russian Mars Lander, did contain a fiber-optic chemical microsensor suite called Mars Oxidant Experiment (MOx) for this purpose. Unfortunately, that spacecraft was destroyed on launch, but the instrumentation developed is described later in this chapter. For future missions, one could imagine proliferating the surface of a planet with small, wireless chemical microsensors that would provide local chemical analysis of soil and gases with data relayed back to the host spacecraft.

Table 11.2. Spacecraft Maximum Allowable Concentrations for Selected Airborne Contaminants

Chemical Formula		SMAC Limits (ppm)[a]				
		1 h	24 h	7 days	30 days	180 days
Acetaldehyde	CH_3CHO	6	2	2	2	2
Acrolein	CH_2CHCHO	75	35	15	15	15
Ammonia	NH_3	30	20	10	10	10
Benzene	C_6H_6	10	3	0.5	0.1	0.07
Carbon dioxide	CO_2	13,000	13,000	7000	7000	7000
Carbon monoxide	CO	55	20	10	10	10
2-ethoxyethanol	$EtOC_2H_4OH$	10	10	0.8	0.5	0.07
Formaldehyde	HCHO	0.4	0.1	0.04	0.04	0.04
Freon 113	CCl_2FCClF_2	50	50	50	50	50
Hydrogen	H_2	4100	4100	4100	4100	4100
Hydrazine	N_2H_4	4	0.3	0.04	0.02	0.004
Indole	C_8H_7N	1	0.3	0.05	0.05	0.05
Mercury	Hg	0.01	0.002	0.001	0.001	0.001
Methane	CH_4	5300	5300	5300	5300	5300
Methanol	CH_3OH	30	10	7	7	7
Methyl ethyl ketone	MeCOEt	π50	50	10	10	10
Methylene chloride	CH_2Cl_2	100	35	15	5	3
Nitromethane	CH_3NO_2	25	15	7	7	5
Octamethytrisiloxane	$C_8H_{24}O_2Si_3$	400	200	100	20	4
2-propanol	$CH_3CHOH(CH_3)$	400	100	60	60	60
Toluene	C_7H_8	16	16	16	16	16
Trimethansilanol	$(CH_3)_3SiOH$	150	20	10	10	10
Vinyl chloride	CH_2CHCl	130	30	1	1	1

[a]Ref. 5

11.1.2 General Issues and Parameters Affecting Sensor Performance

Prior to embarking on a detailed discussion of chemical-microsensor technologies, some general comments on the important issues and parameters affecting sensor performance are appropriate. This section was adapted from Ballantine *et al.*[8]

11.1.2.1 Selectivity

One of the most important parameters is selectivity for the molecule of interest. This is the ability of a sensor to discriminate the molecule of interest (the analyte) in a mixture of other chemicals. For almost any chemical-detection application, the molecule of interest is a minor constituent in a complex chemical mixture. Even for a simple application requiring detection of a single component in air, there are potential interferences from O_2, CO_2, H_2O, and many other trace gases.

There are three general approaches to the issue of selectivity with microsensors. The ideal device is completely selective for the chemical of interest with no interferences from unwanted chemicals. This is virtually impossible to achieve, but given certain detector/analyte combinations with limited environmental exposure, one can come close. A good example is H_2 detection using Pd alloy films.

The second approach is to use an array of sensors, each relatively unselective but with somewhat different responsivity to a range of compounds. The data analysis is coupled with pattern-recognition techniques to deconvolute chemical information from large data sets. Often this requires advanced computational techniques, including neural networks and appropriate algorithms that can be "trained" to deconvolute the data. The analysis of mixtures of organic compounds using acoustic-wave sensors with polymer coatings is one application where this approach works well.

The third approach is to use a completely unselective detector at the end of a device that provides temporal or spatial separation of individual chemical components. A prime example of this is gas chromatography, where chemicals are separated by flowing a gas mixture over a packed column of adsorbent that causes temporal separation of the components by interaction with the absorbent. Another example is mass spectrometry, where ions are separated spatially or temporally by electric or magnetic fields. In these cases, one wants very nonspecific detectors in order to detect as many compounds as possible. A problem with this approach is that the separation element adds significant size and complexity and results in what is really an analytical instrument and not a chemical sensor. Therefore, this approach is outside the scope of this work.

11.1.2.2 Reversibility

Reversibility is an important characteristic of chemical microsensors. This refers to the reversibility of a sensor response after exposure to the analyte. An important distinction between reversible sensors and irreversible dosimeters needs to be made. Coatings that bind a molecule so strongly that no desorption is observed at room temperature can be very selective but are irreversible. However, the device then integrates over the total exposure (dose), and the result is a dosimeter, not a sensor.

There is often a trade-off between selectivity and reversibility. In general, the reason is that selectivity requires strong adsorption of the analyte; whereas reversibility generally implies weaker binding. The ideal sensor is one that is selective and reversible. Strong absorption of the target molecule is required for selectivity, but for continuous detection of changing concentrations, reversible adsorption of the analyte to the coating is required. This can allow an interferent (not the molecule of interest) to displace the analyte. It is rare to find a reversible chemical sensor with high selectivity.

Reversible sensors are generally preferred to dosimeters, since real-time information can be acquired instead of average values typically obtained with dosimeters. For example, an 80 ppb-h dose could have occurred over 1 h or an 8 h period. Furthermore, information about the peak concentration is lost in a dosimeter. Dosimeters also have a limited lifetime, determined by their dynamic range; once the exposure limit is reached, the dosimeter either must be regenerated or replaced. Reversible sensors, in principle, have unlimited lifetimes and, unlike dosimeters, can be individually calibrated. If dose measurements are required, a dosimeter can be made from a reversible sensor by combining the sensor with a microprocessor for data logging and averaging with the same result as a chemical average.

Dosimeters are useful for short-term applications such as monitoring the exposure of personnel to toxic chemicals as long as the device does not become saturated. Often chemical exposure limits are defined in terms of a dose. For example, the TLV for hydrazine is 10 ppb over an 8-h period, or 80 ppb-h. Dosimeters are typically more selective than sensors, and this is an advantage. Because a dosimeter has a "memory" of its exposure, the reading of dosimeters can be done at the experimenter's convenience, even at times much later than a chemical event. Finally, if the time response of a dosimeter is fast, temporal information can be obtained by fast data acquisition with an integrated microprocessor.

11.1.2.3 Sensitivity

Sensitivity is a critical parameter in evaluating sensor performance for a particular application. Sensitivity is defined here as the minimum detectable quantity of a chemical and is also known as the limit of detection (LOD). Sensitivity is defined differently for reversible sensors and dosimeters. For reversible sensors, sensitivity is defined as a concentration value, typically given in a mole ratio such as ppm or ppb or as a mass concentration such as mg/m^3. For irreversible dosimeters, the sensitivity is given as a time-integrated concentration or a dose such as ppb-h. Sensitivity is determined by the fundamental sensitivity of the transducer but is also determined by the coating. A good example is from SAW devices, where the chemical sensitivity is determined by the fundamental mass sensitivity and by the fraction of analyte that can be bound to the sensor surface through chemical interactions. The LOD is determined by the noise level in the system and is typically defined as a signal-to-noise ratio of 3. The sensitivity is calculated from the noise level divided by the responsivity (see below). The required sensitivity for a particular application is usually determined by a set regulatory value, and one must be careful to consider that some of these limits are doses and others are concentration values.

Sensitivity can be increased by the use of a sample concentrator. This is simply a tube containing a porous absorbent that is used to accumulate sample that is then thermally desorbed. Note that this does not change the fundamental sensitivity of the sensor and will compromise time response. The use of a sample concentrator actually belongs in the instrument category and will not be described in detail here. However, it has found considerable utility with SAW sensor systems.

11.1.2.4 Responsivity

Responsivity is defined here by the slope of a sensor response versus concentration and has units of sensor response per concentration unit. For example, the responsivity of acoustic-wave sensors is given in terms of change in frequency/change in concentration ($\Delta f/\Delta c$) or Hz/ppm. This can be defined for reversible, linear-response sensors; however, for sensors where the response is nonlinear, the responsivity is harder to define. It is possible to define the responsivity in terms of logarithmic response.

It should be noted that both responsivity and sensitivity can depend on the temperature, and the effect can be in opposite directions depending on the type of coating applied. For example,

absorption coatings such as the polymers used for SAW devices typically show an inverse temperature dependence of both responsivity and sensitivity. Conversely, chemisorption coatings, in which the reaction rate is key to sensor response, typically show a positive temperature dependence.

11.1.2.5 Dynamic Range

Dynamic range refers to the range of response of a sensor from the LOD to the saturation limit, be it concentration or dose. This is a parameter of concern for applications where high concentrations of analyte may be encountered, such as with hydrogen gas. There are applications such as contamination or leak detection where one would like to detect very low concentrations of hydrogen and others such as explosion hazard where one would like to detect high concentrations of hydrogen. The lower end is determined by the sensitivity, as discussed above. The upper end is usually given by a saturation limit, which is often determined by the capacity of the coating to take up analyte. While it may be possible to increase the thickness of the coating to increase its capacity, there are limits and trade-offs with sensitivity. The saturation limit can also be determined by the electronics associated with the sensor. The dynamic range is typically presented as the ratio of the saturation limit to the LOD. As noted above, the useful lifetime of dosimeters is limited by the dynamic range, but in some cases it may be possible to regenerate the dosimeter chemically or thermally.

11.1.2.6 Response Time

Response time is an important parameter of a chemical sensor that needs to be considered. The response time for a chemical sensor is typically defined as the time required for a sensor to reach 90% of its final value. This differs from other types of sensors, which are often characterized by (1/e) times for exponential response functions. This is a somewhat arbitrary definition, which is probably related to the fact that the time dependence of chemical sensors is typically not exponential. There can be different rise and fall response times, and both should be defined for a sensor. Response times can be determined by surface reaction kinetics or by the time to reach equilibrium, which is determined by kinetics or diffusion. The rate of surface reaction can be affected by the surface area of the coating and the diffusion through the coating, which becomes more rapid for thin porous films. Whether diffusion or kinetics is the rate-limiting process, higher temperatures will accelerate both processes and therefore increase the response time. When response times are measured, it is important to ensure that diffusion of the gas to the sensor is not rate limiting and that the true response time of the sensor is measured.

11.1.3 Thermodynamics and Kinetics of Molecules at Sensor Surfaces

Chemical sensor behavior is determined by the chemical reactions of molecules at the sensor surfaces. Therefore, in order to better understand how chemical sensors operate and what parameters affect their response, some elementary thermodynamics and kinetics of surface reactions are worth reviewing. Figure 11.1 shows a reaction energy profile for a reaction of a gas-phase molecule with a surface. The y axis is energy, and the x axis is a reaction coordinate that represents the extent of reaction. Reactants lie at the left-hand side of the plot, and products lie at the right-hand side. The activation energy for adsorption is E_a, and the activation energy for desorption is given by E_d. In this example, the products are at lower energy than the reactants, and the reaction is exothermic, $\Delta H_a < 0$, as is typical of most surface adsorption reactions. The enthalpy of adsorption is given by $\Delta H_a = E_a - E_d$. The size of E_a determines the type of surface reaction. When E_a is small or negligible, the reaction is simple physisorption and no significant energy is required to cause reaction. Furthermore, the energy required to desorb molecules is simply the activation

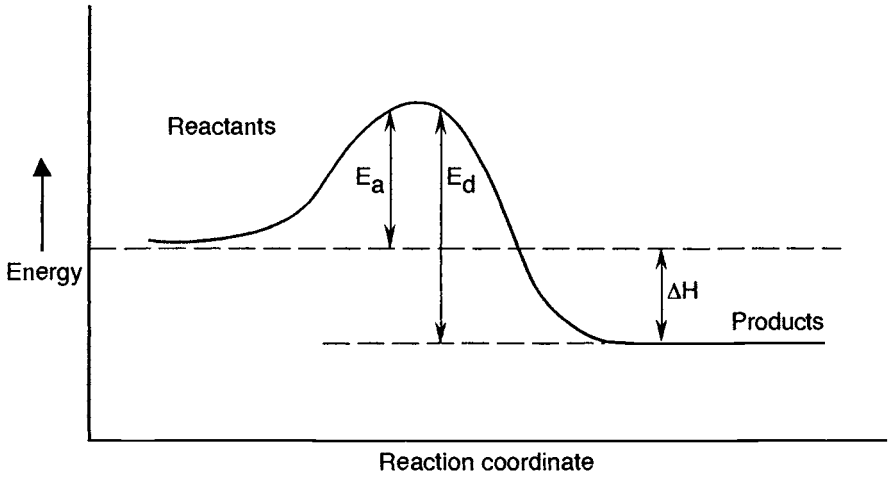


Fig. 11.1. Energy profile for a surface reaction.

energy for desorption, $\Delta H_d \sim E_d$. When E_a is significant, the reaction requires energy to occur, and the sensor will show an enhanced sensitivity with higher temperatures.

The kinetics of molecular adsorption to form a monolayer on a surface are described by following simple kinetic expressions, given some assumptions. Reaction at a surface is assumed to occur only at a limited number of surface sites, and reaction probability decreases as the reaction progresses since these sites become occupied with reactant. The fraction of occupied surface sites is called the coverage, θ. Therefore, the rate of surface adsorption depends on the partial pressure of gaseous analyte (p) and the fraction of unoccupied surface sites ($1 - \theta$):

$$r_a = k_a p(1-\theta) = p(1-\theta)A_a \exp(-E_a/RT) \tag{11.1}$$

where k_a is the rate constant for surface reaction, R is the gas constant, and A_a is the frequency factor for adsorption. When E_a is significant, the rate of surface adsorption is strongly temperature dependent, as given by the exponential term. The rate of desorption from a surface is simply proportional to the fraction of occupied surface sites and is given by:

$$r_d = k_d\theta = \theta A_d \exp(-E_d/RT), \tag{11.2}$$

where k_d is the rate constant for surface desorption and A_d is the frequency factor for desorption. A chemical reaction is at equilibrium when the forward and reverse reaction rates are equal. Setting $r_a = r_d$, an expression for the equilibrium constant for surface adsorption K_a is obtained:

$$K_a = \frac{k_a}{k_a} = \frac{\theta}{p(1-\theta)} = \frac{A_a}{A_d}\exp\left[-\frac{(E_a - E_d)}{RT}\right] \tag{11.3}$$

This equation can be rearranged to give:

$$\theta = \left[\frac{pK_a}{1+pK_a}\right] \tag{11.4}$$

which is called the Langmuir adsorption isotherm, a plot of which is shown in Fig. 11.2. As the partial pressure of reactant is increased, the surface coverage asymptotically approaches complete

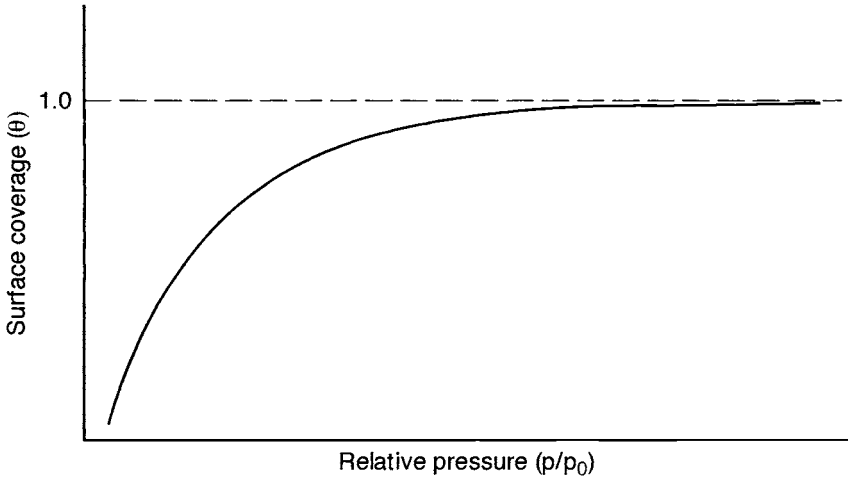


Fig. 11.2. Langmuir adsorption isotherm.

coverage ($\theta = 1$). For chemical sensors where surface adsorption to form a monolayer is important, the response is related to the surface coverage and therefore should be proportional to θ. It should be noted that Langmuir adsorption does not account for multilayer adsorption. Other adsorption isotherms such as Brunauer-Emmett-Teller (BET) have been derived for this situation.[9]

The equilibrium constant K_a is related to the enthalpy (ΔH_a) and entropy (ΔS_a) of adsorption as follows:

$$\Delta G_a = -RTln(K_a) = \Delta H_a - T\Delta S_a \tag{11.5}$$

$$K_a = \exp\left(\frac{-\Delta H_a}{RT}\right)\exp\left(\frac{\Delta S_a}{R}\right) \tag{11.6}$$

where ΔG_a is the free energy of adsorption. K_a is strongly dependent on the temperature determined by ΔH_a. By comparing Eqs. (11.3) and (11.6), we can see the relationship between kinetic and equilibrium parameters:

$$\Delta H_a = E_a - E_d, \tag{11.7}$$

$$\frac{A_a}{A_d} = \exp\left(\frac{\Delta S_a}{R}\right) \tag{11.8}$$

In order for a reaction to proceed spontaneously, the free-energy change must be negative, $\Delta G_a < 0$. Most surface reactions are exothermic ($\Delta H_a < 0$), and ΔS_a is also negative; a molecule in the gas phase is more random that one adsorbed on a surface. Therefore, in order for $\Delta G_a < 0$, $T\Delta S_a$ should be less negative than ΔH_a.

The Langmuir adsorption isotherm provides an instructional but simplified example of the physical principles operating at sensor surfaces. However, monolayer formation is usually the first step in a series of reactions that may include diffusion into a thick film or chemisorption. One typical scenario is surface adsorption followed by chemical reaction, as shown in Fig. 11.3, and is called the Langmuir-Hinschelwood model. This model applies to the chemisorption coatings

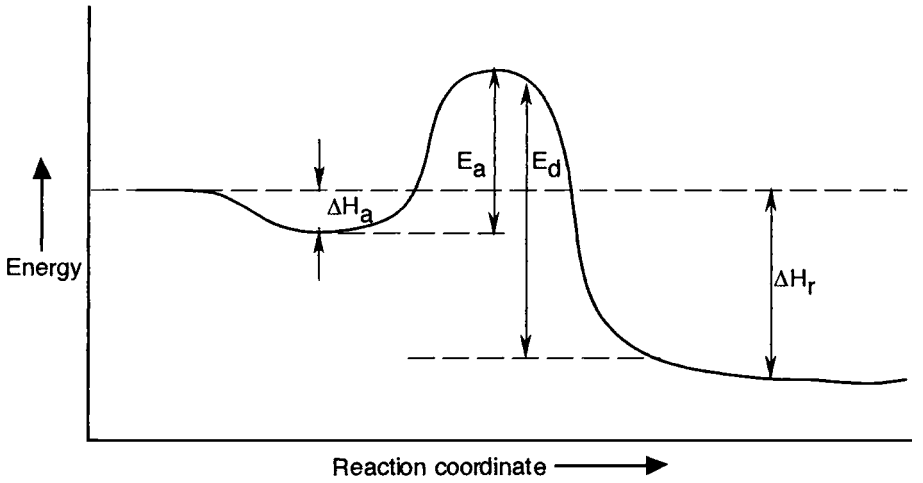


Fig. 11.3. Energy profile for a chemisorption surface reaction.

described below. In this case, it is assumed that the absorption is at equilibrium and follows a Langmuir adsorption isotherm and that the subsequent surface reaction is the rate-limiting step.

$$G \overset{K_a}{=} S \overset{k_s}{\rightarrow} P \tag{11.9}$$

Here, G is the gas-phase analyte, S is the surface-adsorbed analyte, and P is the product from surface reaction. The rate of surface reaction is given by:

$$r_s = \frac{k_s K_a p}{(1 + K_a p)} \tag{11.10}$$

where k_s is the surface-reaction rate constant. An important feature of this scenario is that both k_s and K_a are temperature dependent in opposite directions. The surface-reaction rate constant k_s increases with temperature, but the adsorption constant K_a decreases with temperature. With some approximations, it is easier to see the mathematical effect of temperature. In the limit of very small surface reaction where $K_a p << 1$ or $\theta << 1$, Eq. (11.10) reduces to:

$$r_s = k_s K_a p = p A_s \exp\left(\frac{\Delta S_a}{R}\right) \exp\left[-\frac{(\Delta H_a + E_a)}{RT}\right] \tag{11.11}$$

where A_s is the frequency factor for the surface reaction. Thus the temperature dependence is determined by the difference in ΔH_a and E_a since they are of opposite signs ($\Delta H_a < 0$). One perspective of this result is that the energy released by surface adsorption drives the surface reaction. For example, this equation predicts that when $\Delta H_a = -E_a$, the rate of reaction will have no temperature dependence.

Until now, reaction with only the surface layer has been considered. For many coatings used in chemical sensors, diffusion into the bulk of the film is important, and it is often the rate-limiting process that determines the time response of the sensor.

11.2 Chemical Microsensor Technologies

11.2.1 Chemiresistors

The chemiresistor is one of the most common chemical microsensor devices. As the name implies, the operating principle is a change in the resistance (ΔR) of a thin-film material. Therefore, the goal in designing a chemiresistor is to optimize ΔR of a material in response to a particular gas. The materials found to be useful here include metals (Pd and Ag), organic semiconductors (conducting polymers, phthalocyanines), and inorganic semiconductors such as metal oxides (e.g., WO_3, SnO_2) (see Table 11.3), which vary widely in their baseline resistance from 10 to $10^9\ \Omega$. Therefore, the configuration of the devices and the measurement techniques are considerably different for metals and semiconductors.

Metal oxides, especially SnO_2, are widely used materials for many chemical sensors, including most of those on the market today. This is a result of the range of chemicals that can be detected with one material and various dopants. However, most of the gases detected using these materials are CO and hydrocarbons, and they have not found as much use for space applications for several reasons. They are not the best sensors for detecting any of the propellants, oxidizers, or exhaust gases found in rockets (hydrazine, MMH, UDMH, NO_2, H_2, HCl). Furthermore, these sensors function by oxidation of the measured gas with concomitant reduction of the oxide sensor. A background of ambient O_2 is required in order to maintain a constant stoichiometry of the oxide. This occurs via a catalytic cycle whereby the sensor is reduced by the analyte and oxidized by ambient O_2 to regenerate the sensor. Not only is O_2 required, but the sensor response is sensitive to changes in the O_2 partial pressure, which is not constant in cabin environments such as the Space Shuttle. The sensors operate at fairly high temperatures (~500 °C) and therefore require significant power for operation. Finally, the sensor materials are vibration- and shock-sensitive ceramics and may not withstand the launch environment.* For these reasons, their use in space applications is limited, and their principles of operation will not be discussed. Interested readers are referred to other sources.[18]

Table 11.3. Materials Used as Chemiresistors

Material	Molecules	Source
Palladium	H_2	Ref. 10
Silver	O_3, O atoms	Ref. 11
Phthalocyanines	NO_2, O_3, F_2, BC_{13}, HCl	Ref. 12
Polyaniline	NH_3	Ref. 13
Polythiophene	Hydrazines	Ref. 14
Polypyrrole	NH_3, hydrazines, methanol	Refs. 15, 16
Polymers/carbon	Various organics	Ref. 17
WOx:M	H_2S	a
SnOx:M	Various inorganics	b

[a]Armstrong Monitoring Corporation, Ontario, Canada.
[b]Figaro Corporation, Japan.

*The use of a micromachined substrate may overcome these limitations and is described at the end of this chapter.

11.2.1.1 Metals

11.2.1.1.1 Hydrogen Chemiresistor

A metallic chemiresistor originally developed at Sandia National Labs for the detection of hydrogen gas is one of the most relevant for space-related applications.[10] It has been known for many years that the resistance of palladium foils increases upon exposure to hydrogen gas. This is due to the well-known rapid formation of a hydride in palladium metal. A problem with using pure Pd metal is that above a certain H_2 pressure, there is a phase change in Pd and a corresponding lattice expansion. Not only is this phase transformation irreversible, but delamination of films can occur. In Pd-metal alloys such as with silver and nickel, the solubility of hydrogen is substantially reduced, which allows exposure to much higher H_2 concentrations and avoids this problem. The Sandia group showed that thin films of Pd/Ni alloy (~8%) are highly effective H_2 sensors with a very reproducible and reversible change in resistivity over the range from 100 ppm to 100% H_2. The films can be deposited by sputtering or e-beam evaporation on an oxidized silicon wafer compatible with CMOS (complementary metal oxide semiconductor) processing. Therefore it is possible to combine the chemiresistor with more sensitive catalytic gate FET (field effect transistor) sensors to increase the range of response. The addition of analog and digital circuitry results in a smart integrated sensor with a very wide dynamic range. At The Aerospace Corporation (Aerospace), we have developed a chemiresistor sensor for H_2 based on this technology for a satellite application. A serious problem in the reliability of wide-bandwidth GaAs devices is contamination by H_2 gas.[19,20] In hermetically sealed device packages, large partial pressures of hydrogen have been found and apparently evolve from the metallic materials used in the packaging. Hydrogen has been suggested as the cause for premature failure of these devices. One possible failure mechanism is that the platinum (Pt) used in the device is catalytically active for the dissociation of H_2. This could lead to the formation of a metal hydride, which alters the device's performance. The fact that a hydrogen sensor has been made recently from a Pt/GaAs Schottky diode supports this hypothesis.[21] To confirm that H_2 is responsible, we developed and tested a chemiresistor for *in-situ* detection of H_2 in these device packages. Furthermore, such sensors could be incorporated into the packages to serve as simple passive monitors for H_2 levels to verify their health prior to launch.

A photograph of the Aerospace sensor is shown in Fig. 11.4. The active regions consist of thin films of Pd/Ni alloy (12% Ni) deposited onto a 1.0-cm^2 silicon die with 250 Å of thermal oxide. The metal was deposited directly on the patterned photoresist, and the resulting metal pattern was produced by liftoff. The stripes are for resistance measurements and the dots are for MOS capacitance measurements. Gold pads were deposited onto the capacitors and resistor ends for wire bonding. The device was mounted in a standard 24-pin dual in-line package. This small area contains 24 separate sensor elements that respond to H_2 through changes in their electrical characteristics. Our results focus on the chemiresistor that shows changes in resistance proportional to the H_2 pressure.

Figure 11.5 shows the long-term response of a resistor to cycling between high-purity argon and a mixture of 10.5-ppm hydrogen in nitrogen in the flow cell. Drift in the baseline was observed (corrected in the figure) and is probably the result of the gradual purging of the experimental apparatus and temperature drifts. Given the relatively high signal-to-noise ratio of the data, with temperature compensation incorporated into the device, it should be possible to detect H_2 pressures down to 1 ppm and up to 100% H_2. This means we would have a single sensor with a potential dynamic range of six orders of magnitude.

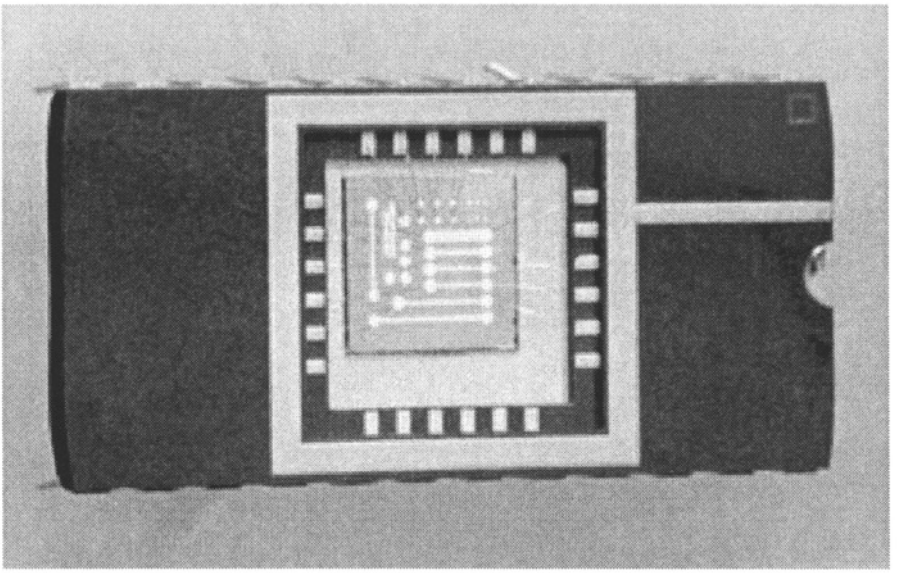


Fig. 11.4. Photograph of the chemical microsensor for H_2. The active areas are a Pd-Ni alloy, which changes resistance with H_2 pressure. There are 24 sensor elements on the 1-cm^2 Si substrate.

Figure 11.6 shows all of our data obtained from the different sensor elements, devices, apparatus, and laboratories plotted on the same graph. It is significant that all of the data lie on a smooth curve, indicating that the sensor response is insensitive to variations in environment and data acquisition. Therefore, the calibration for H_2 response should be robust and independent of

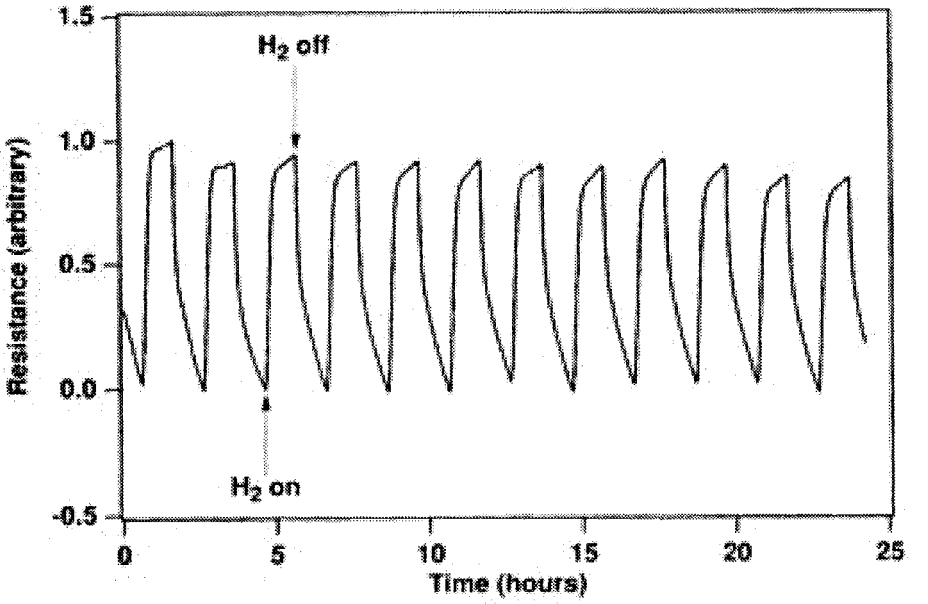


Fig. 11.5. Response of the sensor to alternating flows of pure N_2 and 10.5 ppm H_2 in N_2 in the flow cell. Cycle between 10.5 ppm H_2 mixture and pure argon. Rise time faster than fall time.

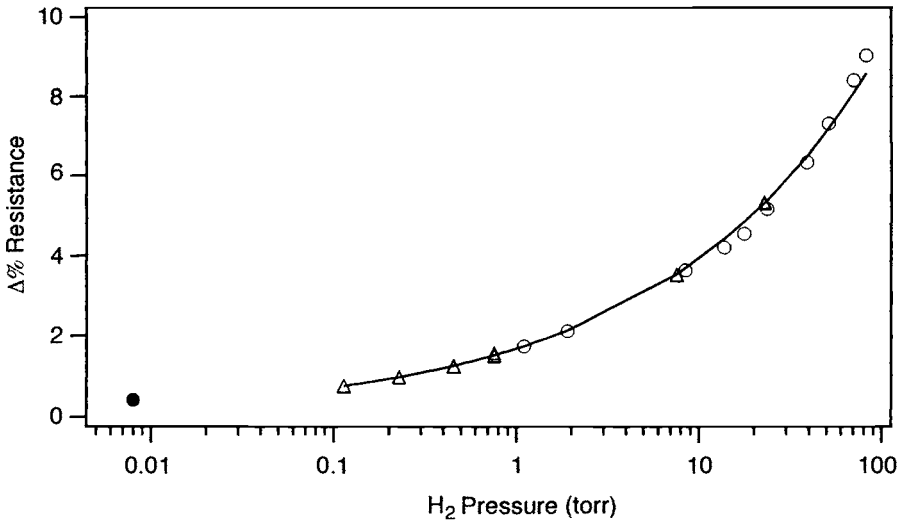


Fig. 11.6. Combined response curve for data obtained in static and flow cells and for different sensor elements, devices, and background temperatures. The observation of a smooth response to all implies a robust calibration. We have measured response from 10 ppm to ~0.1 atm for a dynamic range of 10^4.

experimental conditions. The form of the data in Fig. 11.6 is similar to that obtained by Hughes and Schubert at Sandia for a similar device (referenced above). As discussed elsewhere,[22] the expected functional form for this data is $\Delta R/R \sim (pH_2)^{0.5}$, where pH_2 is the hydrogen pressure. This can be understood from the following chemical reaction and equations:

$$H_2 = 2\,H_b, \tag{11.12}$$

$$K = [H_b]^2/[H_2] \tag{11.13}$$

$$\Delta R \sim [H_b] = \{K[H_2]\}^{1/2} \tag{11.14}$$

where a molecule of H_2 dissociates on the surface of palladium to form two hydrogen atoms that diffuse into the bulk, H_b. The reaction is at equilibrium at room temperature, and therefore the equilibrium constant can be used to calculate relative concentrations of H_b and H_2. The resistance of the alloy is proportional to the concentration of hydride and therefore is proportional to the square root of $[H_2]$. This is known as Sievert's law, and when the data of Fig. 11.6 are plotted versus the square root of H_2 pressure, a straight line is observed consistent with the formation of a bulk hydride.

The H_2/Pd system is a chemisorption reaction since the H-H bond is broken, but the activation energy is so low that the reaction reaches equilibrium at room temperature. Furthermore, the thermodynamics of Pd-H formation are such that the hydride is favored at lower temperature. Therefore, sensor response is inversely proportional to temperature, unlike other chemisorption coatings.

A significant issue in the development of this and other chemical sensors is the selectivity for the target molecule. The number of potential interferences is limited for the application of interest, especially if the devices are sealed in a controlled inert atmosphere. Water is a particular concern, since as much as 5000 ppm is allowed in such packages. Therefore, we examined the response of the sensors to this level of partial pressure of H_2O. The response to 4.1 torr of pure H_2O

(equivalent to 5395 ppm of H_2O at atmospheric pressure) is less than the observed response to 10 ppm H_2. Alternatively stated, the sensor is at least 500 times more sensitive to H_2 than H_2O. Furthermore, the presence of this partial pressure of H_2O does not inhibit the response to H_2, and therefore humidity is not a problem for this application.

Oxygen is another concern for this sensor. It has been noted that the response of Pd-based chemical sensors to H_2 is strongly affected by the presence of O_2.[23] It is believed that H_2 reacts catalytically with O_2 to form H_2O, thereby reducing the hydride concentration. Although there should be little O_2 present in the hermetically sealed packages, we have examined the effect of O_2 on the sensor response to a fixed concentration of H_2 (1000 ppm). As expected, O_2 significantly inhibits the response of the sensor. When the data are plotted in terms of the normalized $\Delta R/R$ versus O_2 fraction, we see that 1% O_2 gives a 10% reduction in response, while 8% O_2 gives a 50% reduction. Therefore, in order to determine the H_2 concentration reliably, the O_2 concentration in the atmosphere must be accurately known. A simple solution to this issue would be to seal the device in a well-controlled, oxygen-free atmosphere. Alternatively, it may be possible to use a separate O_2 sensor or pattern-recognition techniques to deconvolute the H_2 and O_2 concentrations.[24]

11.2.1.1.2 O and O_3 Chemiresistor

Using similar deposition and patterning techniques, we have also developed a sensor for the detection of atomic oxygen (AO) or O_3 based on thin metallic silver films that can be integrated with electronics. An issue of interest for spacecraft in low Earth orbits is the effect of AO on spacecraft materials, so the detection of AO fluxes is important in quantification of spacecraft aging. Detection of AO can be accomplished by spectroscopic techniques, but this approach requires hardware that is relatively heavy and complicated. Chemical sensors offer significant advantages in simplicity, weight saving, and measurement redundancy. Thin films of silver metal have been shown to be effective for the detection of AO by detecting changes in either resistivity or mass.[11,25] This technology has also been shown to be useful for the detection of ozone by mass changes but not resistance.[26] Therefore, we pursued the development of resistive sensors that could be used for the detection of AO or ozone.

The sensors were fabricated using standard silicon microelectronics techniques, as used for the hydrogen sensor developed earlier. Planar magnetron sputtering was used to deposit 350 Å of silver over patterned photoresist on oxidized silicon wafers. Electrical contacts were made via gold contact pads and wire bonding to a standard 24-pin chip carrier. Witness plates were used for film characterization. Quartz crystal microbalance (QCM) crystals were also coated with silver for mass measurements. The sensors were tested by flowing mixtures of oxygen and nitrogen through a photolytic ozone generator and into a cell that contained the sensor and the silver-coated QCM. A computer controlled the flow rates of oxygen and nitrogen and switched the ozone concentration on and off. The mass changes on the QCM and the resistance changes were measured and recorded simultaneously. The concentration of ozone was separately measured by IR absorption to be 140 ppm.

Figure 11.7 shows the response of the sensor to cycling the ozone by switching the lamp on and off. The resistance increases simultaneously with the mass increase on the QCM, indicating rapid oxidation of the silver film. However, the response on the QCM is faster than the resistance change. Both signals decay slowly at room temperature when the ozone is turned off. These results are intriguing, since they are quite different from what has been previously reported for AO—that a stable oxide layer is formed. One possible explanation is the formation of a different oxide phase upon ozone exposure that is metastable and decays slowly at room temperature.

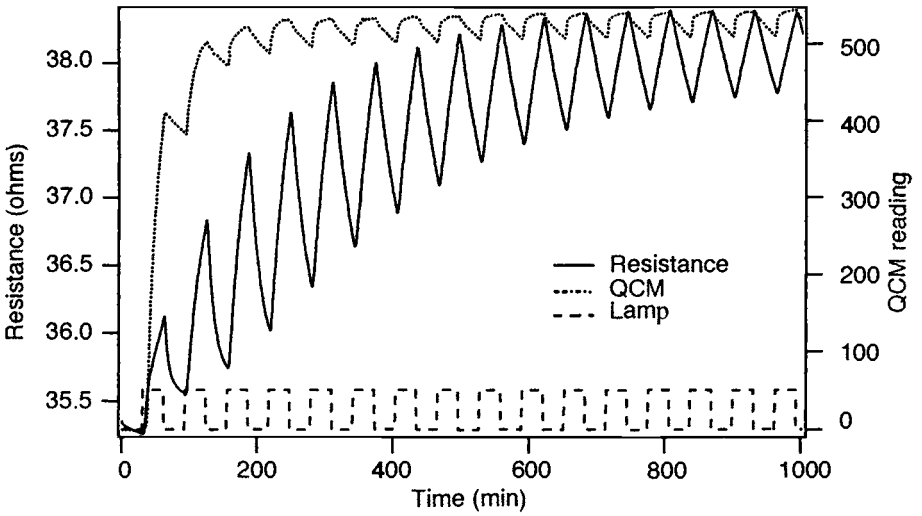


Fig. 11.7. Response of the sensor to ozone cycling on and off in a flow cell. The ozone is generated from photolysis of oxygen and is toggled on/off by the lamp. A silver-coated QCM is used for comparison monitoring of the ozone. Note the simultaneous mass and resistance change and the decay at room temperature.

Further studies are in progress to address this question and determine sensor feasibility. An important general result of this work is the utility of using two types of detection methods in elucidating mechanisms of sensor response. For example, many possible explanations for the observed resistance changes can be discarded because of the observed coincident mass changes on the QCM sensor. These sensors are also being tested for AO response for potential use in a sounding rocket.

11.2.1.2 Organic Conductors

11.2.1.2.1 Conducting Polymers

Conducting organic polymers have proven useful as chemical sensors. There are two types of conducting polymers: intrinsic conductors, in which the polymer material itself is conducting; and extrinsic conductors, in which a nonconducting polymer is made conductive by loading the polymer with appropriate material such as carbon black.[27]

The prototypical intrinsic conductor is polyacetylene (Fig. 11.8) in which conductivity is the result of an extensive pi bond conjugation and results in the formation of a band structure much like a semiconductor. The conductivity is extremely sensitive to doping and can be varied over many orders of magnitude by exposure to chemicals such as $NOPF_6$ or HCl. Other polymers that fit this class are also shown in Fig. 11.8. A particular issue with these materials for sensor applications is that some are not stable in air. Some of the more stable intrinsic conducting polymers are polyaniline, polypyrrole, and polythiophene.

Intrinsic conducting polymers have been known to be sensitive to gases for many years, and there have been reports of sensors for gases such as NH_3, HCl, and hydrazines. The primary principle of operation is that the analyte acts as a dopant, although swelling can also play a role. Whether the response is reversible or not depends on the specific interaction of the analyte with the polymer. For example, the response of polypyrrole to NH_3 is reversible;[13] whereas the response of polythiophene to hydrazine is irreversible.[14]

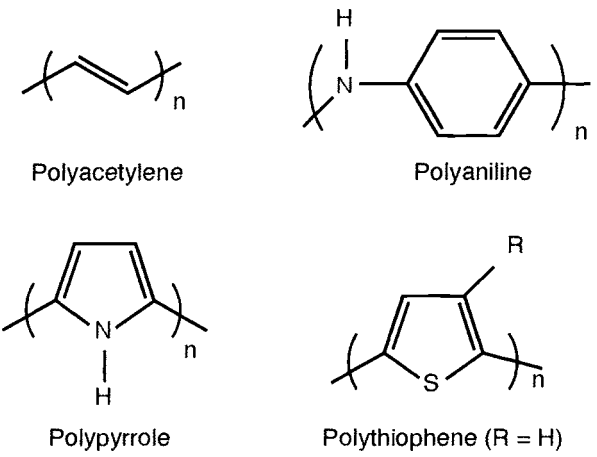


Fig. 11.8. Examples of conducting polymers.

The most relevant intrinsic polymer sensor here is a very sensitive dosimeter for hydrazine detection made from poly(3-hexylthiophene) films under Air Force and NASA funding.[14] In this case, the hydrazines act as reducing agents for the polymer, and irreversible changes in the conductivity are observed. The reported sensitivity for threshold detection is 1 ppb-h. This sensor has been integrated with a microcontroller into a small instrument that can be worn as a personal dosimeter. The time response is quite rapid, and the rejection ratio for NH_3 is very high. Limitations of the device are that, since it is an irreversible dosimeter, the sensor element has a limited range of response and must be replaced when the limit is reached. Also, the sensor elements have a short lifetime at room temperature and are even limited at low temperature. Therefore, this technology may be useful for personal dosimetry, but it is not clear if it will ever be useful for long-term monitoring applications.

Extrinsic conducting polymers are perhaps more versatile than intrinsic conductors because of the wide range of polymers that can be used. The mechanism of their response to gases is sorption, which causes a swelling of the polymer. Swelling lowers the conductivity by essentially reducing the density of carriers in the material, strictly due to volume changes. This provides a simple transduction mechanism that permits utilization of the vast body of gas sorption knowledge, as in the development of polymer-coated SAW sensors. Furthermore, almost all of these types of sensors will be reversible, since the energetics of the interaction are relatively small.

Most recently, a group at California Institute of Technology (CalTech) has demonstrated the utility of using an array of extrinsic conducting polymers as sensors.[17] The use of polymer coatings for chemical sensors has received much recent attention, especially in using the sensor array approach to chemical identification, as discussed above. A particularly powerful combination is an array of SAW sensors with polymer coatings. However, high-frequency SAW sensors require fairly complicated electronics, which limits the size, cost, and complexity of the final devices. The Caltech group showed that with the simple but clever approach of using extrinsic conductive polymers with a range of sorptive properties, an effective array sensor could be realized which takes advantage of simple chemiresistor technology. They fabricated an array of 17 polymers made conductive with carbon black. This gives a very stable sensor coating and allows practically any polymer to be used. Control over individual sensing elements could be realized by varying the content of carbon black as well as the polymer thickness. Similar molecules such as benzene/

toluene and methanol/ethanol could be distinguished with this simple technology. However, the demonstrated sensitivity was only in the 0.1% range. As with any sensor array, especially where no one element is very selective, great attention must be paid to the data analysis in order to properly identify the chemicals. Furthermore, this approach works well when the set of possible chemicals is known, but problems may occur when there is exposure from chemical unknowns outside this set. A group at NASA Jet Propulsion Laboratory (JPL) is developing this approach as an artificial nose for use on the ISS.[6] An array of eight sensor elements is being used with pattern recognition algorithms for chemical identification. Sensitivity in the 25- to 50-ppm range has been demonstrated for some chemicals on the SMAC list. However, it is not clear at this point if chemical identification can be accomplished with this device. The electronic nose is scheduled for a flight on the Space Shuttle in 1998, and a small company has been formed to explore this technology.*

11.2.1.2.2 Phthalocyanines

Phthalocyanines (Pc) are conjugated, macrocyclic, organic molecules with the basic structure shown in Fig. 11.9. At the center of a Pc is a metal ion, which can vary across the periodic table. In addition to being extremely stable up to temperatures of 500°C and intensely colored, thin films of Pc's are molecular organic semiconductors. Their conductivity displays an increasing exponential dependence on temperature typical of semiconductors:

$$\sigma = \sigma_0 \exp(-E/2kT), \tag{11.15}$$

where σ is the specific electrical conductivity, σ_0 is the intrinsic electrical conductivity, E is an energy gap, and k is the Boltzmann constant. This is the result of thermal enhancement of carrier concentration in the conduction band. The electrical conductivity of Pc's is sensitive to exposure to different gases, presumably through a doping mechanism that depends on the electronegativity of the gas and the work function of the film. This, however, is a very crude picture and does not entirely account for the observed behavior of Pc's. In any event, the observation of gas sensitivity has led to extensive investigations into the potential for Pc's for use as chemiresistors. Most notable in the context of this book are the demonstrated sensors for NO_2,[12] and the related molecular semiconductor, dithiolene, for hydrazines.[28]

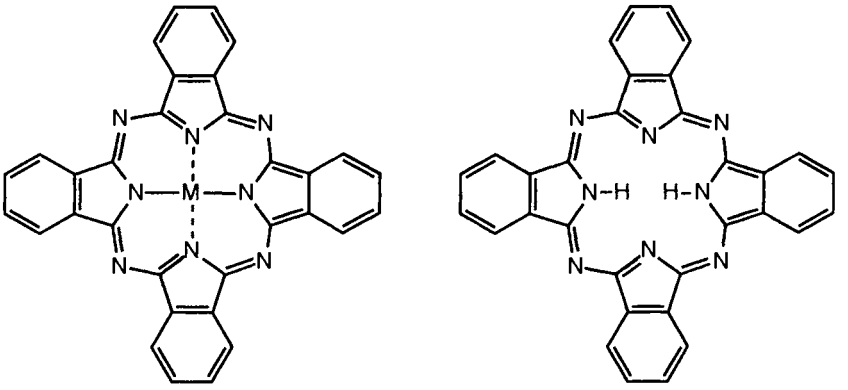


Fig. 11.9. Structures of phthalocyanines, metal substituted (left) and metal free (right).

*Cyrano Sciences, Inc. (http://www.cyranosciences.com)

The electrical conductivity of Pc's is quite low, especially at room temperature. In order to increase the conductivity and response time and to obtain a reversible response, it is necessary to heat the sensors in the range of 120°C to 200°C. Furthermore, interdigitated electrodes are generally used to further lower the resistance of the films. A typical sensor substrate used for this application is shown in Fig. 11.10. It is a 3- × 3-mm piece of alumina with interdigitated electrodes on the top surface and an integrated platinum heater and temperature sensor on the back side. For thermal insulation, the substrate is suspended by leads to posts in a standard TO-5 header. The substrate can be heated to 500°C, while the package remains at room temperature.

At Aerospace, we have developed a sensor for HCl gas based on Pc. To fabricate the sensors, thermal evaporation was used to deposit phthalocyanines on the substrate. To evaluate the quality and quantity of material deposited, a quartz witness plate was used for optical measurements. For quantitative deposition rate data, a QCM was used to provide real-time, *in situ* measurements. The sensors were tested by exposure to known concentrations of HCl and other gases under automated control for measurements of the long-term response. Dilution of a calibrated HCl gas mixture with N_2 was accomplished with computer-controlled mass flow controllers. A sensitive electrometer measured resistance, a stable power supply drove the heater, and a voltmeter/ammeter measured the heater resistance in order to determine the substrate temperature. The substrate temperature was previously checked against IR emission measurements with an IR camera.

Figure 11.11 shows the response of a sensor to cycling between high purity N_2 and a mixture of 2 ppm HCl. The temperature was held constant at 150°C with a power input of ~250 mW. The resistivity decreases upon exposure to HCl, and the response is very reproducible with relatively low noise. It should be noted that the actual HCl concentration is probably much lower, since wall

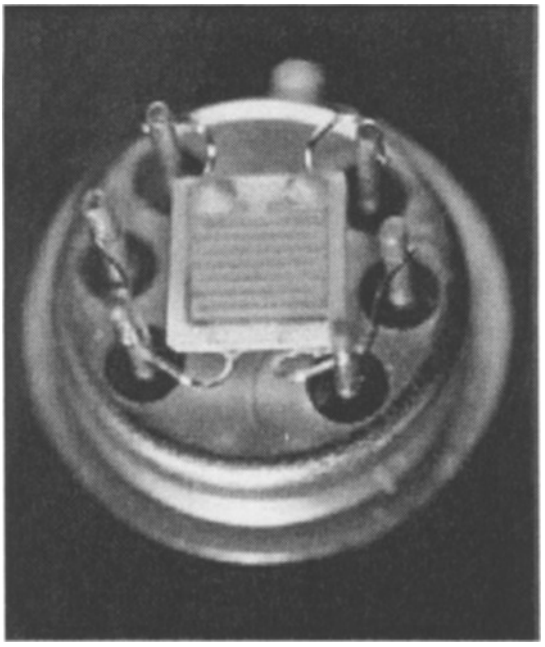


Fig. 11.10. Photograph of the HCl sensor. The substrate is 3 × 3 mm alumina with interdigitated electrodes on top and integrated heater and temperature sensor on bottom. The electrodes are coated with a phthalocyanine.

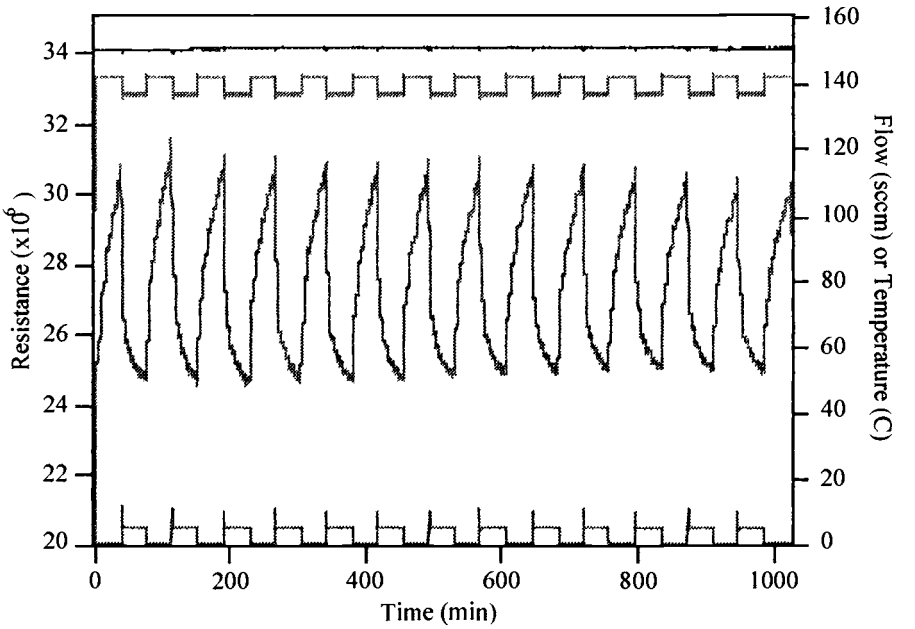


Fig. 11.11. Alternating response of the HCl sensor to 2 ppm HCl and pure N_2.

losses are expected. In addition, preliminary data show a much greater response at higher temperatures. Given these observations, it should be possible to detect HCl down to 0.5 ppm. This makes it a promising candidate for HCl sensing applications. We are currently working on improving the response time and reducing the power requirements of this sensor via the use of micromachined sensor substrates.

11.2.2 Catalytic Gate MOS Devices

The catalytic gate metal oxide semiconductor (MOS) device is a well-developed chemical sensor, especially for hydrogen detection. Lundstrum *et al.* were the first to demonstrate in the early 1970s that replacement of the gate metal in field effect devices (i.e., transistor and capacitor) with a catalytic metal such as Pd gave very sensitive, selective sensors for hydrogen gas.[29] Under UHV conditions, detection limits as low as 30 ppt are possible. Figure 11.12 shows two typical MOS devices used for chemical sensing, a capacitor and a transistor; their principle of operation is as follows. A voltage is applied to the gate, and both devices function by the induced formation of a layer of charges (electrons or holes) at the interface of the gate metal with the semiconductor. This buildup of charge at the interface can either form a capacitor or can induce conductivity by allowing current to flow from the source to the drain in the transistor. As discussed earlier, Pd has well-known, unique properties in regard to the facile dissociation of H_2. When the normal gate metal (aluminum) is replaced with palladium, hydrogen dissociates on its surface to form atomic hydrogen, which diffuses into the film to form a hydride. Some of the hydrogen atoms adsorb at the interface with the SiO_2 gate oxide. These hydrogen atoms form a dipole layer at the interface and effectively give rise to an additional voltage in series with the applied voltage, which either causes a shift in the C-V curve or in the threshold voltage. Figure 11.13 shows the response of the devices to gate voltage with the effect of hydrogen gas on that response. The sensors can also be made to respond to ammonia or hydrogen sulfide by modifying the gate metal.[29]

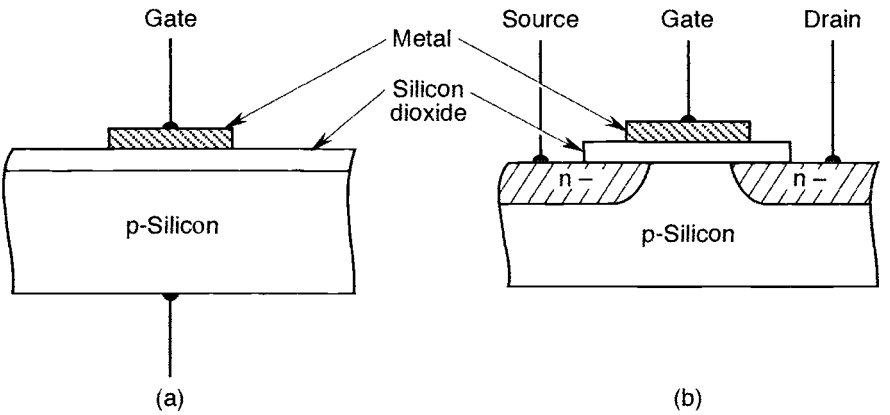


Fig. 11.12. Typical field effect devices used as chemical sensors on a p-doped silicon substrate. (a) Capacitor and (b) transistor.[29]

Some issues with these devices are that the response time at room temperature is somewhat slow, and interference from adsorbed water is a problem. Therefore the devices are operated at elevated temperatures (150°C), which can require significant power (0.5 to 1 W). A resistive heater, temperature-sensing diode, and control circuitry can be readily implemented in silicon on the same chip. In addition, the range of hydrogen concentration that can be sensed with these devices is limited due to saturation effects. On the other hand, a great advantage of this sensor is that it is fabricated using standard silicon microelectronics processing techniques. Therefore, the sensor can be readily integrated at the chip level with CMOS circuitry for signal conditioning. Cost savings are inherent in the batch wafer processing approach as well. This has been demonstrated very nicely by J. L. Rodriguez *et al.* at Sandia National Lab.[30] They integrated a Pd/Ni chemiresistor with a Pd/Ni transistor, temperature control, and data acquisition circuitry. Pd/Ni was used because of material issues with Pd at high H_2 concentrations (see above). The use of two hydrogen sensors provides usable response over a wide range of hydrogen concentration, from the low ppm level well into the explosive region (> 4% H_2 in air). This device has been licensed and is now being produced by a small company.*

As with the Pd chemiresistors, an important issue for these devices is their sensitivity to oxygen. Oxygen acts as a scavenger for hydrogen by reacting with the dissociated hydrogen to form water. The response of the sensor to hydrogen in air is 4 to 5 orders of magnitude less than in argon. Therefore, the oxygen concentration must be constant, or if it is variable, then it must be measured to separate the hydrogen and oxygen concentration changes.

Similar sensor technology has been developed at NASA Lewis using Pd/Ag in a Schottky diode for the detection of hydrogen propellant leaks in launch vehicles.[31] The Sandia sensor has also been evaluated for this application.

11.2.3 Fiber-Optic Chemical Sensors

Fiber-optic sensors provide an important method for chemical detection. There are numerous approaches, all of which take advantage of the ability of a fiber to carry light long distances from a source to a detector. In the simplest implementation, fibers are used as light conduits for absorbance or emission spectroscopy. Coupled with recent developments in micromachined

*DCH Technology, Valencia, CA.

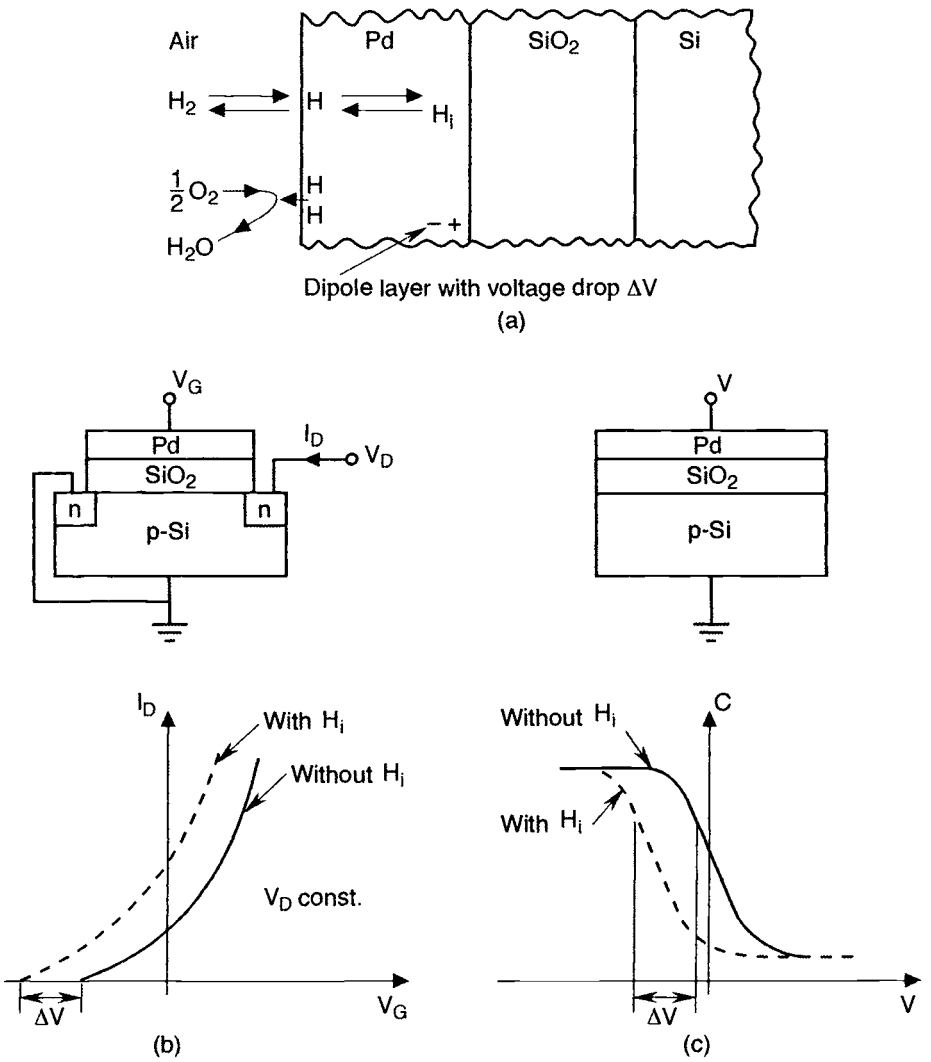


Fig. 11.13. Change in applied gate voltage as function of H_2 pressure. (a) Dipole layer with voltage drop ΔV, (b) capacitor, and (c) transistor.[29]

spectrometers and array detectors, fibers could provide a useful method of chemical analysis for space applications. However, in this application, the fiber is a light conduit and not a sensor.

Fiber-optic chemical sensors result when thin-film coatings are used to modify the optical properties of a surface to produce a change in response to chemical exposure. The surface that is modified can be the end of a cleaved fiber, in the fiber cladding, or an external micromirror. These approaches are discussed according to the type of interaction of light with the surface: (1) evanescent wave absorption and (2) reflectivity of micromirrors.

In the evanescent wave approach, the multiple internal reflections inside a fiber are used to perform highly sensitive absorbance measurements of coatings embedded in the cladding of fibers. At each reflection, there is an evanescent wave that propagates a short distance into the cladding. Although the individual propagation distance is short, there are many reflections that

add up to give significant total path, and therefore good sensitivity is obtained for absorbance measurements. By chemical modification of the cladding so that the light absorption changes upon exposure to a particular chemical, the change in transmitted light intensity can be used for chemical detection. This is probably the most common approach used for fiber-optic chemical sensors.[32] Another advantage of this approach is that, with a pulsed laser and time domain measurements, a single fiber can be used for multipoint detection. Locations along the fiber are distinguished by the pulse arrival time at the detector. This is a great advantage in obtaining distributed measurements over a wide area with a single fiber.

This approach is the basis of a proposed hydrazine dosimeter system under development at Aerospace for use at Cape Canaveral.[33] The proposed system has a 1-km trunk length and a fiber-optic network of 100 sensors in a parallel star configuration. With a 10-mW, 680-nm diode laser, it is estimated that a 5% change in light intensity could be measured for an individual sensor with a signal-to-noise ratio of 5. As with the other chemical sensors discussed, the key to chemical detection is an appropriate coating for selective chemical transduction. In this case, 12-molybdophosphoric acid ($H_3PO_4 \cdot 12MoO_3 \cdot xH_2O$) was found to be a useful indicator for the presence of hydrazine. It changes color from bright yellow to blue upon reduction by hydrazine, but the chemical reaction is not spontaneously reversible (a dosimeter). However, the reaction can be reversed by subsequent oxidation by NO_2. (This could be a potential problem, since NO_2 is the oxidant used in combination with the hypergolic rocket fuels.) The fiber cladding was replaced with a sol-gel coating containing molybdophosphoric acid. This gave dosimeters with sensitivity of 2.3 ppb-h at the 5% level and a reproducibility of ± 1 ppb-h. The sensitivity for NH_3 was determined to be 9200 times less than for hydrazine. While this system holds promise for hydrazine detection given a fiber-optic network infrastructure, issues to be resolved are reproducibility and chemical selectivity. The use of a fiber-optic network eliminates many of the wiring problems and is equivalent to having networked sensors on a simple twisted pair line. However, installation of a fiber is required and is a significant implementation hurdle for most applications. In this case, a fiber-optic network will be built for communication purposes, and this system will take advantage of that infrastructure. But without such an infrastructure, it is much simpler to implement a wireless data-transmission scheme.

In the reflectivity or micromirror approach, which has been exploited with much success at Sandia National Lab, coatings are used that change reflectivity when exposed to the chemical of interest.[34] Either the end of a single fiber is coated or an external surface is used as the micromirror. The advantage of coating the end of a fiber is that a single fiber is used to transmit the probe and reflected light beams. The coating is typically a metal which changes in reflectivity upon exposure to a particular chemical. Examples of micromirror sensors that have been developed include H_2 (Pd) and others. With an external micromirror, two fibers are used, one for illumination of the micromirror and a second to carry the reflected light to the detector. This configuration allows a large number of micromirrors to be probed separately with a fiber-optic bundle, a single light source, and an array detector. This provides a powerful way to sense many different chemicals with a relatively simple optical arrangement. This approach was used in the MOx chemical sensor instrument designed to measure soil samples on Mars.

The MOx was to have been flown on the Russian Mars 1994 mission and was designed to investigate the soil chemistry on Mars. It was the first planetary probe or spacecraft to incorporate chemical microsensor technology.[35] The instrument was designed to characterize the chemical nature of the Martian soil, in particular its proposed oxidizing nature. This was to be accomplished using fiber-optic technology to study the reflectivity changes in a suite of thin-film materials. The design constraints on the instrument were very challenging: small volume, mass less

than 850 g, power consumption less than 25 to 50 mW for short time intervals, shock impact of 250 G, temperature variations of 100°C. The chemical information desired was diverse as well: the oxidizing nature of the soil, effect on organic materials, potential biological activity, and pH. The solution was an instrument that used an array of various micromirrors, which were optically probed using fiber optics and only two light sources and a single detector. The optical layout used two LED sources to provide dual wavelength capability for reference and probe. The detector was a 256-pixel linear diode array detector, which allowed for as many as 256 separate micromirror samples with this simple arrangement. Key features of the design are fiber guides that were micromachined in silicon to provide precise and reproducible alignment with the plate containing the micromirrors and a micromachined SiN membrane that would provide a hermetic seal until landing on Mars. The array of films used includes reference films for constant reflectivity (Au/Pt/Ti) and temperature (Au/Pt/Ti/Si) and to check for dust accumulation (bare SiO_2). An array of metal films probe for frost (Au/Pt/Ti), H_2, H_2S or unsaturated organics (Pd), O and O_3 (Ag), and other oxidants (V, Ti, Al, Mg). Kerogen-like films are used to simulate the carbonaceous material thought to be deposited on Mars by meteors. A range of organic coatings is used to detect the following: polybutadiene for O_3 detection, D- and L-glucose for a chirality preference in Martian reactivity, pH (thymol blue, bromothymol blue, 2,6-dichloroindophenol, fluorescein), O_2 or CO_2 (hematin), and reducing agents (methylene blue). Pattern recognition techniques were to be used for chemical identification. Unfortunately, the Russian rocket failed and the Mars Lander landed in the ocean. In any event, the successful development of flight hardware shows that the technology is viable and can meet the stringent mission requirements for planetary exploration.

11.2.4 Acoustic-Wave Sensors

Another important class of chemical microsensors is acoustic-wave sensors. These devices are ultrasensitive mass sensors and include the well-known QCM and SAW devices, shown in Fig. 11.14. Both devices rely on the generation of acoustic waves in piezoelectric material, typically quartz, via metal electrodes applied to both sides of the crystal. Mass accumulation at the surface of both devices is detected as a frequency shift in the acoustic waves. These mass sensors can be made into chemical sensors by the application of thin-film coatings that selectively adsorb chemicals of interest and are detected by the mass change at the surface. This section was adapted from Ballantine *et al.*,[8] and the interested reader is referred to this excellent reference for more details on acoustic sensors.

11.2.4.1 Quartz Crystal Microbalances

Although the QCM is technically not a microsensor, its operation will be described since it is closely related to the SAW device. The QCM is currently used on several spacecraft as a contamination detector, and provides a sensitivity benchmark. The QCM is a resonator generating shear mode acoustic waves. Shear waves of opposite polarities are created at the electrodes on each face of the crystal. Resonance occurs as a result of constructive interference between the incident and reflected waves. The resonance condition is that the crystal thickness (h_s) is equal to multiples of half acoustic wavelengths: $h_s = n(\lambda/2)$. This can be expressed in terms of the phase velocity, which is related to the crystal properties:

$$f_n = v_s/2h_s \qquad v_s = (\mu_q/\rho_q)^{1/2}, \tag{11.16}$$

where f_n is the frequency of the nth mode, v_s is the phase velocity of the shear wave, h_s is the crystal thickness, μ_q and ρ_q are the shear stiffness and density. The resonant frequency for a crystal with $h_s = 0.033$ cm, $\mu_q = 2.95 \times 10^{11}$ dynes/cm^2, and $\rho_q = 2.65$ g/cm^3 is $f_1 = 5.06$ MHz, a typical number for QCM devices.

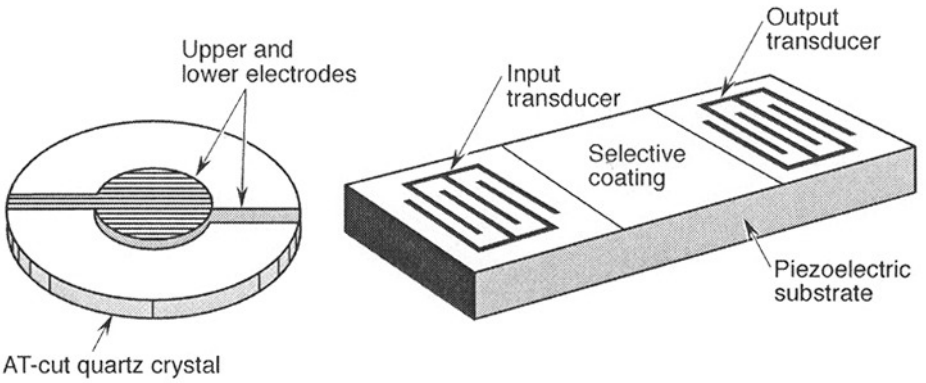


Fig. 11.14. QCM and SAW devices.[8]

Mass accumulation at the surface of a QCM results in a shift in the resonant frequency of the crystal. This is a result of the requirement that the kinetic and potential energy densities of the wave balance. By equating expressions for the energy densities and making a linear approximation, the following equation results:

$$\frac{\Delta f}{f} = -\left(\frac{\rho_s}{h_s \rho_q}\right) \tag{11.17}$$

where ρ_s is the density of the surface film. Equation (11.17) shows that the fractional change in resonant frequency is proportional to the fractional change in mass due to the accumulated layer. From Eq. (11.17), a responsivity factor R for the QCM can be derived:

$$R = df/d\rho_s = -f_0/h_s\rho_q. \tag{11.18}$$

For the crystal parameters used above, $R = -57$ Hz-cm^2/μg; or for every 1 μg/cm^2 mass change, the frequency decreases by 57 Hz. The sensitivity can be calculated, given a noise level that is typically due to the stability of the oscillator used. If the oscillator stability is 0.1 Hz and the limit of detection is a signal-to-noise ratio of 3, then the limit of mass detection (LOD) is:

$$LOD = 0.3\ Hz/(57\ Hz\text{-}cm^2/\mu g) = 5\ ng/cm^2. \tag{11.19}$$

11.2.4.2 Surface Acoustic-Wave Devices

The SAW device shown in Fig. 11.14 is a true microsensor in that micron-sized electrodes are required to excite the surface acoustic waves. Surface acoustic waves are Rayleigh waves in which the acoustic energy is confined to the surface of the solid. In 1970, R. M. White of University of California, Berkeley, discovered that surface acoustic waves could be generated and detected using interdigitated electrodes on the surface of a piezoelectric material.[36] Figure 11.15 shows the relationship between the electrode spacing and the resultant surface acoustic wave. The efficiency of a SAW device is maximized when the electrode spacing matches the SAW wavelength. For ST cut quartz, the propagation velocity is $v = 3.158 \times 10^3$ m/s, and the frequency and wavelength are related to the wave velocity by $f\lambda = v$. For a 97-MHz device, the acoustic wavelength is 32 μm, and this should be the spacing between fingers on one electrode. In order to achieve these dimensions, SAW devices are manufactured using microelectronics processing techniques with quartz substrates. Photolithography is used to pattern the electrodes, and metal is deposited by evaporation or sputtering. Generally, aluminum or gold is used, although gold is preferred because it is more chemically inert.

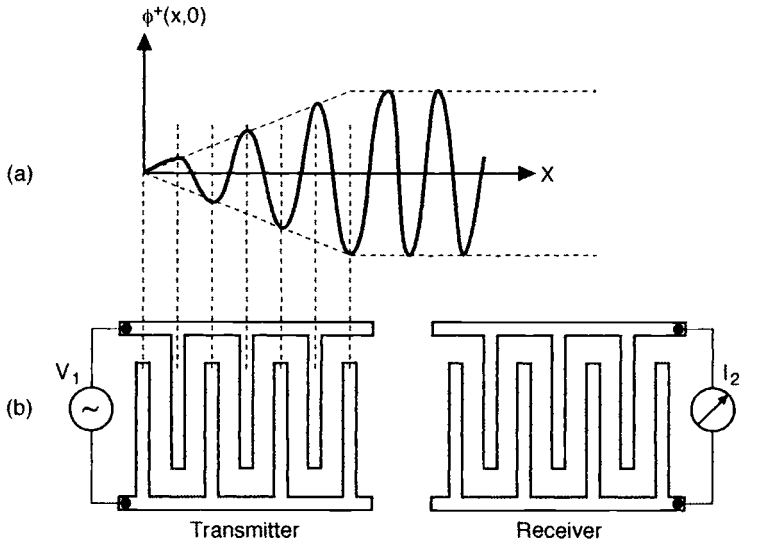


Fig. 11.15. Relationship between electrode spacing and the generated surface acoustic wave.[8]

Two types of SAW devices are typically used as chemical sensors, the delay line and the resonator, both of which are shown in Fig. 11.16. The delay line has two sets of electrodes; one set is used to create an acoustic wave defined by the electrode spacing and is detected by a second set of electrodes after propagation delay along the crystal. The resonator has a high Q acoustic cavity created by a series of ridges in the surface of the SAW crystal on either side of the interdigitated electrodes. One small electrode is used to launch the SAW, and the second is used to sample the wave and feeds back into the driving circuit. Fewer fingers are required for the electrodes in a resonator compared with the delay line, since the acoustic wave is defined by the resonant cavity.

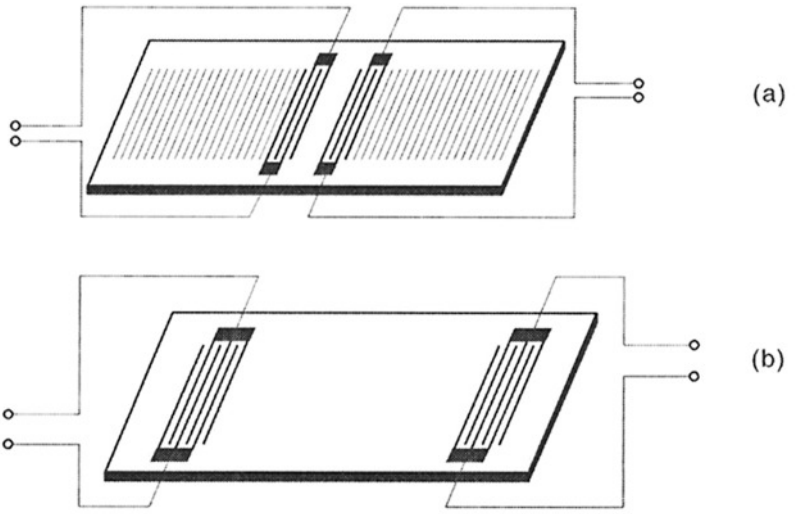


Fig. 11.16. SAW (a) resonator and (b) delay line.[8]

The sensitivity factors for SAW devices can be calculated* and are shown in Table 11.4 along with the value for the QCM. One can see that the sensitivity factors for the SAW devices are an order of magnitude greater than for the QCM, and this is largely driven by the higher frequency of operation of SAW ($S \sim f_0^2$). However, this is somewhat misleading, and it is important to consider the limits of detection for the various devices. These are determined by the noise levels of the measurement setup, which is typically about 0.1 Hz for QCMs and about 1 Hz for SAWs. A trade-off in using the higher frequency SAW devices is an order of magnitude higher noise level. The limit of detection is considered to be a signal-to-noise ratio of 3, and the limits of mass detection are calculated from:

$$LOD = N_f/R_m \tag{11.20}$$

where N_f is the noise associated with the frequency measurement and R_m is the responsivity of the device. Using noise levels of 0.3 Hz for QCMs and 3 Hz for the SAW devices, the results are shown in Table 11.4. Thus, the 97-MHz delay line is only 17 times more sensitive than a 6-MHz CM; whereas the 200-MHz resonator is 58 times more sensitive. It should be noted that noise levels in experimental setups are highly dependent on many factors. For example, thermal effects can be important in causing long-term drift and can severely degrade the limits of detection.

11.2.4.3 Acoustic-Wave Chemical Sensors

The key to converting the above mass sensors into chemical sensors is the application of a chemically selective coating that selectively absorbs the chemical of interest and causes a mass change at the sensor surface. The useful coatings range from porous polymers that simply absorb gases to inorganic reagents that irreversibly react with a specific gas. For all acoustic-wave sensors in which mass loading is the dominant effect, the frequency shift is linearly proportional to the mass change at the surface:

$$\Delta f = -kS_m\Delta m_a \tag{11.21}$$

where Δf is the frequency shift, Δm_a is the change in mass per unit area at the device surface, R_m is the mass responsivity of the device, and k is a constant that contains geometric and other response factors. The important point is that for identical devices and geometries, the frequency shift is directly proportional to mass change at the surface. However, it should be noted that in addition to the mass effect, there can be other response mechanisms of response related to electrical and viscoelastic properties that are not accounted for by the above equation.

Table 11.4. Mass Sensitivities of Acoustic-Wave Devices[a]

Device	Theoretical Mass Sensitivity (S_m) (Hz-cm^2/mg)	Frequency (MHz)	Experimental Mass Sensitivity (S_m) (Hz-cm^2/μg)	Limit of Detection (ng/cm^2)
QCM (AT quartz)	$2.3*f_0^2$	6	84	3.5
SAW delay line (ST quartz)	$1.32*f_0^2$	97	12,200	0.2
SAW resonator (ST quartz)	$1.26*f_0^2$	200	50,400	0.06

[a]From Ref. 8.

*This is beyond the scope of this chapter, but the interested reader is referred to Ref. 8.

The types of coatings used for acoustic-wave chemical sensors can be divided into three categories based on the physical and chemical interaction between analyte and the coating: physisorption, chemisorption, and absorption. Physisorption is a weak molecular interaction that is typically not specific and is not very useful in the development of chemical sensors. Coatings that have been used for this include activated charcoal, silica and alumina gels, zeolites, and porous polymers (Tenax, XAD, Chromsorb). Attachment of the particulate materials to create a uniform film without plugging the pores is a challenge.

11.2.4.4 Chemisorption Coatings

As described earlier, chemisorption reactions at surfaces require significant energy input to occur. Therefore, the sensor responsivity can increase with increasing temperature, and the resulting sensors are usually irreversible dosimeters at room temperature. A sampling of chemisorption coatings used for acoustic-wave sensors is listed in Table 11.5.

Some examples are the use of metals such as Zn, Pd, Au, and Ag for the detection of HCl, H_2, Hg and O_3/O, respectively. QCMs coated with Zn were found to be useful sensors for HCl through a corrosion reaction to form $ZnCl_2$ irreversibly.[37] $ZnCl_2$ is hygroscopic, and the formation of $ZnCl_2 \cdot (H_2O)_x$ gives a mass amplification and increased sensitivity. This sensor was developed as a fire detector for electrical equipment that contains polyvinyl chloride. However, this sensor is useful only as a threshold alarm detector due to the reactivity of the $ZnCl_2$ with H_2O. The ability of O and O_3 to oxidize Ag has been known for many years and has been used to make chemiresistor sensors, as described above. The same effect can be used as the basis of acoustic-wave sensors. The formation of the oxide is generally irreversible. As discussed earlier in Sec. 11.2.1.1.1, hydrogen is quite soluble in Pd films, and this effect has been the basis ofChemresistor and ChemFET sensors. This same effect has been exploited on a SAW platform to give a H_2 sensor.[41] Mercury can be detected with gold coatings via the formation of a stable amalgam.[40]

Polymer films have also been used to form chemisorption sensors. An ozone sensor has been made from a coating of polybutadiene on a QCM.[38] The ozone reacts irreversibly with carbon

Table 11.5. Chemisorption Coatings for Acoustic-Wave Sensors[a]

Analyte	Coating	Limit of Detection	Comments	Ref.
HCl	Zn	5 ppm	$ZnCl_2$ is hydroscopic, gives mass amplication	37
Ozone	Ag	10 ppm	--	38
Ozone	Polybutadiene	10 ppb	Irreversible	
Olefins	$PtCl_2$(ethylene)(pyridine)	0.6 ppm	Reversible with ethylene	39
Hg	Au		Amalgamation, thermally reversible	40
H_2	Pd	50 ppm	--	41
H_2S	WO_3	10 ppb	High temperature	42
Cyclopentadiene	Poly(ethylene maleate)	200 ppm	Irreversible	43

[a]From Ref. 8.

double bonds to form ozonides. The sensitivity of the sensor was 10 ppb/min. A cyclopentadiene detector was made from poly(ethylene maleate) (PEM) on a SAW substrate.[45] PEM is a useful reversible sensor for a range of organic vapors such as acetone, methylene chloride benzene, and methanol. However with cyclopentadiene, an irreversible Diels Alder reaction occurs, resulting in a strong chemical bond with the coating. While this sensor does not have the selectivity one would like in the chemisorption sensor, the response to cyclopentadiene is considerably greater than that of the other organic vapors.

Semiconductors have also been used as chemisorption coatings for acoustic-wave sensors. Phthalocyanines have been used to detect NO_2. Recall that these organic semiconductors are useful chemiresistor materials that change in conductivity upon exposure to various gases. It was found that with PbPc, NO_2 could be detected with sensitivity of <1 ppm.[46] The oxide semiconductor WO_3 has been used to detect H_2S.[44] The sensitivity was very good at 10 ppb, although the sensor requires operation at high temperature. However, there can be a significant electrical component to the response of these films.

It has been shown that electrical effects can dominate over mass response for certain combinations of coatings and substrates. Because there is an electrical field associated with an acoustic wave in a piezoelectric material, electrical effects can be prominent with metallic or semiconductor films with conductivity changes upon chemical exposure. $LiNbO_3$ has an electromechanical coupling coefficient almost 40 times larger than ST quartz and a mass sensitivity that is almost half. Using this material with phthalocyanine (Pc) coatings, electrical effects have been shown to dominate over mass effects.[47] In this case it was possible to separate mass and conductivity effects by the use of two SAW devices, one in which the Pc was applied directly to the SAW substrate and the other with a Cr underlayer. The Cr layer served to short out any electrical effects, and with that sensor, no response of PbPc to NO_2 was observed. The sensor with PbPc applied directly to the $LiNbO_3$ gave a good response to 10 ppm of NO_2.

11.2.4.5 Polymer-based Sorption Coatings

Polymer films are one of the most common coatings used for SAW chemical sensors. These films function by sorption of gases into the bulk of the polymer involving both interfacial and bulk interactions. Because the interaction of analyte with the polymer film is relatively weak, sorption is at equilibrium at room temperature and sensors based on these films are reversible. The analytes targeted with these films are organic vapors, since they have the highest affinity for polymers. A significant drawback with these films is the lack of selectivity: sorption occurs to some degree with virtually all pairs of volatile organics and polymer films. In order to resolve mixtures of organic vapors, several SAW sensors, each coated with different polymers, are required.

Sorption of gases in polymers is at equilibrium at room temperature, and thermodynamics can be used to quantify the concentrations of absorbed species in the sensor film.* The relative concentrations of gas phase and absorbed species is given by the partition coefficient, K_c:

$$K_c = \frac{C_s}{C_g} = \frac{m_s}{V_s C_g} \tag{11.22}$$

where C_s is the concentration of sorbed species, C_g is the gas phase concentration, m_s is the mass of sorbed species, and V_s is the volume of the coating. This is the same partition coefficient used to describe retention behavior in gas chromatography. For acoustic sensors where mass effects

*These principles apply to the extrinsic conducting polymer sensors described in Sec. 11.2.1.2.1.

dominate, Eq. (11.22) can be used to calculate K_c from the frequency shifts associated with the analyte and that due to the application of the coating to the sensor.* Typically, K_c calculated from SAW devices is significantly larger than values derived from gas chromatography, indicating that the other mechanisms of response are significant. For nonconducting polymer films, this is largely due to viscoelastic effects.

The partition coefficient is an equilibrium constant and can be related to the enthalpy of solution:

$$K_c = \exp\left(\frac{-\Delta G_s}{RT}\right) = \exp\left[-\frac{(\Delta H_s - T\Delta S_s)}{RT}\right] \sim \exp\left(\frac{-\Delta H_s}{RT}\right) \tag{11.23}$$

where ΔG_s and ΔH_s are the free energy and enthalpy of solution, respectively. The entropy of solution of very dilute solutions is negligible. The enthalpy of solution is related to the enthalpies of condensation and of mixing:

$$\Delta H_s = \Delta H_c + \Delta H_m \sim \Delta H_c \tag{11.24}$$

For ideal solutions, ΔH_m is zero and there is no interaction between the analyte and the coating. Clearly this is not the case for real polymer coatings, and a goal of sensor development is to maximize this interaction. However, the condensation enthalpy is significant, and it is instructive to consider its effect on sensor response.

Boiling point is a simple, readily available parameter that can be used to estimate sensitivity for similar molecules. The enthalpy of the condensation is proportional to the boiling point of a simple liquid according to Trouton's rule:

$$\Delta H_c = -\Delta S_c T_b \tag{11.25}$$

where T_b is the boiling point of the liquid, and ΔS_c is the average entropy of condensation for simple, nonpolar liquids (20.3 cal/molK). Therefore, the concentration of the sorbed analyte can be related to its boiling point:

$$C_s = C_g K_c = C_g \exp\left(\frac{-T_b \Delta S_c}{RT}\right) \tag{11.26}$$

From this expression, one can examine the effect boiling point of analyte has on the relative response of a particular sensor/coating combination. For given gas phase concentrations for two analytes, C_s and C_s', the ratio of Δf for one analyte to Δf for a second analyte with a higher boiling point is given:

$$\frac{\Delta f'}{\Delta f} = \frac{\Delta m'}{\Delta m} = \frac{C_s}{C_{s'}} = \exp\left(\frac{\Delta T_b \Delta S_c}{RT}\right) \tag{11.27}$$

where $\Delta T_b = T_b' - T_b$. Based on condensation enthalpy alone, for two similar molecules with a difference in boiling points of 20°, a particular sensor will have at least a twofold greater response for the higher boiling molecule. For example, consider pentane and hexane, which have boiling points of 36.1°C and 68.7°C. The response of a given sensor will be 3.1 times greater for hexane

*The partition coefficient is calculated from the Δf for the coating, the density of the coating, which gives the volume of the coating. $K_c = \Delta f_s r_c / \Delta f_c C_s$.

than for pentane. Sorption-based sensors are much more sensitive for higher boiling point molecules. Conversely, it is quite difficult to detect low-boiling molecules with sorption films.

It is possible to use partition coefficients derived from gas chromatography to estimate SAW sensor response, given various molecules and polymer coating combinations. Grate *et al.* have done a very nice job of estimating limits of detection for various molecules for 14 different polymer films on SAW sensors.[48]

11.2.5 Micromachined Chemical Sensors

MEMS and silicon micromachining have been used to create sensors with current or potential applications to space systems. These include MOx fiber-optic sensor systems described above that contained micromachined components. MEMS sensors based on cantilever beams[49,50] have been developed, but currently these are only laboratory novelties. Perhaps the most relevant micromachined sensors are small, heated elements for conductive sensors.

11.2.5.1 Micromachined Sensor Substrates

Micromachined sensor substrates have been made in order to reduce the power consumption of heated chemiresistors and to take advantage of batch processing possible with silicon microelectronic fabrication techniques. In addition to the phthalocyanine-based chemiresistors described above, there are several other chemiresistor materials that need to be heated for operation. The so-called Taguchi tin oxide sensors are an important product, and also require high temperature (~500°C) and considerable power for operation. By reducing the size of the sensor element, it should be possible to reduce the power consumption. The sensors should also be stronger and more resistant to the shock and vibration associated with launch. Batch processing should give better uniformity and reduce costs, and with silicon microfabrication, signal conditioning circuitry can be integrated on the die with the sensor element. In addition, silicon provides a much smoother surface that should allow much thinner films to be used and could greatly improve the time response.

Micromachined chemiresistor substrates (Figs.11.17 and 11.18) are in essence micro hot plates. Thermal conduction pathways from the hot substrate to the chip are reduced with the use of an active area suspended by silicon support beams (Figure 11.17) or with a very thin membrane for the active area (Fig. 11.18). Compare the macroscopic hot plate sensor shown in Figure 11.10 suspended by wire leads. The first microfabricated hot plate and membrane structures for chemical sensor applications appear to be those presented in 1986 by Grisel and Demarne.[51] Since then, the membrane device in Fig. 11.18 was developed at University of Michigan (UM) by K. Wise and co-workers,[52] and the suspended hot plate in Fig. 11.17 was developed at the National Institute of Standards and Technology (NIST) by S. Semancik and co-workers. [53] Technically, the main difference between the membrane and hot plate is whether the wafer is processed only on the front (hot plate), or is additionally etched from the back (membrane).

The NIST micro hot plate is shown in Figure 11.17. The element is small (~ 60 mm) and has impressive thermal response times (2–5 ms). The thermal efficiency of this device is good, with a temperature rise of 8°C/mW. Moreover, the device was designed to be fabricated using the MOSIS foundry service. The design files are available on the Internet. Standard single-sided wafers were used, and anisotropic etching was used to form pyramidal pits under a suspended "hot plate." The NIST group has made tin oxide chemical sensors from these substrates. An advantage of using the MOSIS foundry is that CMOS electronics can readily be integrated with the sensors on a single die. On the other hand, the design rules of the process place strict limits on the available materials and processing steps. For example, the only metal available through MOSIS

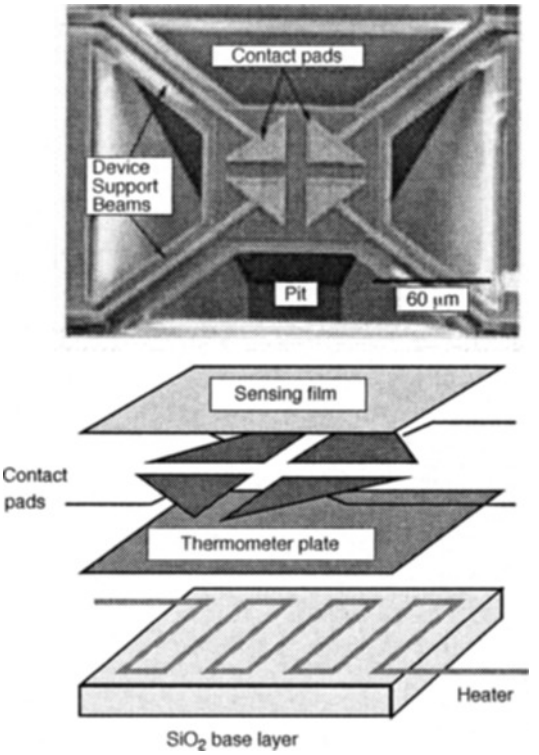


Fig. 11.17. NIST Micro Hot Plate Device. (a) SEM photograph and (b) Schematic diagram showing the different layers in the device. (Reprinted with permission from *Accounts of Chemical Research* 31, 279–287, May 1998) (© 1998 American Chemical Society).

is aluminum, which is not a good choice for a chemical sensor because it is quite reactive. Inert metals like Au and Pt are preferable, and Au is available through other foundries, including the MUMPS (Multiuser MEMS process) foundry at the Microelectronics Center of North Carolina.

The UM membrane device shown in Fig. 11.18 was fabricated using double-sided wafers in a six-mask, in-house process. The die size was 2.8 × 2.8 mm, the total window size was 1 mm^2, and the active area was 250 × 250 mm. The heaters were polysilicon with a thermal coefficient of resistance (TCR) of 1800 ppm/°C, and a temperature uniformity of ± 0.1°C was achieved. The thermal efficiency of the device is good, with 6°C/mW in air and 20°C/mW in vacuum. The metallization was Ir with a Ti adhesion layer. The active sensing layer was Pt/Ti, which was used to detect oxygen.

A problem in making sensors from these micromachined substrates is in applying the sensor coating selectively to the active area. This is especially true if one wants to make array sensors using these structures on a single die. Some materials are amenable to using photoresist and lift-off for patterning, such as the SnO_2 used above. If the coating is to be applied on a single die, as when an outside foundry is used, then it is generally not feasible to use photoresist, given the small size of the die and alignment issues. The UM and NIST groups have cleverly patterned the sensor elements by using the heated substrate as the deposit material by chemical vapor deposition (CVD). This method provides a simple, self-lithographic process for sensor deposition. Furthermore, the electrical properties of the material can be monitored *in situ* during the deposition process.

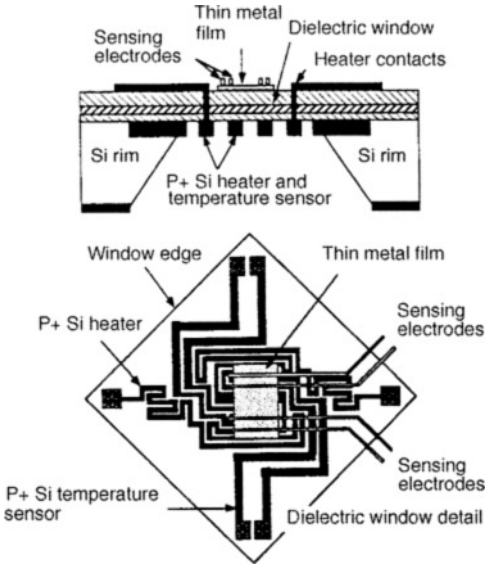


Fig. 11.18. University of Michigan membrane device (© 1994 IEEE).[52]

At Aerospace, we are fabricating a micro hot plate array sensor using the MUMPS foundry service. The goal of this effort is to integrate the phthalocyanine coatings described in section onto a hot plate array sensor in order to detect a range of analyte molecules relevant to the space launch environment. A schematic of the sensor array is shown in Fig. 11.19. The die dimensions are 2.5 × 2.5 mm, and each hot plate is ~250 × 250 mm. There are a total of 16 sensor elements. Each element has gold interdigitated electrodes on the top surface and an embedded polysilicon heater. The device is a membrane similar to the UM design, and the back side will be etched. This etching requires local processing of the back side of the MUMPS die. Processing individual die, as opposed to wafers, is very challenging because of difficulties in handling and masking. (Photolithography is not possible because of alignment issues.) We are currently developing a laser-based process which, if successful, will be useful as a general tool for the in-house processing of die. We hope to report on this process in the near future.

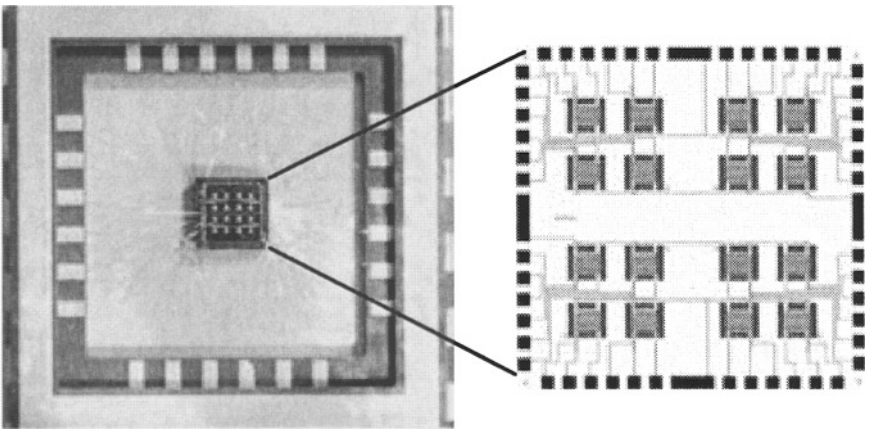


Fig. 11.19. Aerospace micro hot plate array.

11.3 Conclusions

Chemical microsensors comprise very diverse technologies at the intersection of chemistry, materials science, and electronics, and as such it is almost impossible to present an exhaustive overview in a compact and readable form. The author hopes that this chapter has given the reader an overview of the technologies relevant to the detection of space-related gas phase molecules and has provided direction for further investigation by interested readers. The goal is to further the acceptance of chemical microsensors by the space community and to foster more interest and subsequent applications.

Chemical microsensors have been demonstrated to be useful for the detection of space-related molecules and will become more important in the push toward smaller, more power-efficient and smarter sensing systems. This chapter has described their uses on the ground around launch bases, on the Space Shuttle, and on interplanetary probes. As chemical microsensor technologies mature and become more accepted we will undoubtedly see more use of these important devices in future space systems. This will be especially true as progress occurs in the integration of chemical microsensors with analog and digital electronics for sensor control, data acquisition, microprocessing, and wireless communication. In the near future, we will see the development of powerful autonomous sensing systems that provide real-time data on chemical concentrations either on the ground, in spacecraft, or on other planets.

11.4 Acknowledgments

I would like to thank all those who assisted with this review by providing reprints or preprints of their work: Drs. A. Ricco of Sandia, M. Homer of JPL, S. Semancik of NIST, N. Lewis of Caltech, and C. Klimcak of Aerospace.

11.5 References

1. *1993–1994 Threshold Limit Values for Chemical Substances and Physical Agents and Biological Exposure Indices*. American Conference of Governmental and Industrial Hygienists.
2. *Emergency and Continuous Exposure Guidance Levels for Selected Airborne Contaminants*, Vol. 5 (National Research Council. Committee on Toxicology, 1985, dist. 1998); The SPEGL values for N_2H_4 and UDMH were raised to levels shown in 1989, letter from John Doull, M.D., Chairman, NRC Committee on Toxicology.
3. R. K. Kumar. *J. Fire Science* 3, 245 (1985).
4. B. B. Brady, E.W. Fournier, L. R. Martin, and R. B. Cohen, *Stratospheric Ozone Reactive Chemicals Generated by Space Launches Worldwide*. Aerospace Corp. Report no. TR-94(4231)-6 (June 1994).
5. *Spacecraft Maximum Allowable Concentrations for Selected Airborne Contaminants*, Vol. 2 (National Academy Press, Washington, D.C., 1996).
6. M. A. Ryan, M. L. Homer, M. G. Buehler, K. S. Manatt, F. Zee, and J. Graf. "Monitoring the Air Quality in a Closed Chamber Using an Electronic Nose," SAE Paper 972493, SAE 27th International Conference on Environmental Systems. July 1997.
7. A. P. M. Glassford, "Application of the Quartz Crystal Microbalance to Space System Contamination Studies," in *Applications of Piezoelectric Quartz Crystal Microbalance*, edited by C. Lu and A. W. Czanderna (Elsevier, New York, 1994). Chap. 9.
8. D. S. Ballantine, Jr., S. J. Matin, A. J. Ricco, G. C. Frye, H. Wohltjen, R. M. White, and E. T. Zellers, *Acoustic Wave Sensors: Theory, Design and Physico-Chemical Applications* (Academic Press, San Diego, CA, 1997).
9. A. Adamson and A. P. Gast, *Physical Chemistry of Surfaces* (Wiley, New York, 1997).
10. R. C. Hughes and W. K. Schubert. "Thin Films of Pd/Ni Alloys for Detection of High Hydrogen Concentrations," *J. Appl Phys.* 71 (1992).

11. W. M. Moore and P. J. Codella, "Oxidation of Silver Films by Atomic Oxygen," *J. Phys. Chem.* 92, 4421–4426 (1988); V. Matijasevic, E. L. Garwin, and R. H. Hammond, "Atomic Oxygen Detection by a Silver Coated Quartz Deposition Monitor," *Rev. Sci. Instrum.* 61, 1747–1949 (1990).
12. T. A. Jones and B. Bott, "Gas-Induced Conductivity Changes in Metal Phthalocyanines," *Sensors and Actuators* 9, 27–37 (1986).
13. Mistutoshi Hirata and Liangyan Sun, "Characteristics of an Organic Semiconductor Polyaniline Film as a Sensor for NH_3 Gas," *Sensors and Actuators A* 40, 159–163 (1994).
14. D. L. Ellis, M. R. Zakin, L. S. Bernstein, and M. F. Rubner, "Conductive Polymer Films as Ultrasensitive Chemical Sensors for Hydrazine and Monomethylhydrazine Vapor," *Anal. Chem.* 68, 817–822 (1996).
15. N. M. Ratcliffe, "Polypyrrole-based Sensor for Hydrazine and Ammonia," *Analytica Chemica Acta* 239, 257–262 (1990).
16. P. Topart and M. Josowicz, "Characterization of the Interaction Between Poly(pyrrole) Films and Methanol Vapor," *J. Phys. Chem.* 96, 7824–7830 (1992).
17. M. C. Lonergan, E. J. Severin, B. J. Doleman, S. A. Beaber, R. H. Grubbs, and N. S. Lewis, "Array-Based Vapor Sensing Using Chemically Sensitive, Carbon Black-Polymer Resistors," *Chemistry of Materials* 8 (9), 2298–2312 (1996).
18. M. J. Madou and S. R. Morrison, *Chemical Sensing with Solid State Chemical Devices* (Academic Press, 1989).
19. W. O. Camp, R. Lasater, V. Genova, and R. Hume, "Hydrogen Effects on Reliability of GaAs MMICs," *IEEE GaAs IC Symposium* (22–25 October 1989), pp. 203–206.
20. P. Schuesslar and D. Feliciano-Welpe, "The Effects of Hydrogen on Device Reliability and Insights on Preventing These Effects," *Hybrid Circuit Technology* 8 (1), 19–26 (January 1991).
21. L. M. Lechuga, A. Calle, D. Golmayo, P. Tejedor, and F. Briones, "New Hydrogen Sensor Based on a Pt/GaAs Schottky Diode," *J. Electrochem. Soc.* 138 (1), 159–162 (January 1991).
22. W. Auer and H. J. Grabke, "The Kinetics of Hydrogen Adsorption in Palladium (alpha- and beta-phase) and Palladium-Silver-Alloys," *Ber. Bunsenges. Phys. Chem.* 78 (1), 58–67 (January 1974).
23. B. H. Weiller, J. D. Barrie, K. A. Aitchison, and P. D. Chaffee, "Chemical Microsensors for Satellite Applications," *Materials Research Society Symposium Proceedings* 360, 535–540 (1995).
24. R. C. Hughes, D. J. Moreno, M. W. Jenkins, and J. L. Rodriquez, *Solid-State Sensors and Actuators Workshop* (13–16 June 1994, Hilton Head, S.C.); report No. SAND-94-0047C.
25. W. R. Henderson and H. I. Schiff, "A Simple Sensor for the Measurement of Atomic Oxygen Height Profiles in the Upper Atmosphere," *Planet. Space Sci.* 18, 1527–1534 (1970).
26. J. T. Kucera, J. D. Perkins, K. Uwai, J. M. Graybeal, and T. P. Orlando, "Detection of Ozone Using a Silver Coated Quartz Crystal Rate Monitor," *Rev. Sci. Instrum.* 62, 1630–1632 (1991).
27. W. J. Feast, "Conducting Polymers," in *Chemical Sensors*, edited by T. E. Edmonds (Chapman and Hall, New York, 1988), pp. 117–131.
28. J. W. Grate, S. Rose-Pehrsson, and W. Barger, "Langmuir-Blodgett Films of a Nickel Dithiolene Complex on Chemical Microsensors for the Detection of Hydrazine," *Langmuir* 4, 1293–1301 (1988).
29. I. Lundstrom and C. Svensson in *Solid-State Chemical Sensors*, edited by Jiri Janata and Robert J. Huber (Academic Press, Orlando, FL, 1985), pp. 2–61.
30. J. L. Rodriguez, R. C. Hughes, W. T. Corbett, and P. J. McWhorter, "Robust, Wide Range Hydrogen Sensor," *Proceedings of IEEE International Electron Devices Meeting* (1992), pp. 521–524.
31. G. W. Hunter, R L. Bickford, E. D. Jansa, D. B. Makel, C. C. Liu, Q. H. Wu, and W. T. Powers, "Microfabricated Hydrogen Sensor Technology for Aerospace and Commercial Applications," *NASA Technical Memorandum* 106703 (1994).
32. A. J. Rogers in *Sensors, A Comprehensive Survey: Optical Sensor*, Vol. 6, edited by W. Gopel, J. Hesse, and J. N. Zemel (VCH, Weinheim, Germany, 1992).
33. C. Klimcak, G. Radhakrishnan, and B. Jaduszliwer, "A Remote Fiber Optic Dosimeter Network for Detecting Hydrazine Vapor," *Optical Sensors for Environmental and Chemical Process Monitoring*, SPIE Vol. 2367 (9–10 Nov 1994), pp. 80–88; C. Klimcak *et al.*, "Development of a Fiber Optic

Chemical Dosimeter Network for Use in the Remote Detection of Hydrazine Propellant Vapor Leaks at Cape Canaveral AFS," *Chemical, Biochemical, and Environmental Fiber Sensors VI*, SPIE 2293 (26–27 July), pp. 209–219.

34. M. A. Butler and A. J. Ricco, "Chemisorption-Induced Reflectivity Changes in Optically Thin Silver Films," *Appl. Phys. Lett.* 53(16), 1471–1473 (17 October 1988).
35. "Investigating the Surface Chemistry of Mars," *Anal. Chem.* 67, 605–610 (1 October 1995); F. J. Grunthaner, A. J. Ricco, M. A. Butler, A. L. Lane, C. P. McKay, A. P. Zent, R. C. Quinn, B. Murray, H. P. Klein, G. V. Levin, R.W. Terhune, M. L. Homer, A. Ksendzov, and P. Niedermann, "Mars Surface Chemistry Investigated with the Mo_x Probe: a 1-kg Optical Microsensor-based Chemical Analysis Instrument," to be published.
36. R. M. White, "Surface Elastic Waves," *Proceedings of the IEEE* 58(8), 1238–1276 (1970).
37. G. G. Neuberger, *Anal. Chem.* 61, 1559 (1989).
38. H. M. Fog and B. Reitz, "Piezoelectric Crystal Detector for the Monitoring of Ozone in Working Environments," *Anal. Chem.* 57(13), 2634–2638 (1985).
39. E. T. Zellers, N. C. Hassold, R. M. White, S. M. Rapport, "Selective Real-Time Measurement of Styrene Vapor Using a Surface-Acoustic-Wave Sensor with a Regenerable Organoplatinum Coating," *Anal. Chem.* 62(13), 1227–1232 (1990).
40. O. Scheide and J. K. Taylor, "Piezoelectric Sensor for Mercury in Air," *Env. Sci. & Technol.* 8(13), 1097–1099 (13 December 1974).
41. A. D'Amico, A. Palma, E. Verona, "Palladium-Surface Acoustic Wave Interaction for Hydrogen Detection," *Appl. Phys. Lett.*
42.
43. (3), 300–301 (1 August 1982).
44. J. F. Veletino, R. Lade, and R. S. Falconer, "Hydrogen Sulfide Surface Acoustic Wave Gas Detector," *IEEE Ultrason. Ferro. & Freq. Control*, **UFFC-34**(2), 156 (2 March 1987), pp. 157–162.
45. A. Snow and H. Wohltjen, "Poly(ethylene maleate)-Cyclopentadiene: A Model Reactive Polymer-Vapor System for Evaluation of a SAW Microsensor," Anal. *Chem.* 56(8), 1411–1416 (July 1984).
46. M. S. Nieuwenhuizen, A. J. Nederlof, and A. W. Barendsz, "Acoustic Wave Gas Sensor for Nitrogen Dioxide," *Anal. Chem.* 60(3), 230–235 (1988).
47. A. J. Ricco, S. J. Martin, and T. E. Zipperian, "Surface Acoustic Wave Gas Sensor Based on Film Conductivity Changes, *Sensors and Actuators*, 8(4), 319–333 (4 December 1985).
48. J. W. Grate, S. J. Patrash, and M. H. Abraham, "Method of Estimating Polymer Coated Acoustic Wave Vapor Sensor Responses," *Anal. Chem.* 67, 2162–2169 (1995).
49. J. K. Gimzewski, C. Gerber, E. Meyer, and R. R. Schlittler, "Observation of a Chemical Reaction Using a Micromechanical Sensor," *Chem. Phys. Lett.* 217, 589–594 (1994).
50. T. Thundat, G. Y. Chen, R. M. Warmack, D. P. Allison, and E. A. Wachter, "Vapor Detection Using Resonating Microcantilevers," *Anal. Chem.* 67, 519–521 (1995).
51. A. Grisel and V. Demarne, "Fabrication of Integrated Thin Film Semiconductor Gas Sensors," in *Chemical Sensor Technology*, Vol. 2, edited by T. Seiyama, (Kodansha Ltd, Tokyo, 1989), p. 43.
52. N. Najafi, K. D. Wise, and J. W. Schwank, "A Micromachined Ultra-Thin-Film Gas Detector," *IEEE Trans. on Electron Devices* 41 (10) 1770–1777 (1994).
53. (a) R. E. Cavicchi, J. S. Suehle, K. G. Kreider, M. Gaitan, and P. Chaparala, "Fast Temperature Programmed Sensing for Micro-Hotplate Gas Sensors," *IEEE Electron Devices Lett.* 16 (6) (1995); (b) S. Semancik and R. Cavicchi, "Kinetically Controlled Chemical Sensing Using Micromachined Structures," *Accounts of Chemical Research* 31(5), 279–287 (1998).

12

Surface Micromachined Optical Systems

V. M. Bright*

12.1 Introduction

Recent advances in micromachining readily allow implementation of micromachined microopto-electromechanical structures on the order of 100 μm in size. Applications of such miniaturized mirrors, gratings, lenses, and shutters include laser radar imaging, free-space optical communication, optical switching, holographic data storage and retrieval, and adaptive/corrective optics. Subcomponents, such as hinged and rotating structures, are combined with powerful and compact microactuators for positioning and operating the micro-optical devices. The individual micro-optical components and control mechanisms can be arranged into complex systems like fiber-optic switches, optical scanners, and interferometers. The micromachined optical systems, which fit on a single integrated circuit chip, have many potential practical advantages, including integration with control and signal processing electronics, as well as batch fabrication and low cost. In addition, the low individual mass of the micromachined devices leads to superior ruggedness and fast system response time.

12.2 Fabrication Technology

Until recently, research into microelectromechanical systems (MEMS) was restricted to institutions that had access to private fabrication facilities that could meet the specialized demands of micromachining. However, the growing number of commercial micromachining foundries is making the technology widely available. Such commercial services can impose restrictions on the types of devices that can be built. However, the low cost and frequently scheduled fabrication runs of commercial foundry processes reduce the scheduling and financial risk of prototyping useful MEMS solutions. Although micromachined optical devices have been demonstrated using metal electroplating processes[1] and bulk silicon micromachining,[2-4] these technologies are limited to specific applications and the types of components that can be produced. The discussion here will thus focus on micro-optical components produced in a surface micromachining technology, which allows much flexibility in design and integration.

A popular commercial surface micromachining process for MEMS in the United States is the Multi-User MEMS Processes (MUMPs) sponsored by the Defense Advanced Research Projects Agency (DARPA).[5] MUMPs is a three-layer polycrystalline silicon (polysilicon) process. It is intended for prototyping MEMS, using surface-micromachined thin films on a silicon wafer. MUMPs offers three patternable layers of polysilicon and two sacrificial layers of phosphosilicate glass on a base layer of silicon nitride. A top layer of gold is provided as the reflective and/or conductive surface. Table 12.1 identifies the layer thickness for each of the films used in MUMPs, and Fig. 12.1 illustrates the MUMPs layers. The order of the entries in Table 12.1 is consistent with the deposition order of the films on the silicon wafer substrate, with silicon nitride being the first layer. Gold is evaporated onto the device after all other layers have been deposited by low pressure chemical vapor deposition. The polysilicon layers and the <100>-cut silicon substrate are highly doped with phosphorus (approximately 10^{20} atoms-cm^{-3}) to decrease electrical

*Department of Mechanical Engineering, University of Colorado, Boulder, Colorado.

Table 12.1. Structural and Sacrificial Layers Used in MUMPs[5]

Layer Name	Nominal Thickness (μm)
Nitride (silicon nitride)	0.60
Poly-0 (bottom polysilicon layer)	0.50
1st Oxide (sacrificial layer—phosphosilicate glass)	2.00
Poly-1 (middle polysilicon layer)	2.00
2nd Oxide (sacrificial layer—phosphosilicate glass)	0.75
Poly-2 (top polysilicon layer)	1.50
Metal (gold)	0.50

resistance. After construction, the micromachined device is "released" by removing the sacrificial glass layers in a bath of buffered hydrofluoric acid.

After the devices are released, residual material stresses in the gold layer (tensile) and Poly-2 layer (compressive) may cause the reflective surface to curl slightly into a concave shape.[6] Residual material stress varies somewhat after each production run. For example, for the 16th MUMPs production run, typical peak-to-valley curvature for a 100-μm-wide gold on a Poly-2 micromirror was measured as 495 nm.[6] Combining the top and middle polysilicon layers (stacked poly) as support for the gold layer reduced peak-to-valley curvature of a 100-μm-wide micromirror to 140 nm.[6] Peak-to-valley curvature was further reduced to 63 nm by retaining the second oxide layer between the top and middle polysilicon layers (trapped oxide).[6] The second oxide layer is isolated from hydrofluoric acid using a via around the edge of the device to connect the top and middle polysilicon layers and to seal in the oxide layer.[7] Although trapped oxide has the lowest curvature, use of a via forces the gold layer to be indented further from the edge of the device because the surface over the via is not optically flat. The indentation of the gold layer reduces total reflective surface area of a device.

In surface micromachining, the thin-film layers conform closely to the topology of the previously deposited and patterned layers (Fig. 12.1). Unless the designer makes sure a layer is flat by controlling the pattern of the layers beneath it, the induced topology can have detrimental effects on the layer's uniformity and on the effective elastic modulus of mechanical structures. In extreme cases, the topology can trap part of a structure that was intended to move freely. This problem of surface topology can be controlled in more sophisticated surface-micromachining processes, where layers are chemically mechanically polished prior to subsequent layer deposition.[8]

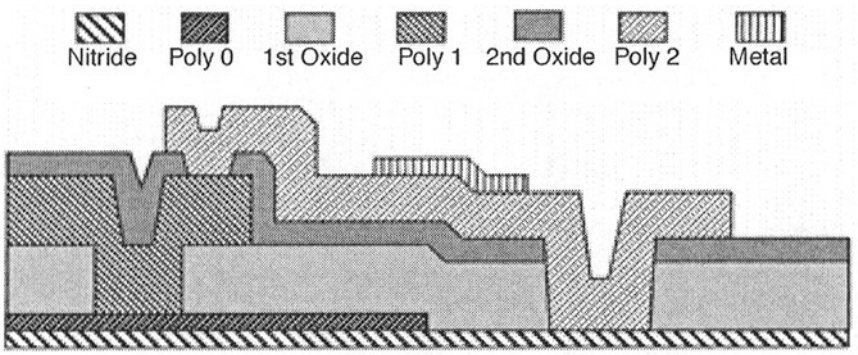


Fig. 12.1. Cross-section of notional MUMPs layout.[5] (Courtesy MCNC MEMS Technology Applications Center, Research Triangle Park, North Carolina.)

12.3 Sliders, Rotating Hubs, Microhinges, and Microlatches

Surface micromachined designs are planar in nature and require some assembly and/or actuation after fabrication. Much of the surface-micromachined optics is based on the use of movable systems. The primary components of movable systems are sliders, rotating hubs, microhinges, and microlatches.

A surface-micromachined slider is shown in Fig. 12.2. The slider consists of a releasable plate and a guide. The releasable plate is formed from the Poly-1 layer in the MUMPs process. The plate is not anchored to the substrate and is free to move. The guide is formed using the Poly-2 layer. One side of the guide projects over the edge of the released plate; while the other side is anchored to the substrate. By placing guides on opposing sides of the plate, the movement of the plate is restricted to the direction defined by the guides.

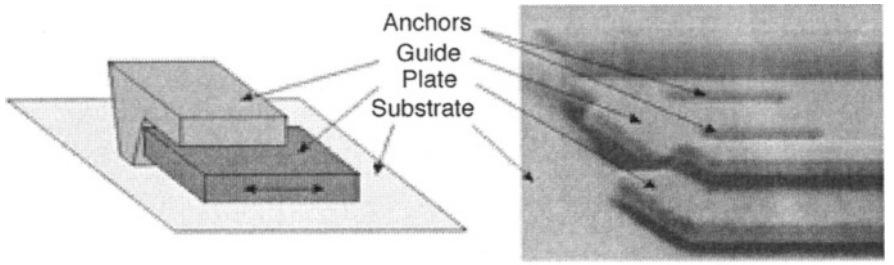


Fig. 12.2. Sliders that allow linear motion of a plate are fabricated using two releasable structural layers. The basic design is shown in the schematic view (a); while a fabricated polysilicon slider is shown in (b).[9]

The construction of a rotating hub is similar to that of a slider, except that the guide is now circular, as shown in Fig. 12.3. The center of the guide is anchored, while the outside edge projects over the rotating plate. A circular hole is cut into the center of the rotating plate in conjunction with the circular anchor of the rotating hub.

Microhinges enable surface-micromachined parts to be rotated out of the plane of the substrate, as illustrated in Fig. 12.4. This allows fabrication of flip-up components such as plates, mirrors, gratings, and lenses. There are three types of microhinges: the substrate hinge, the floating

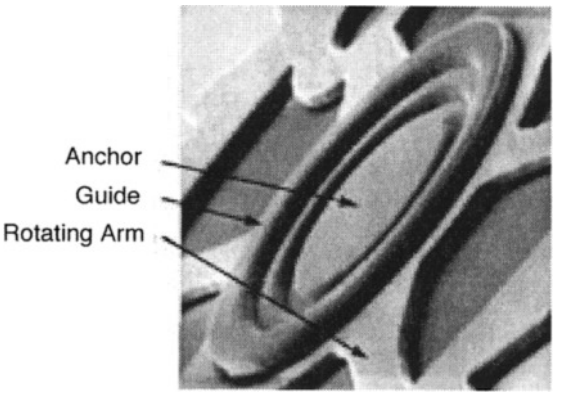


Fig. 12.3. Scanning electron micrograph of a rotating hub.[9]

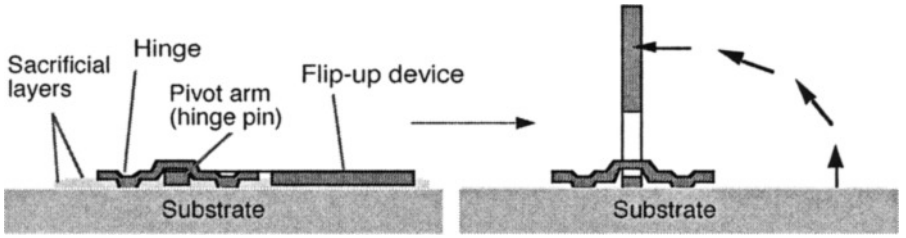


Fig. 12.4. Assembly of flip-up optical components using microhinges (after Ref. 10).

substrate hinge, and the scissors hinge. The substrate and scissors hinges were originally designed and fabricated by Pister *et al.*;[10,11] while the floating substrate hinge is a modification of the substrate hinge. All three hinge designs require two releasable structural layers for their fabrication. A substrate hinge, shown in Fig. 12.5, is used to hinge released structures to substrate. This hinge consists of a pivot arm and a staple. The pivot arm is fabricated from the Poly-1 layer. The structure that is to be flipped up off of the substrate is connected at the ends of the pivot arm. The staple is fabricated from the Poly-2 layer, and forms a bridge over the pivot arm. The staple, which is anchored to the substrate, is not connected to the pivot arm and allows free rotation of the pivot arm. A floating substrate hinge allows the fabrication of movable flip-up structures. This hinge is fabricated by replacing the anchors of the substrate hinge staple with connections to a movable Poly-1 plate (see Fig. 12.6).

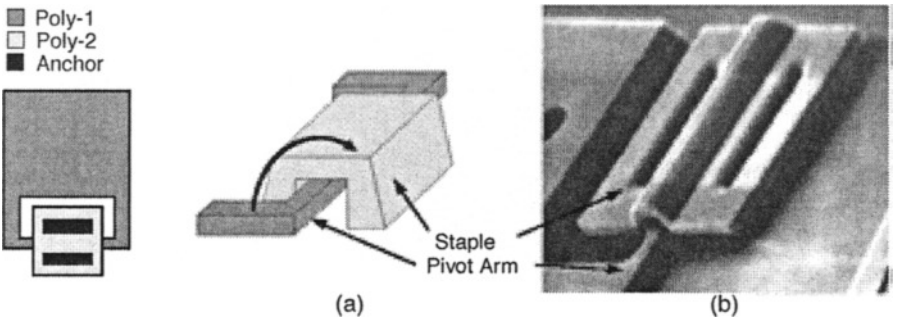


Fig. 12.5. Schematic (a) and scanning electron micrograph of a substrate hinge (b).[9]

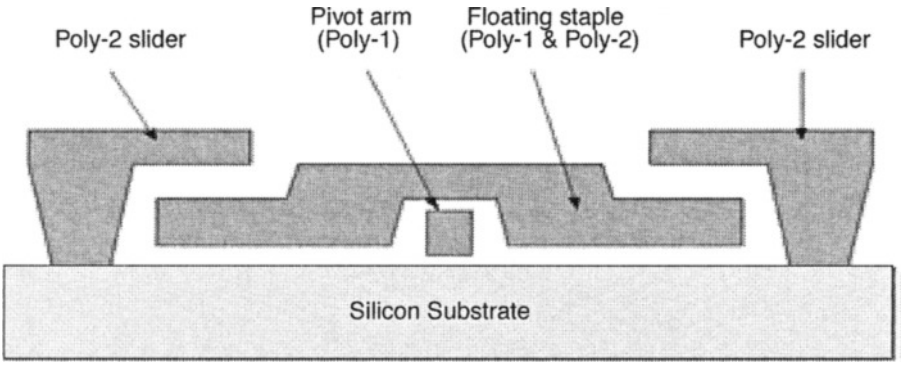


Fig. 12.6. Schematic diagram of a floating substrate hinge.[9]

The floating staple is incorporated into a rotating hub or a slider, which allows the hinge to both rotate and slide across the substrate. A scissors hinge is fabricated by interlocking fingers of two structural layers, as shown in Fig. 12.7. This hinge allows two plates to be connected while still allowing them to pivot at the connection.

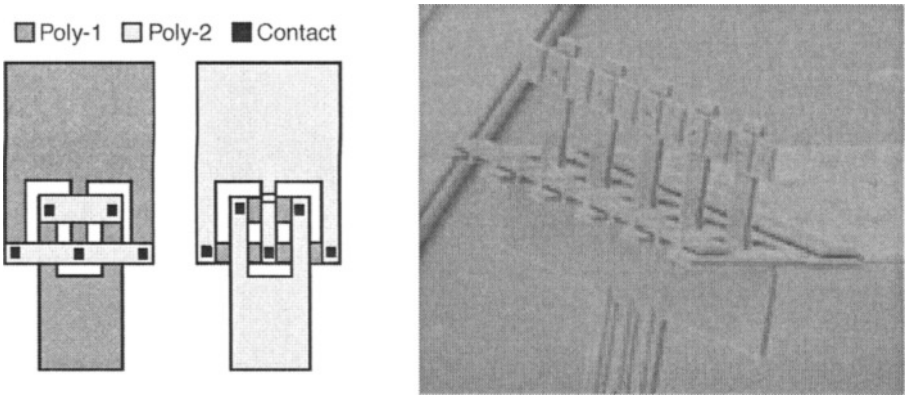


Fig. 12.7. Scissors hinges are used to connect two released plates still allowing them to pivot.[9,10]

Microlatches enable flip-up structures to be latched into specific positions. Several different locking mechanisms exist. One approach is to use a second flip-up plate positioned at a 90-deg angle with respect to the first plate (Fig. 12.8).[10] The lock works by initially flipping the first plate into the desired position, then flipping up the second plate. A notch in the second plate engages with the first plate, locking both plates into position. The disadvantage of this design is that both plates must be flipped up and interlocked into position. Another approach of locking flip-up components into position is to use a self-engaging locking mechanism (microlatch),[9,12] which is shown in Fig. 12.9. When the flip-up plate is raised into position, the latch moves along the surface of the plate until it drops into position in the locking slot. The locking slot has a wide opening at the top and a narrow opening at the bottom. The latch end is tapered into a triangular shape. Below the triangular tip, slots are cut into both sides of the latch, corresponding to the size of the

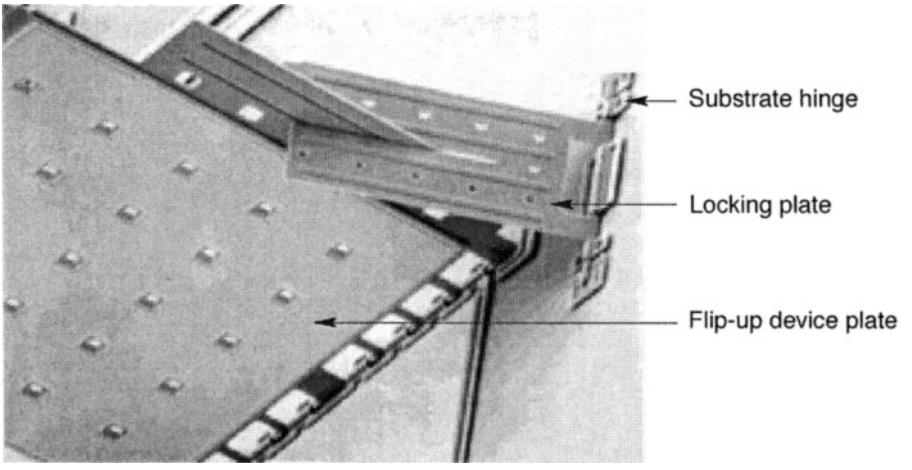


Fig. 12.8. Locking of two flip-up plates positioned at a 90-deg angle with respect to each other.[13]

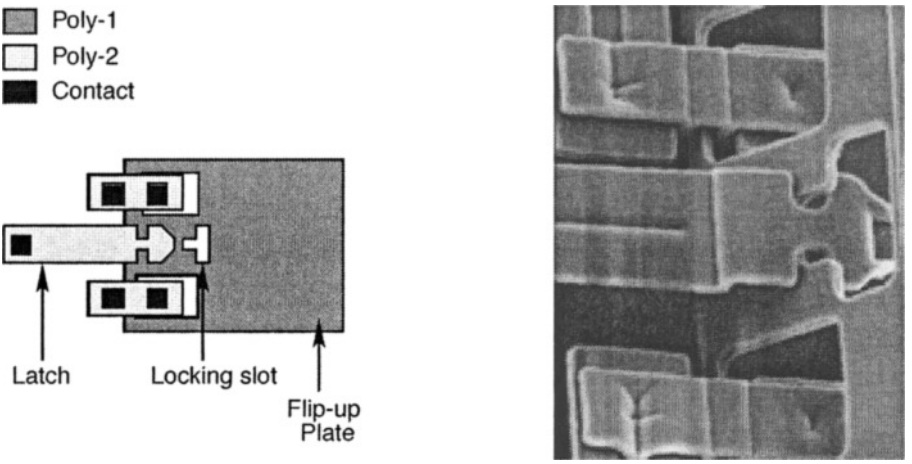


Fig. 12.9. A self-engaging locking mechanism (microlatch).[9]

narrow bottom of the lock opening in the flip-up plate. As the hinged plate is rotated off the substrate, the latch slides into the lock opening, and the slots in the latch engage the narrow bottom of the lock opening in the plate. By anchoring the latch to a floating plate instead of the substrate, this type of lock can be easily implemented for floating structures. The advantage of the self-engaging lock is that it is only necessary to flip up one plate. The simplicity of this design is important in self-assembled systems.[14]

12.4 Actuators

Complex optical MEMS may include moving components such as shutters and mirrors. To provide motion of these structures by electrical means, a variety of microactuators have been developed. The most common modes of actuation are electrostatic attraction and thermal expansion.

A common electrostatic actuator is the parallel plate actuator shown in Fig. 12.10. This device takes advantage of the planar nature of the surface-micromachining process, which easily forms parallel-plate capacitor structures. The upper plate of the structure can be metallized to create a moving micromirror. The list of potential applications for a moving micromirror includes projection display devices, printers, optical switching networks, maskless photolithography, and adaptive optics. Each application imposes a unique set of micromirror requirements. Important requirements have to do with speed of operation, desired optical modulation, and optical efficiency. Driven by commercial applications in print and display technology, torsion micromirrors

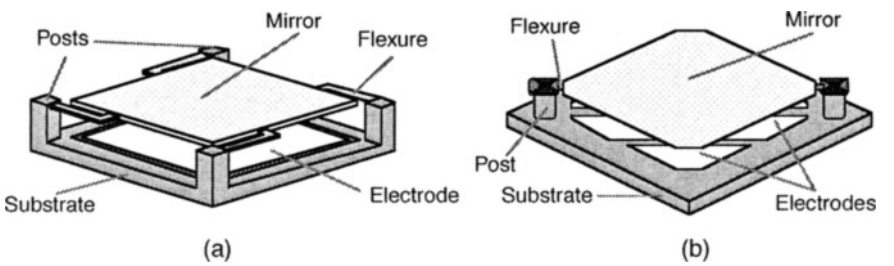


Fig. 12.10. Piston (a) and torsional (b) parallel plate actuators/mirrors.[15] The suspended mirror plate moves when a voltage is applied between the plate and underlying address electrode(s).

dominate the technical literature.[16] Adaptive optical systems require amplitude modulation (torsion mirror motion) for stabilization or tracking and phase modulation (piston mirror motion) to compensate for higher order aberrations.[17,18]

When a voltage is applied between the two plates (a mirror and the underlying address electrode), an attractive electrostatic force is developed, which is balanced by the restoring mechanical force of the flexures, as illustrated in Fig. 12.11. A detailed analytical model has been developed to compute the electrostatic force on an electrostatically driven piston actuator.[19] The model incorporates the effects of cross-talk from neighboring devices, ambient temperature, fringing electric fields, and deformations of the actuator plate surface. If thermal and fringing effects are ignored, which are minor, and if deformation of the actuator surface is ignored, the address voltage versus displacement model simplifies to[19,20]

$$V = (z_0 - d_f)\sqrt{\frac{2k_{sp}d_f}{\varepsilon_0 A}} \tag{12.1}$$

where V is the positive address voltage, z_0 is the resting gap between the actuator plate and the addressing electrode, d_f is the desired downward actuator displacement from this resting position, ε_0 is the free space dielectric constant, and A is the addressing electrode area. k_{sp} is a total spring constant, which accounts for the number, geometry, and material of the actuator flexures.[19,20]

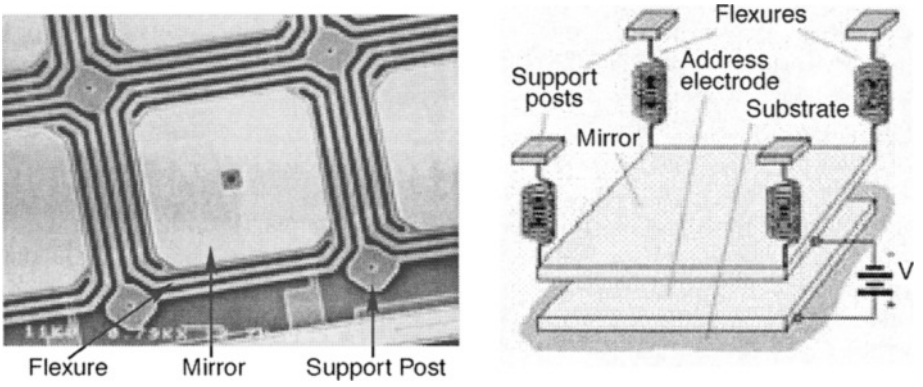


Fig. 12.11. Micrograph and representation of the square piston micromirror.[19]

Motion of a parallel plate actuator is limited by the gap separating the suspended plate from the address electrode. In addition, electrostatic parallel plate actuators exhibit a "snap-through instability" behavior when the deflection corresponds to $z_0/3$.[21] After an actuator is deflected approximately one-third of the separation gap between the actuator plate and the electrode, the actuator plate quickly snaps down toward the substrate. This phenomenon results from using a linear force (the supporting flexure) to counter a nonlinear force (electrostatic attraction). Similar parallel plate electrostatic actuators include the gap-closing actuators[22] and the scratch-drive actuators.[23] However, these devices require high voltages to actuate (greater than 50 V).

Another common type of electrostatic actuator is the comb actuator (Fig. 12.12), often operated at resonance.[24] The device has one or two interdigitated electrodes and a suspended platform. Applying a potential between the electrode and the platform results in an attractive force and corresponding lateral motion of the platform. Unfortunately, the force is relatively small (less than a micronewton). To achieve large displacements of the platform, it is necessary to apply a large voltage or to operate the device in a resonant mode by applying an ac drive signal to the drive electrode.[25,26]

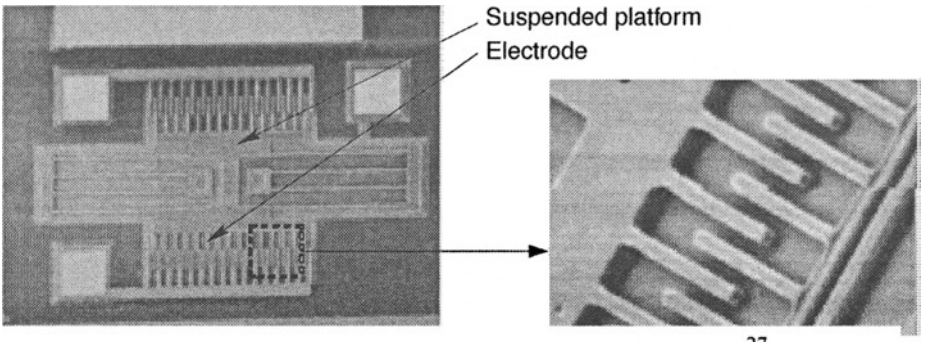


Fig. 12.12. Scanning electron micrograph of a comb actuator.[27]

A polysilicon thermal actuator represents a complementary class of actuators. A thermal actuator uses ohmic heating to generate thermal expansion and movement. The basic designs are shown in Figs. 12.13 and 12.14. When current is applied to the thermal actuator, it causes the hot arm to expand more rapidly than the cold arm. Removing the current from the thermal actuator returns it to its normal state. Actuators fabricated in the MUMPs process are capable of providing deflections of greater than 10 μm, while typically requiring drive voltages less than 5 V. The actuators can take on a wide variety of geometries and can be fabricated in any MEMS process that has at least one releasable, current-carrying layer. One of the benefits of constructing thermal actuators in the MUMPs process is that the polysilicon layers are conductive. As a result, the actuators can be designed to operate at voltages and currents compatible with complementary metal oxide semiconductor (CMOS) electronics.[28]

The thermal actuators can be operated in two modes. In the basic mode, current is passed through the actuator from anchor to anchor, and the higher current density in the narrower "hot" arm causes it to heat and expand more than the wider "cold" arm. The arms are joined at the free end, which forces the actuator tip to move in an arcing motion. Another mode of operation is to create a permanent deformation in the hot arm of the actuator, which is accomplished by applying enough current to cause plastic deformation of the polysilicon. In general, the amount of current necessary to create a permanent deformation is slightly higher (within 10%) than the current needed to generate the maximum tip deflection. When the current is removed, the actuator is permanently "back-bent" from its original position, due to bowing or buckling of the hot arm.

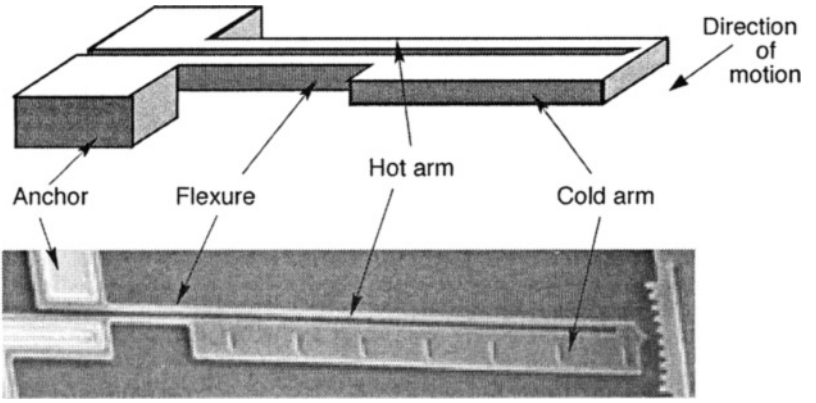


Fig. 12.13. A lateral thermal actuator.[29,30]

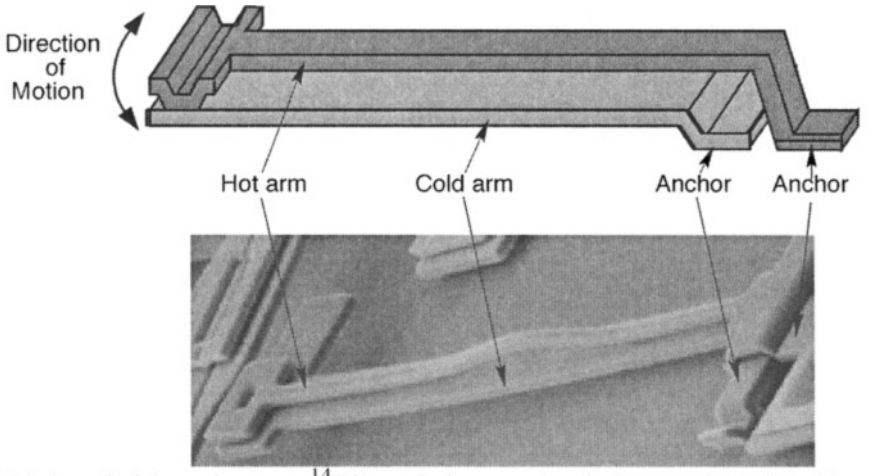


Fig. 12.14. A vertical thermal actuator.[14] The vertical actuator has the hot arm above the cold arm so it will deflect downward when powered and upward when back-bent.

The amount of deformation or back-bending depends on the amount of over-current that is applied. After back-bending, the actuator can be operated in the basic mode. Back-bending is particularly useful for one-time positioning of actuators and as a tool in automated-assembly of complex devices.[13,14,30]

Thermal actuators can be arrayed to obtain higher forces. For fabrication processes with only one or two releasable layers, a compliant structure can be attached perpendicularly at the free ends of parallel thermal actuators, such as a yoke with flexures to accommodate the arc-like motion of the actuators. The yoke cancels out arcing and expanding motions, producing a purely linear motion of the yoke, in addition to combining the forces of the actuators. Each actuator has its own attaching flexure, so adding more actuators has little effect on the overall array stiffness; each additional actuator adds a fixed increment of force. There is no discernible upper limit on the number of actuators that can be arrayed together, so achieving extremely high forces is possible with these arrays (4.8 μN per individual actuator is typical).[31] Arrays of up to 60 actuators have been successfully operated.[13,31] An array of 10 actuators with a compliant yoke is shown in Fig. 12.15.

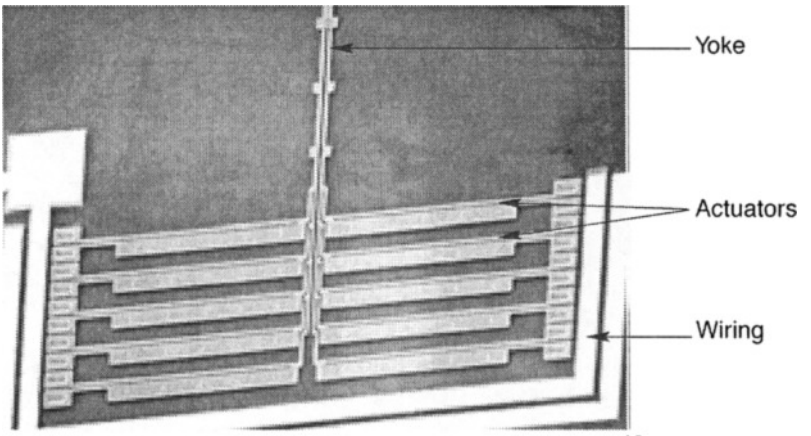


Fig. 12.15. Ten-element thermal actuator array.[13]

Thermal actuator arrays can be built into rotary or linear stepper motors, which can be used to position optical components. Figure 12.16 shows a rotary stepper motor. A main array of 10 thermal actuators is attached to a drive pawl, which is engaged and disengaged from the rotor by a second array of three pusher actuators set at 90 deg to the main array. This arrangement of actuators leaves much of the rotor edge free to connect to other mechanical structures such as gear trains. The same two-array actuation scheme can be applied to linear motors by replacing the rotor with a sliding, toothed rack (Fig. 12.17). In the motor of Fig. 12.17, the rack can be moved 200 μm between stops. However, for component positioning or automated-assembly applications, the rack could be as long as needed and could be incorporated into the device being positioned.

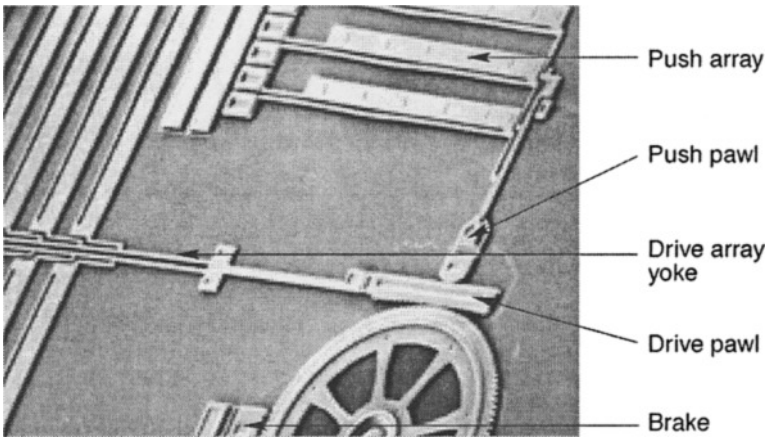


Fig. 12.16. A rotary stepper motor.[13] The three pusher actuators that engage the drive pawl and the 10 actuators that move the rotor are 240 μm long, 21.5 μm wide, and 2 μm thick. The rotor is 200 μm in diameter. The teeth on the pawls and rotor are 3.5 μm thick. The motor successfully operated at 375 rpm.

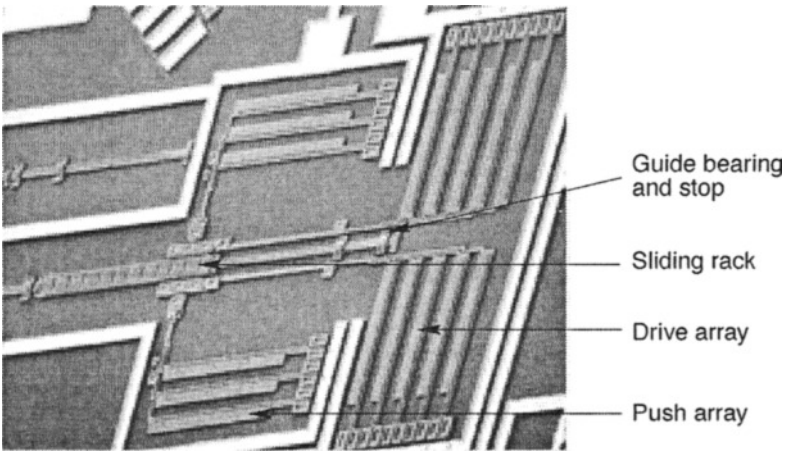


Fig. 12.17. A linear stepper motor with toothed rack driven from both sides by arrays of 240 μm long, 2 μm thick actuators.[13] The rack on this motor can slide 200 μm between stops. Rack and pawls are 3.5 μm thick.

12.5 Micromirrors for Adaptive Optics

Controlling aberrations in optical systems is one of the major problems of optical imaging and beam projection systems. One approach to improving the performance of aberrated systems is to apply the concept of astronomical adaptive optics[32] to correct the aberrations. The essential elements of astronomical adaptive optics are (1) sensing the aberrations using some type of wave front sensor and (2) mechanically correcting for the aberration using a deformable mirror. Deformable mirrors based on piezoelectric actuators and a continuous face-sheet design have been most widely used in astronomy.[33,34] MEMS deformable mirrors offer an alternative low-cost technology for aberration correction.[2–4,17,19,20,35–39]

MEMS deformable mirrors for adaptive optics offer potential of reduced corrector element size; exceptional element uniformity; and drastically reduced mirror system size, weight, power dissipation, and cost. Small deflections required for optical phase modulations are consistent with the dimensions of MEMS processes. MEMS deformable mirrors can be categorized as continuous membrane mirrors or arrays of piston elements. Continuous membranes offer high optical efficiency with limited aberration correction modes. Piston-only arrays offer the potential of high-order aberration correction but suffer from reduced optical efficiency. The segmented, reflective surfaces give rise to larger diffraction effects than those provided by continuous face-sheet deformable mirrors. However, segmented micromirror arrays are attractive because of their low fabrication cost, low drive voltages, and simple control algorithms.

For any micromirror array, intended application and fabrication process constraints guide design decisions. An ideal array of piston micromirrors would have perfectly flat elements with 100% reflectivity, with identical deflection versus voltage response. The mirrors would completely cover the array surface with no gaps between them. Address wiring could be extended to any depth without impacting micromirror properties. Reference 16 describes recent attempts at optical aberration correction using a segmented hexagonal micromirror array. The micromirror array used in Ref. 16 and shown in Fig. 12.18 is not ideal, as a result of both the MUMPs fabrication process and the originally intended application. As this design was originally targeted at beam-forming[40] and beam-steering[41] applications, some design features, such as metallization of the static support structures, are not optimal for aberration correction applications.

The polysilicon layers used in the MUMPs process are relatively thick, so several design features are incorporated to reduce the overall mechanical spring constant of the micromirror element. The mirror, flexures, and support structures are all fabricated in the thinner (1.5-μm-thick) Poly-2 layer. A hexagonal design was chosen for micromirror elements because it allows the minimum number of flexures for a purely vertical piston-like motion. Hexagons also pack efficiently, maximizing the active surface area of the array. The individual mirror elements are 80 μm across (flat to flat), with each mirror 46.9 μm on a side. The large mirror element size increases the active optical area of the array, and, through longer flexure length, reduces flexure stiffness. This reduced stiffness, in turn, lowers the operating voltage (refer to Eq. [12.1]). Each of the three flexures runs along two sides of the hexagonal micromirror, providing the longest flexure length possible without lateral overlap. Longer flexures can be made by wrapping around more sides or by folding the flexure back along the same two sides of a mirror. However, these approaches decrease the useful reflective surface of the array and add more diffracting edges, thus increasing stray light generation. The design rules for the MUMPs process set the minimum dimension at 2 μm for lines and gaps designed in Poly-2. In the mirror design of Fig. 12.18, the flexures are 2 μm wide, and the gaps on either side of the flexures are 3.2 μm wide.

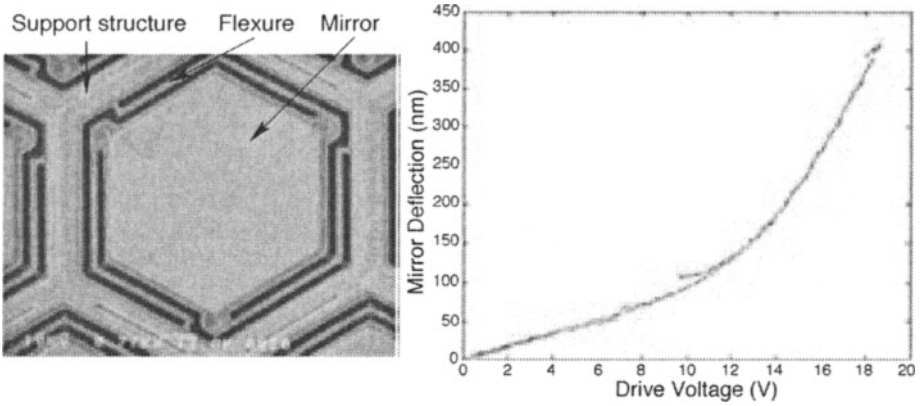


Fig. 12.18. Scanning electron micrograph and deflection vs voltage data for a mirror element in a hexagonal 127-micromirror array.[17] Three sets of raw data are shown as measured with a laser interferometer.[19,39]

Array elements are individually addressable via the Poly-0 wires. The address wires run under the support structure so that the topology induced in the overlying Poly-2 layer does not affect the flatness of the active mirror surfaces or the stiffness of the flexures. A hexagonal array of 127 mirrors requires space for up to three Poly-0 wires between mirrors to address such elements. The final design of the support frame around the hexagonal mirrors allowed space for up to three wires between mirrors, for a mirror edge-to-edge spacing of 36 μm, and a mirror center-to-center distance of 117 μm. This results in an active mirrored area of 48% for the 127-element array. Past the edge of the array, the Poly-0 address lines are extended as Poly-0/Poly-2/gold wires to provide a low resistance path to the bond pads.[17]

Device die were mounted in a 144-pin grid array package and were tested in an adaptive optics system experiment to demonstrate optical aberration correction using a segmented micromirror array.[17] Because of the periodic structure of the array, a large amount of optical power was shifted into the first and higher diffraction orders. To study the effects of aberration control using a segmented micromirror array, only the intensity of the zeroth diffracted order of the far-field diffraction pattern due to the reflected wave front was examined. Laboratory experiments were performed for reducing the effects of an artificially introduced quadratic aberration on a propagating laser beam. It was shown that the effects of correcting aberration with a segmented micromirror array are to decrease the strength of the side lobes of the far-field diffraction pattern and to increase the maximum value of the zeroth diffracted order,[17] thus demonstrating capability for optical image correction and beam shaping.

To overcome the low fill factor and interference with the static background structure in a segmented micromirror array, a refractive lenslet array can be placed directly over the micromirrors, as shown in Fig. 12.19.[18,42] Use of diffractive lenslets and micromirrors in an integrated assembly has also been proposed.[43] The lenslet array focuses incident light onto the center of the reflective surface of the micromirrors, greatly improving optical efficiency of the hybrid correcting element. For mirror deflections much smaller than the lenslet focal length, the lenslet/micromirror combination behaves as a phase-only modulating element. In addition to improved optical efficiency, the micromirror-lenslet system reduces background interference effects by eliminating illumination of the nondeflecting parts of the array. Using a lenslet/micromirror combination, one can correct rather severe quadratic aberrations. For example, correction of a spherical aberration with a 0.35-m radius of curvature at 632.8-nm wavelength was achieved with the micromirror device shown in Fig. 12.20.[18]

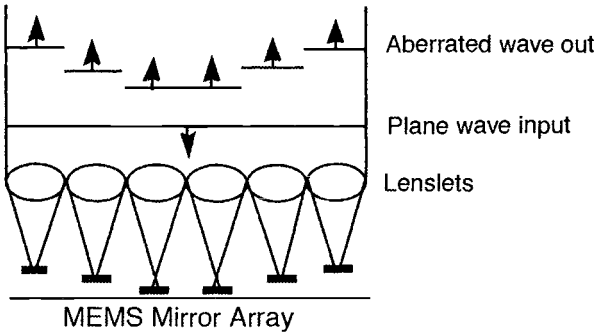


Fig. 12.19. MEMS mirrors/lenslet concept.[18,42] Induced aberration is shown, but aberration correction is also possible with this arrangement.

The micromirror array in Fig. 12.20 is composed of 128 individually controllable micromirror elements on a 12 × 12 square grid (16 elements—4 in each corner of the array were inactive). The reflective gold mirror surface is 60 μm diam. The 203-μm center-to-center spacing of the mirrors in the array was designed to match the refractive lenslet array. Each mirror is controlled by a static electric voltage applied between the movable mirror plate and an underlying address electrode, not visible in Fig.12.20.

An aberrating lens was chosen to provide a 0.35-m radius of curvature on the incident wave at the lenslet array. Control voltages to generate radii of curvature on the micromirror array of 0.5 to 1.0 m in 0.1-m steps were applied, and the far-field diffraction patterns were recorded, as shown in Fig. 12.21.[18] The images in Fig. 12.21 are overexposed to display the off-axis light distribution more clearly. Because the lenslet array is composed of small, square periodic structures, the diffraction patterns are composed of a central diffracted order and many higher diffracted orders, which are distributed periodically on a "rectangular" grid in the observation plane. Best performance was obtained with a micromirror array radius of 0.7-m curvature, which is exactly twice the incident aberration. The factor of 2 is due to the "round-trip" nature of aberration correction using reflective devices.[18] The best aberration correction case (0.7 m) exhibited a peak intensity that was 32% of the peak intensity of the unaberrated case. A computer model predicted the peak intensity of the corrected aberration case at 34% of the unaberrated case.[42] Thus, almost complete correction of a rather severe quadratic aberration was achieved, indicating that diffraction-limited performance is obtainable with the lenslet/micromirror approach.

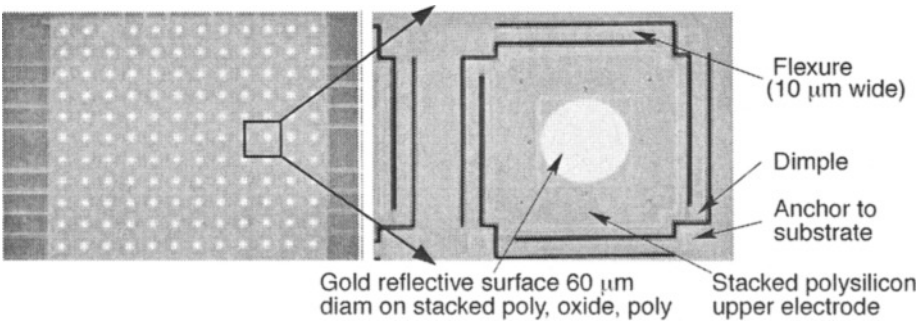


Fig. 12.20. A 128 individually addressed micromirror array on 203-μm grid to match lenslet array.[18] Mirror deflection of 316 nm is achieved at 13.8 V.

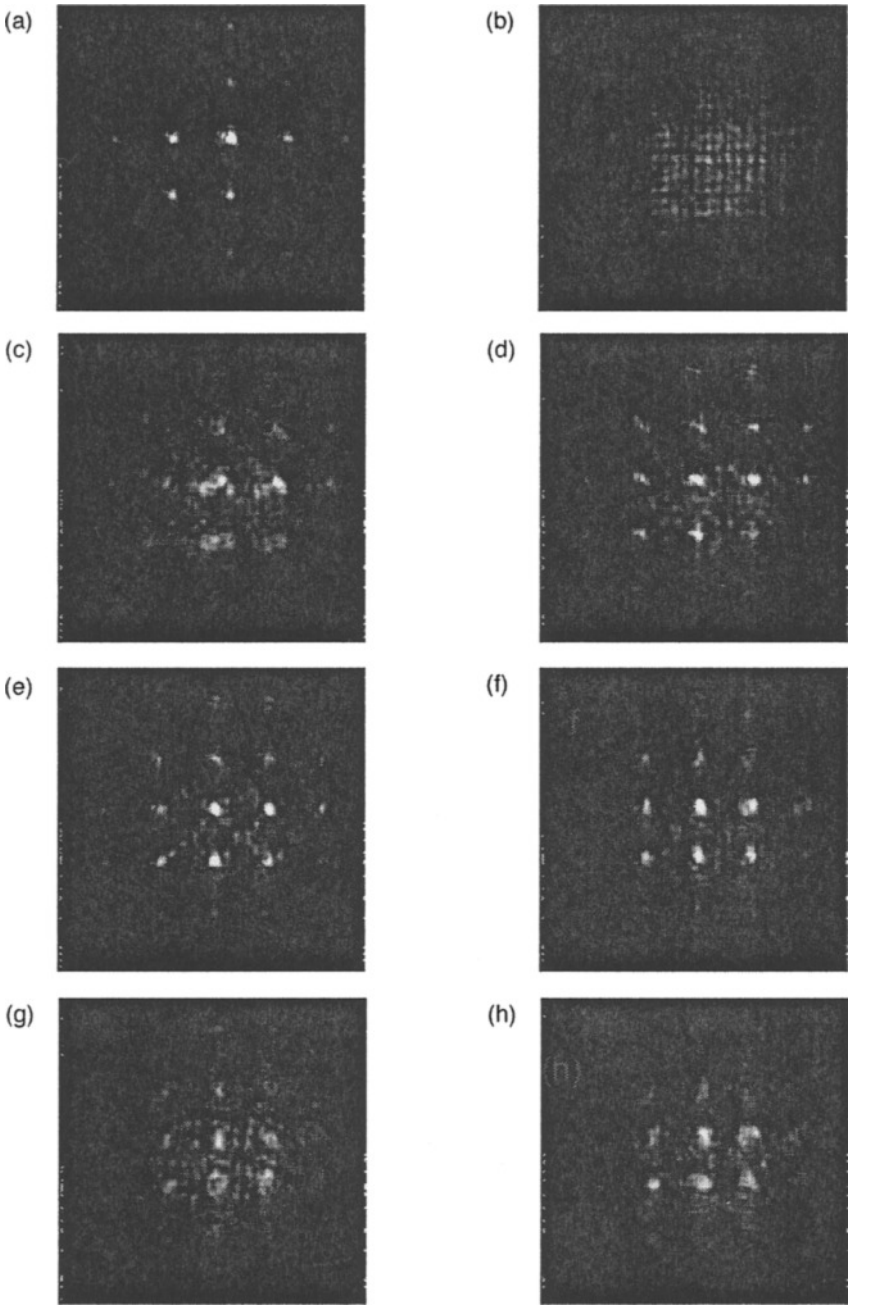


Fig. 12.21. Far-field diffraction patterns for the lenslet/micromirror combination:[18] (a) plane wave on undeflected micromirror array, (b) aberrated wave on undeflected micromirror array, and (c)-(h) aberrated wave on deflected micromirror array with radius of curvature increasing from 0.5 m to 1.0 m in steps of 0.1 m; (e) is the best aberration correction obtained—compare with (a)—with micromirror array radius of curvature of 0.7 m.

The micromirror/lenslet arrangement has been also successfully tested for optical beam steering (Fig. 12.22).[42] The results shown in Figs. 12.21 and 12.22 are the only known experimental demonstrations of aberration control and beam steering with a micromirror/lenslet combination.

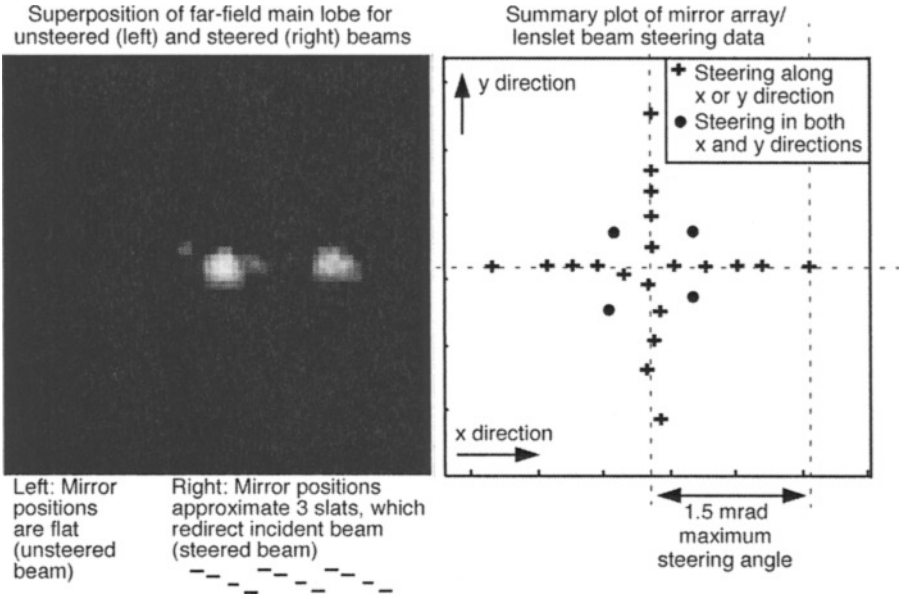


Fig. 12.22. Micromirror array/lenslet beam-steering data.[42]

12.6 Micromirrors for Beam Steering

Beam-steering micromirrors find applications in optical beam-steering systems, corner cube reflectors, optical scanners, and optical couplers. For example, Fig. 12.23 shows a 250-μm-sq mirror flipped to 45 deg off the substrate using vertical thermal actuators.[44] The principal design purpose for this device is coupling of normally incident optical signals onto the substrate plane. After fabrication and the release etch, the parallel-wired vertical thermal actuators are simultaneously back-bent, flipping the mirror plate to an angle of 45 deg from the substrate. Return flexures

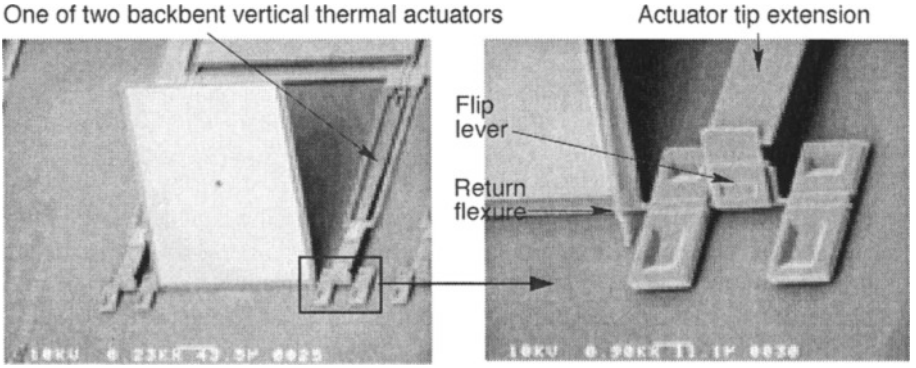


Fig. 12.23. A 250-μm-sq mirror flipped to 45-deg angle with the substrate using two vertical thermal actuators, and detail of the hinge/return flexure mechanism.[44]

prevent the mirror from flipping too far and keep the mirror plate engaged with the actuators through a hinge/lever arrangement. Subsequent drive of the actuators permits adjustment of the mirror's angular position. The micromirror can be scanned from 0–45 deg with a 0–8 V signal.

Figure 12.24 shows a beam-steering mirror that employs vertical thermal actuators for setup.[44] Here, a hexagonal plate 200 μm from corner to corner is raised from a nominal as-fabricated position of 2 μm off the substrate to 10 μm high, using three back-bent actuators driven in parallel. The metallized mirror surface is 100 μm. After setup, the mirror plate is grounded, and three electrodes under the hexagonal plate are used to deflect the plate electrostatically. The control electrodes do not extend under the metallized mirror surface, in order to avoid induced topology in the reflective surface. By proper control of the electrode voltages, piston and/or angular deflection are possible. Increasing the electrostatic gap by a factor of 5 also increases the maximum controllable steering angle by a factor 5. Using one-third the gap distance as a limitation (to avoid snap-through) and a center-to-plate corner dimension of 100 μm results in a maximum steering angle of 1.91 deg. The resonant frequency of this mirror is 6.7 kHz.

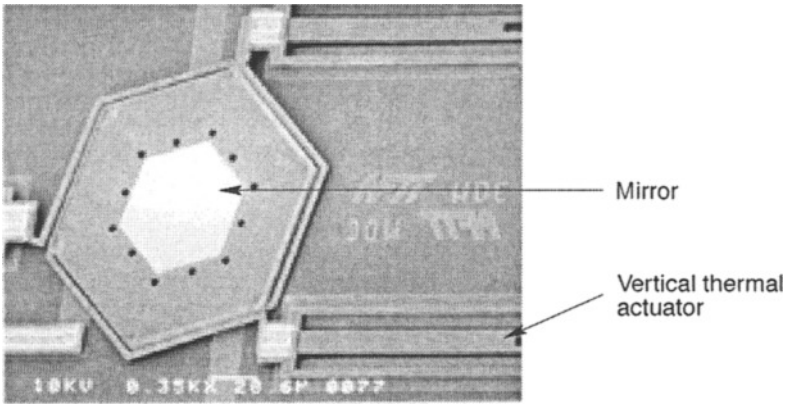


Fig. 12.24. An electrostatically controlled beam-steering mirror positioned 10 μm above substrate by three vertical thermal actuators.[44]

Where minimum power dissipation and high-frequency response are less important system requirements than maximum steering angle, the large deflection of vertical thermal actuators can be directly exploited.[44] Figure 12.25 depicts a thermally actuated beam-steering mirror. The device consists of an 80-μm-diam gold mirror attached to three vertical thermal actuators. After fabrication and release, the three individually wired actuators are driven in parallel so that they will be back-bent equally. Back-bending raises the mirror plate 10 μm off the substrate. Independent control of the actuators then permits angle and/or piston modulation of an optical beam. A maximum steering angle of 7.16 deg was measured for this device.[44] Maximum operating frequency of thermally actuated devices is limited by the ability of the structure to dissipate heat. The operating frequency for this device is several hundred hertz.

Optical systems with arrays of micromirrors often employ bistable actuation to simplify the control scheme. Figure 12.26 shows a portion of a linear array of 10 piston electrostatic micromirrors, positioned for bistable operation by vertical thermal actuators.[44] When back-bent, the vertical thermal actuators pull the mirror plate up against the stops. After back-bending, the thermal actuators can be operated to vary the bistable deflection distance from 0 to 2.0 μm. For this particular design, π modulation of a normally incident HeNe (632.8-nm wavelength) beam requires only 5 V. Postfabrication electrical adjustment of the bistable deflection distance allows the

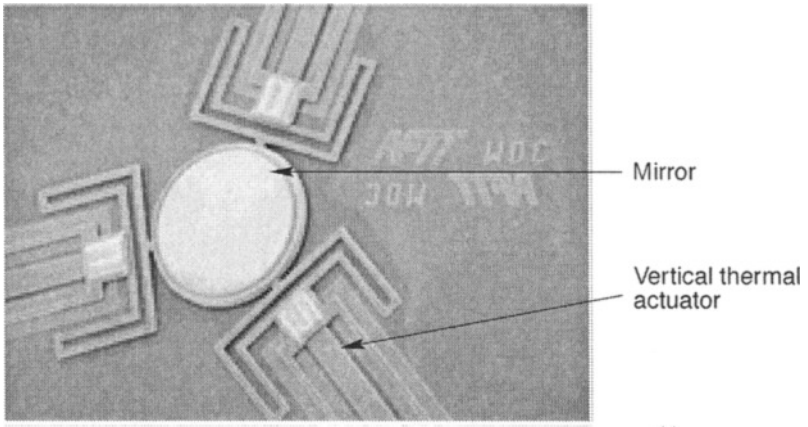


Fig. 12.25. Thermally set up and actuated beam-steering mirror.[44]

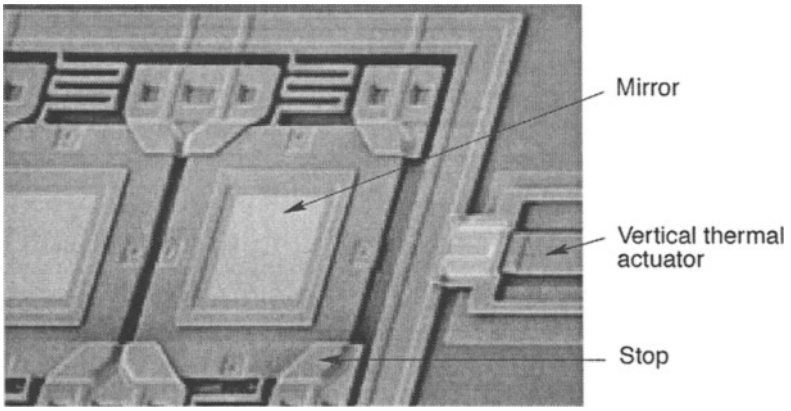


Fig. 12.26. Portion of 10-element linear array of adjustable bistable piston mirrors.[44]

same device to be used in phase modulation applications with different operating wavelengths, permits fine tuning, and could conceivably be used to vary the bistable modulation "on the fly."[44]

Vertical thermal actuators can also be used to position bistable tilting mirrors for amplitude modulation applications. A 100-μm-square mirror shown in Fig. 12.27 illustrates the concept.[44] A hinge mechanism facilitates bistable operation for large steering angles. After release, this mirror has a beam-switching angle of only 1.14 deg. Using the back-bent vertical thermal actuators to lift the mirror plate 10 μm off the substrate yields a switch angle of 22.6 deg.[42] The maximum switching angle is controlled at setup by the degree of actuator back-bending. Setup tolerances on the order of 1% to 2% are feasible. This degree of repeatability is adequate for many beam-steering applications. Electrical drive of the back-bent actuators permits adjustment of the steering angle from 0 deg to the maximum switching angle. This adjustment may be a part of the system setup or tuning, or an integral part of the optical modulation scheme.

Systems of multiple-switching mirrors have also been fabricated. An example is shown in Fig. 12.28. In this system, four vertical thermal actuators driven in parallel raise a polysilicon frame in which four 100 × 400 μm mirrored slats are supported by hinges.[44] The electrostatic drive electrodes for each slat are wired together so that the slats move in unison.

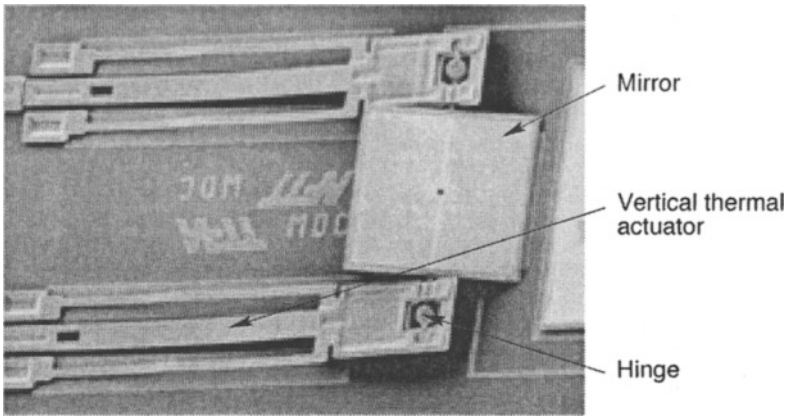


Fig. 12.27. Electrostatically controlled bistable switching mirror setup by vertical thermal actuators.[44]

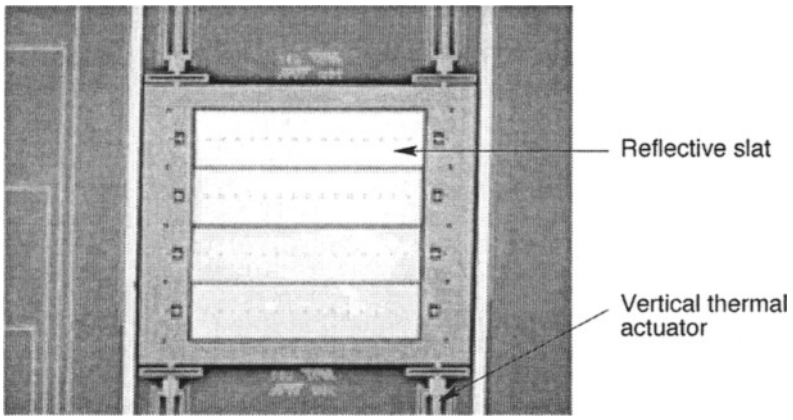


Fig. 12.28. Electrostatically toggled mirror slats (100 μm × 400 μm) setup using backbent vertical thermal actuators.[44]

Other applications for beam-steering micromirrors are in optical-scanning systems. Optical scanners have long been used by the commercial and defense industry, the goal to reduce the size and cost of these scanners. Applications of optical scanners are seen in bar-code scanners, compact disk players and recorders, laser printers, laser radars, laser-guided weapons, holographic storage devices, and data links (routing links or switches) between integrated circuit chips.

A scanning micromirror system is shown in Fig. 12.29.[28] This system has three main elements: an actuator array, a locking tether, and a flip-up mirror. The mirror is connected to the substrate with two substrate hinges, and to the actuator array with a self-locking tether. The actuator array is used to set the angle of the mirror for fine adjustment of an optical system, or to move the plate continuously to create a scanning mirror with a large scan angle (up to 20 deg).[28] The locking tether is 15 μm wide and 100 μm long. The tether is connected to the actuator yoke by a 3-μm-wide, 15-μm-long flexure. At the other end of the tether, a microlatch is used to secure the mirror. The mirror plate is constructed of stacked polysilicon layers and coated with gold. The gold-coated mirror surface is a square with 75-μm sides. A 3 × 3-μm-square etch hole is cut in the center of the mirror to ensure that the mirror is completely released during the oxide-etching process.

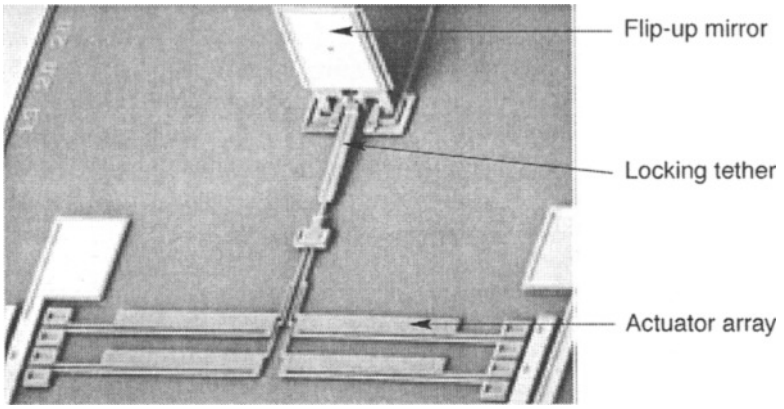


Fig. 12.29. Scanning micromirror system using four-element thermal actuator array.[28]

A rotating scanning mirror system is shown in Fig. 12.30.[28] This system is composed of two thermal actuator arrays, a flip-up mirror plate, and a rotating base. The mirror plate is 100 × 100 μm, and the rotating base has a diameter of 220 μm. The mirror plate is constructed of stacked polysilicon layers and coated with gold. Two actuator arrays are used to rotate and position

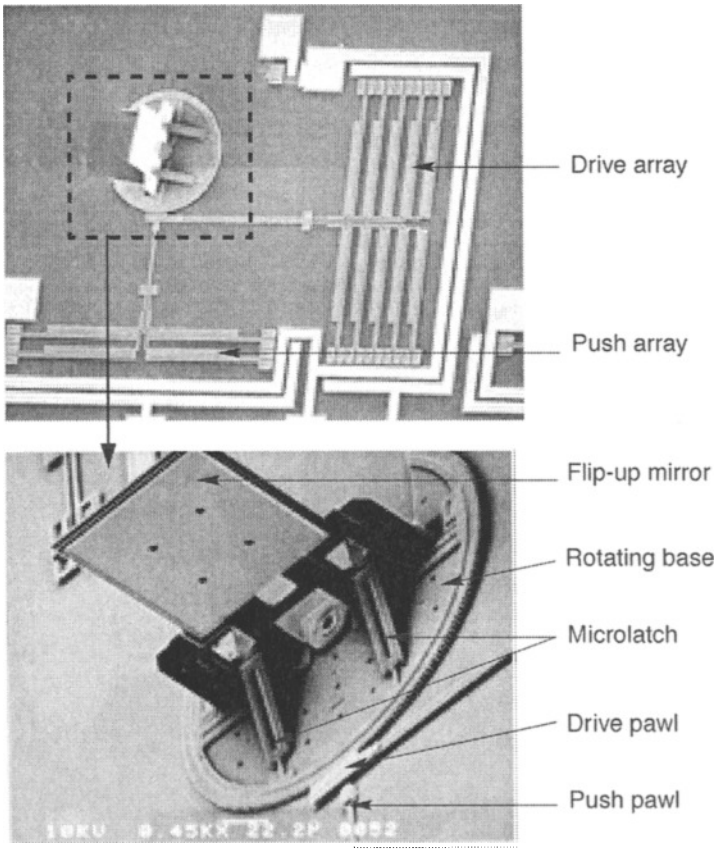


Fig. 12.30. Rotating micromirror system.[28]

Next Page

the mirror. The mirror can be rotated in either direction with a total range of 210 deg. The direction of rotation is controlled by the sequencing of the push and drive thermal actuator arrays.[13]

Computer-controlled electrical interfaces can be developed to automate the positioning of both the scanning and rotating micromirrors shown in Figs. 12.29 and 12.30. The high force of the thermal actuator arrays and the ability to design thermal actuators that operate in the voltage and current regime of digital CMOS technology allow the design of simple control mechanisms for these optical systems. A voltage amplitude control and a pulse modulation (100 KHz) scheme using average power have been demonstrated.[28]

For the scanning micromirror of Fig. 12.29, a computer takes user inputs and converts them to positioning commands. The positioning commands are based on an empirically derived relationship between applied voltage and mirror plate movement. The performance of a scanning micromirror driven by a four-element thermal actuator array is shown in Fig. 12.31. There is little deflection below 1 V because the current flow is not sufficient to heat the thermal actuators. Above 4 V, the mirror deflection is limited by mechanical interference in the design of the micromirror.[28]

The rotating micromirror of Fig. 12.30 has different control requirements than the scanning mirror. Unlike the scanning mirror, the actuator arrays used in the rotating mirror are not required to hold the mirror position. Instead, they are used as a motor to position the rotating base.[13,45] A CMOS-based, application-specific integrated circuit (ASIC) was designed that has large output drivers and a digital interface.[28] In this design, the computer accepts user inputs for positioning the mirror and sends control inputs to the CMOS controller via a computer interface card. The controller then selects the proper output channel and drives the actuator array. The mirror is able to rotate through a range of 210 deg. The micromotor can move the mirror through the entire 210-deg range in 200 steps. This capability resulted in a positioning resolution of slightly greater than 1 deg with less than 3 deg of error.[28] A finer positioning accuracy can be achieved by designing a larger motor gear with more teeth.

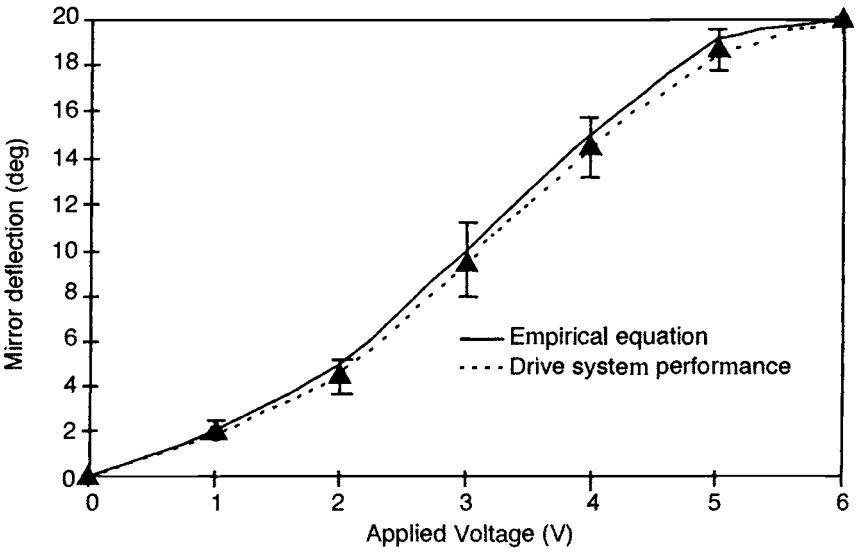


Fig. 12.31. Scanning micromirror deflection vs applied voltage using four-element thermal actuator array.[28] The positioning of the plate is not as precise as desired, as demonstrated by the RMS error bars. This is due to the low tolerances in the design of the micromirror hinges, allowing too much play in the movement of the mirror plate. The mirror is shown in Fig. 12.29.

12.7 Corner-Cube Reflector

A corner cube reflector (CCR) has recently been studied as a possible MEMS communication link.[13,46,47] A CCR has three mutually perpendicular, mirrored walls. This mirror arrangement reflects light back in the direction of its incoming path. A sample CCR design in Fig. 12.32 has a static gold mirror on the substrate and two hinged, gold-covered mirrors. One hinged mirror is positioned and modulated with a thermal actuator array. The other hinged mirror is held by a slotted locking plate designed to position the mirror within 3.5 mrad of perpendicularity with the substrate.[13] The hinged mirror plates are constructed of two releasable MUMPs polysilicon layers with a layer of silicon oxide trapped between them, for a total thickness of 4.25 μm.[48]

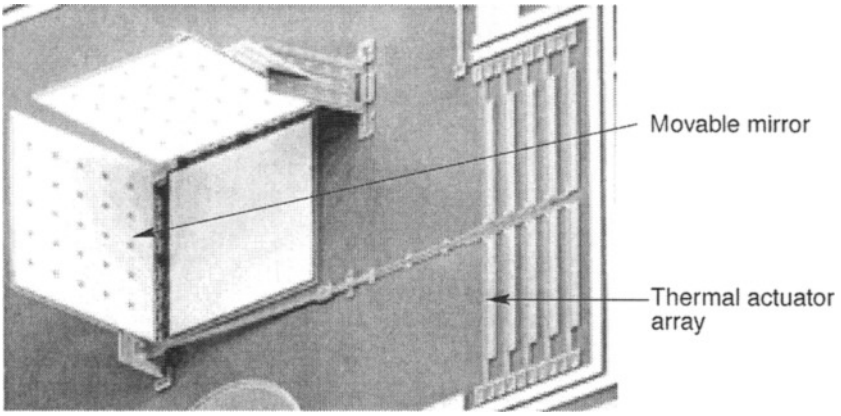


Fig. 12.32. Corner cube reflector with a thermal actuator array to position and modulate a hinged plate mirror.[13] Mirrors are 280 μm square, 4.25 μm thick sandwich of polysilicon surrounding trapped silicon oxide.

12.8 Fresnel Lens

Since surface micromachining limits the designer to materials with uniform layer thicknesses, it is not possible to design curved refracting lenses; however, Fresnel diffracting lenses can be fabricated easily. Fresnel lenses can be used to collimate light from a laser diode.[12,49] Fresnel lenses can also be used to focus light to and from optical fibers. An example of an eight-element array of Fresnel lenses is in Fig. 12.33.[50] The structure is fabricated in the Poly-1 layer and is hinged to

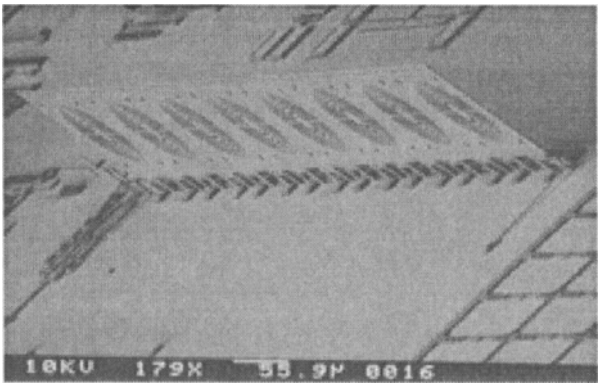


Fig. 12.33. Scanning electron micrograph of an eight-element hinged Fresnel lens array.[50] The plate is 200 μm tall and locks into place.

the substrate so it can be flipped up into the light path. Microlatches catch the lens plate and hold it in position. The Fresnel lenses shown in Fig. 12.33 have a designed focal length of 1000 μm. The focal length was measured to be 1277 μm, with a standard deviation of 275 μm.[50]

Figure 12.34 shows an example of a Fresnel lens designed onto an elevated platform.[9] These types of platforms can be used to support optical components such as mirrors, lenses, and gratings. The base of the side plates are connected to substrate hinges. The elevated platform can be designed to slide over another device that is fabricated on the surface of the substrate.[9] The angle between the suspended platform and the substrate can be also defined during device design by sloping the top of the two supporting plates. This type of device can potentially be used for interfacing optical fibers, semiconductor lasers, and optical components placed on the substrate.

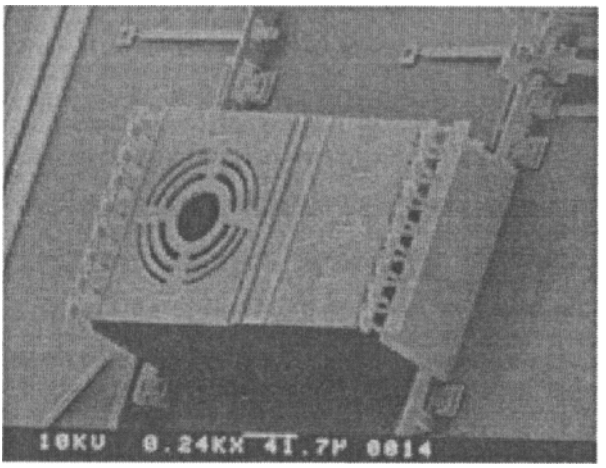


Fig. 12.34. A Fresnel lens raised 100 μm above the substrate, using several flip-up plates connected by microhinges.[9]

12.9 Gratings

Optical gratings can be easily fabricated using micromachining techniques. For example, a grating light valve[51] has been constructed in which individual grating pixels are electrostatically moved out of the substrate plane to diffract light of a particular wavelength at a designed angle. Gratings have also been investigated as tunable optical filters,[1] in which the grating period is adjusted by a linear electromagnetic actuator. Gratings can also serve as beam splitters, optical switches, and beam-steering devices.

Grating designs (Fig. 12.35) can be analyzed using optical diffraction theory. The diffraction angles for a normally incident light are given by the grating equation:

$$\theta_m = \sin^{-1}\left(\frac{m\lambda}{a}\right) \tag{12.2}$$

where θ_m is the direction of the m^{th} diffraction order, λ is the incident wavelength, and a is the grating period.[52] The intensities of the diffracted orders cannot be determined using Eq. (12.2), but they can be obtained from the Fourier transform of the phase description of the grating.[53,54]

Variable blaze gratings (VBGs) can be used to steer monochromatic light in discrete directions corresponding to each diffraction order.[55] Each reflective slat is tilted so specular reflection of

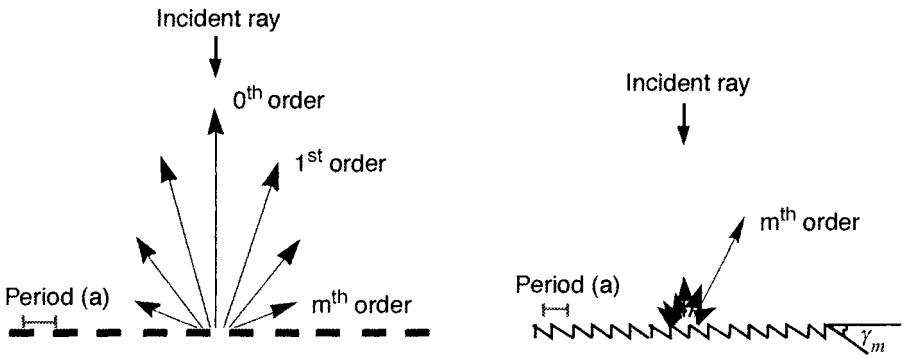


Fig. 12.35. Grating diffraction.[56] The variable blaze grating is shown on the right.

incident light off the slat matches a desired diffraction order. Because the direction of each diffraction order depends on the wavelength of the incident light [Eq. (12.2)], VBGs are used to steer only monochromatic or nearly monochromatic light. The slat tilt angle γ_m in Fig. 12.35 is referred to as the VBG blaze angle, and is equal to half θ_m (for normally incident light). Slats in a VBG are tilted to the same blaze angle.

An electrostatic grating example is shown in Fig. 12.36. This grating moves normal to the plane of the substrate to change the phase relationship between light reflected off the grating lines and the substrate.[54] This device is designed to modulate optical intensity by shifting power from the zero diffracted order to the $\pm 1^{st}$ diffracted orders. The movable grating is attached to the substrate by flexure beams, which provide a restoring mechanical force. The design incorporates drive plates at edges of the grating, so the grating lines themselves are not part of the electrostatic actuation. Since the drive force is applied only to the drive plates, the grating lines do not experience a downward force at their centers, and thus do not sag. Two antisag support lines keep grating lines parallel during actuation. The grating has 0.75-μm-deep dimples in upper electrode

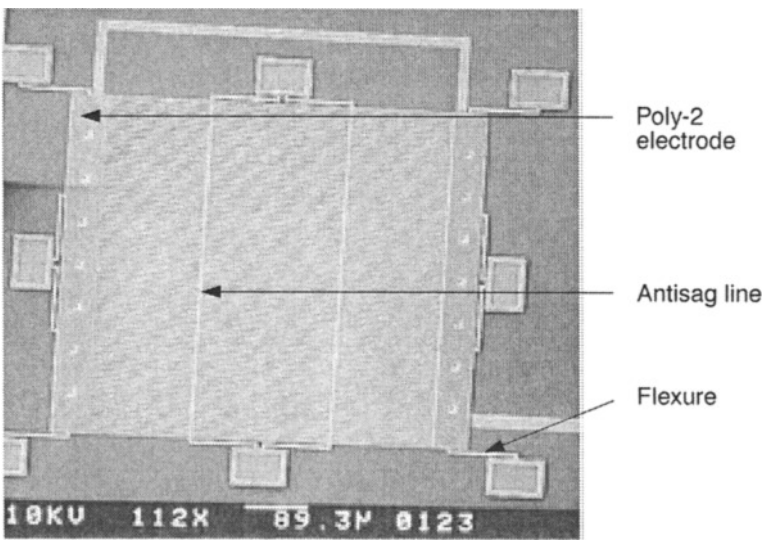


Fig. 12.36. Electrostatically actuated grating.[54] The grating is 500 × 500 μm, with 2-μm-wide lines spaced 4 μm center to center, for a total of 125 periods.

surfaces to prevent adhesion of the device to the substrate after the release etch, and to prevent the upper electrodes from contacting and shorting to the lower electrodes during actuation. This grating has an active area of 500 × 500 μm with 2-μm lines spaced 4 μm center to center. With these dimensions, only diffracted orders up to the third order contain significant diffracted power.[54] The power diffracted into orders higher than ±1 is only 10–15% of the total diffracted power. The angular separation of the diffracted orders agrees with optical diffraction theory within 2%.[54]

Figure 12.37 shows diffracted orders as a function of the drive voltage. As the voltage increases, transition time between maximum and minimum intensities decreases due to the nonlinear deflection vs voltage response of electrostatically actuated devices (see Eq. [12.1]). A higher percentage of the diffracted power is exchanged between the 0 and ±1 orders at higher voltages, since at higher voltages the entire grating is pulled closer to the substrate, better maintaining the phasing between light reflected from the grating and the substrate. The frequency response of the grating (first diffracted order) showed a 3-dB intensity roll-off at 45 kHz.[54] Only 3 V were required to change the first diffracted order from maximum to minimum intensity (18.1 dB contrast ratio), making this device an excellent optical switch.

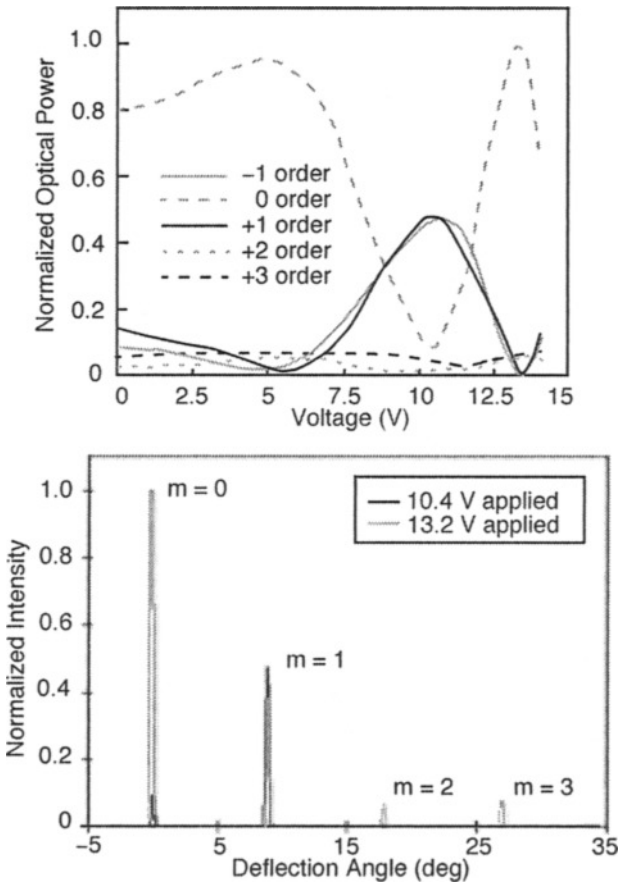


Fig. 12.37. Diffracted orders (m = 0,1,2,3) vs applied voltage for the electrostatic grating shown in Fig. 12.36.[54]

A laterally actuated diffraction grating is shown in Fig. 12.38.[54] This device consists of two gratings. The top grating is moved laterally over the lower grating. The lower grating can also be driven electrostatically normal to the plane of the substrate. In this version, a pair of thermal actuators drive the upper grating, but any actuator that provides lateral motion can be used. The upper sliding and lower static gratings interact to shift the diffracted light intensity between the first and second orders. The p-doped MUMPs polysilicon has a reflectivity of only ~35% at the 632.8 nm illumination wavelength, but the device is designed so the grating area can be plated with a more reflective surface if needed. An evaluation of optical properties of the variable diffraction grating showed that it matched the theory within 0.1 deg for the diffracted angle values and the change in intensity of the diffracted orders versus grating deflection.[54] Figure 12.39 shows the theoretical and measured intensity profiles for the first two diffraction orders, showing the energy switching between orders. Another example of a grating design is shown in Fig. 12.40: a flip-up hinged grating is assembled perpendicularly to a rotating gear.[29] The gear replaces the rotor section of the rotary stepper motor in Fig. 12.16 and allows 180-deg rotation. This device can potentially be used as a beamsplitter or as a diffractive element in a microspectrometer.

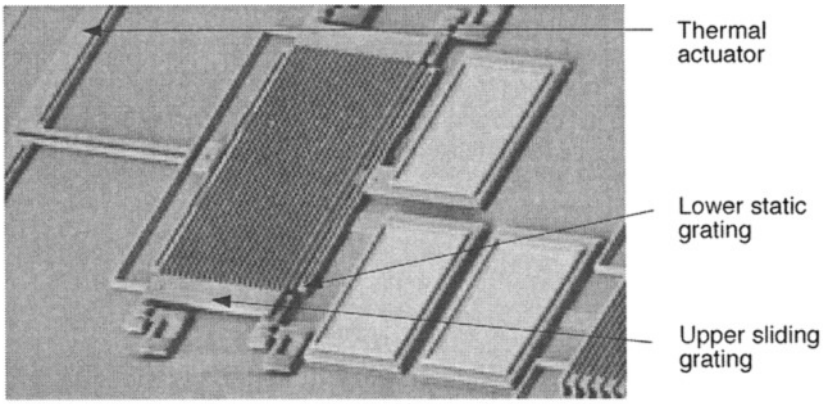


Fig. 12.38. A 90 × 120 μm variable diffraction grating.[54] Grating lines are 2 μm wide with 4 μm period.

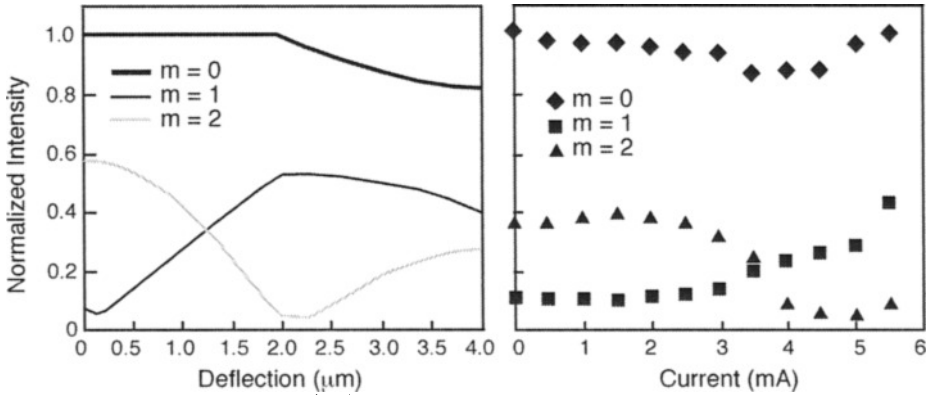


Fig. 12.39. Theoretical (left) and measured (right) intensity for the 0^{th}, 1^{st}, and 2^{nd} diffracted orders for the variable diffraction grating shown in Fig. 12.38.[54] The theoretical model shows lateral deflection of the upper plate vs intensity in the orders; the measured data show the thermal actuator drive current vs intensity in the orders.

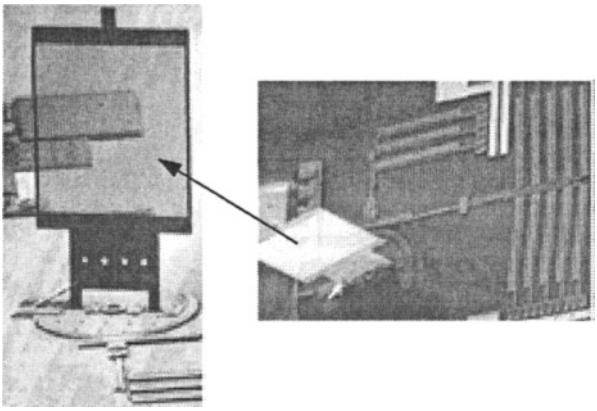
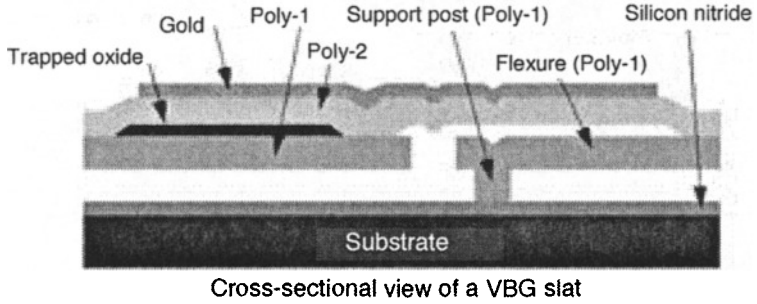


Fig. 12.40. Rotating grating on a 200-μm-diam gear, which allows 180 deg of positioning.[29] Grating is 185 × 200 μm with 2-μm-wide lines and spaces.

An electrostatically actuated VBG example is shown in Figs. 12.41 and 12.42.[55,57] The grating slats are connected with flexible wiring at each end to a power bus. The flexible wiring also supports the ends of the slats and reduces curvature induced by residual material stress along the length of the slat. Support posts and flexures under the slat surface are formed of Poly-1 and covered by layers of Poly-2 and gold (Fig. 12.41). Slats are constructed using trapped oxide. When a voltage is applied between the slat and the substrate, the slat tilts down toward the substrate until



Cross-sectional view of a VBG slat

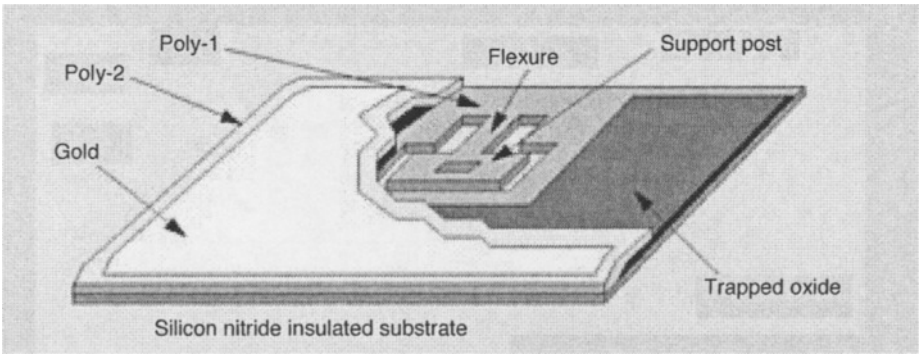


Fig. 12.41. Diagram of the slat support post and flexure used in the electrostatically actuated variable blaze grating.[57]

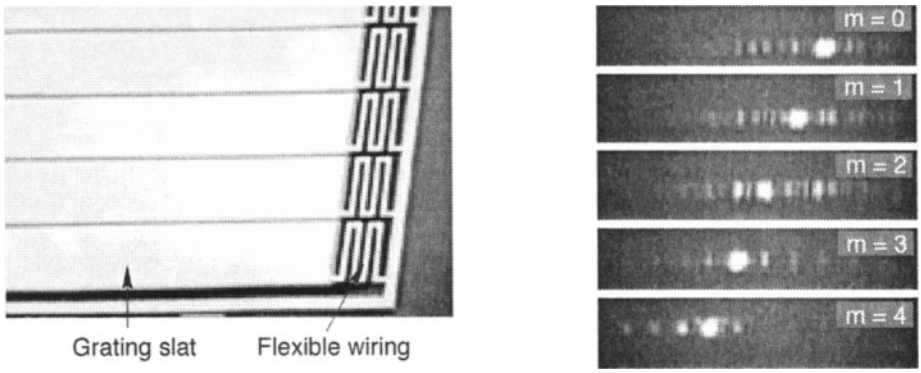


Fig. 12.42. Scanning electron micrograph and far-field intensity patterns of an electrostatically actuated 36-slat VBG with a period of 80 μm.[55] The gap between adjacent 3-mm-long slats of the grating is 2 μm.

the Poly-2 layer rests on the Poly-1 support post. If the voltage is increased further, then the slat rotates on top of the post until the edge of the slat opposite the flexure touches the silicon nitride layer.

The VBG was tested with a normally incident 632.8-nm wavelength HeNe laser in ambient air pressure and temperature. The grating supported five steered beam directions with excellent agreement between calculated and measured diffraction orders.[55] The beam steering range was 1.81deg. The ratio of the energy in the selected diffraction order to energy in nonselected diffraction orders was found to be 15 or higher.[55]

Another method of steering an optical beam is with a phased linear micromirror array, in which individual mirrors are pistoned down to support a nonzero beam-steering angle. Micromirrors can either have continuously or binary adjustable positions Continuously adjustable phase offers more flexibility in steering an optical beam. For continuous mode operation, the height of the array mirror element can be arbitrarily decreased by pulling the array element toward the substrate with electrostatic force. Binary mode operation uses a fixed voltage to pull the array element down toward the substrate. Continuous mode operation is thus defined as deflecting the array element less than the snap-through distance. Binary mode operation is defined as deflecting the array element more than the snap-through distance. Equation (12.3) gives the round-trip optical phase change φ that results from displacing a mirror element distance d under normal illumination by a source of wavelength λ:

$$\varphi = \frac{4d\pi}{\lambda} \tag{12.3}$$

During device operation, the height of each mirror element in the array is adjusted by applying individual voltages to each array element. When the array element height is changed, the optical path difference between the array element and the phase front of the incident light changes. The array element optical path difference determines the phase of incident light reflected off each array element. The phase of light reflected off each array element is set so that the phase front of the reflected light from the array is a plane perpendicular to the desired beam-steering direction, θ. The interelement phase difference is defined as the change in phase between light reflected off an array element and light reflected off an adjacent array element. Equation (12.4) gives the interelement phase difference Ψ required to steer normally incident light to an angle θ.[58]

$$\Psi = \frac{2\pi b}{\lambda}\sin\theta \tag{12.4}$$

where b is the array period (distance between the centers of adjacent mirrors). This technique has been successfully demonstrated to steer the output of a laser diode array using a linear micromirror array with 10 elements.[59]

Linear phased arrays of micromirrors also have potential applications in holographic data storage systems. These systems encode data in interference patterns and store the patterns in photorefractive crystals. The interference patterns may be controlled using angle multiplexing, in which each reference beam is employed separately at a different angle for each data record. Alternatively, the patterns may be controlled using phase code multiplexing, in which all reference beams are employed simultaneously (at many angles), with a different phase pattern for each data record. Linear phased arrays may be particularly appropriate for phase code multiplexing if very low (e.g., less than 1%) systematic errors in micromirror flatness and actuation displacement can be achieved.[60]

12.10 Microoptical Bench and Automated Assembly of Microsystems

The micromachined components can be combined on a micro-optical bench.[12] Figure 12.43 illustrates the concept with construction of a microscanner.[49] The microscanner uses a vertical cavity surface emitting laser (VCSEL) integrated on a MEMS optical bench. Potential applications for this microscanner are bar-code scanners, laser printers, laser radar, laser guided weapons, and data links between integrated circuit chips.

The design of the VCSEL microscanner starts with a surface micromachined MEMS die. Since VCSELs are fabricated using III-V semiconductor materials, the VCSEL device is inserted into a micromachined cavity in the substrate using a hybrid packaging scheme. The VCSEL is 600-μm long × 300-μm wide × 100-μm high, with an aperture of 8–10 μm in diameter. A mirror is placed over the VCSEL at a 45-deg angle. The laser beam from the VCSEL, which is normal to the surface of the substrate, hits the mirror and is reflected 90 deg, resulting in a beam parallel to the substrate. The laser beam then passes through a Fresnel lens that collimates the beam. A rotating mirror then scans the beam laterally. The final optical device is a fan mirror, which scans the beam vertically and reflects the beam normal to the substrate. The fabricated scanner is shown in Fig. 12.44.[49]

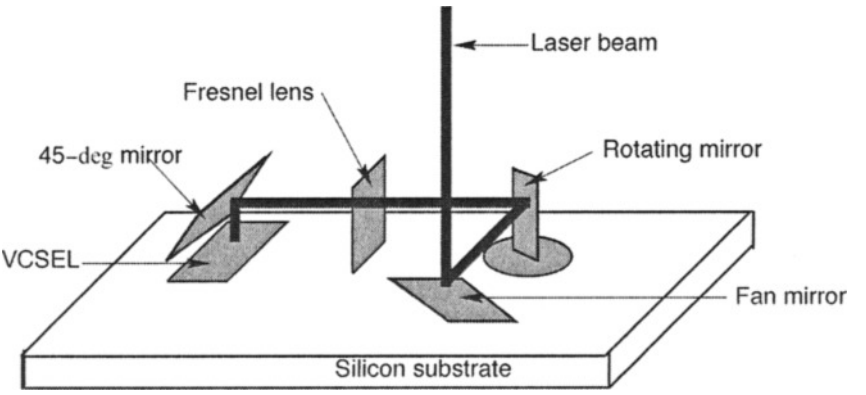


Fig. 12.43. Schematic diagram of VCSEL optical scanner.[49]

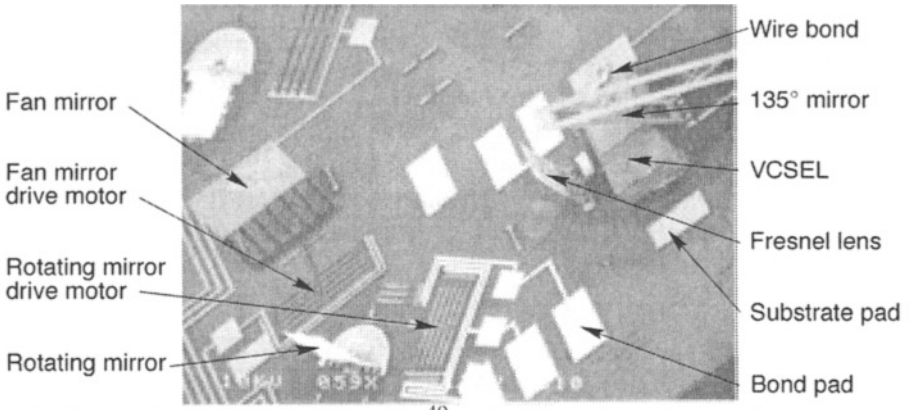


Fig. 12.44. Micrograph of VCSEL optical scanner.[49] Device dimensions: 135-deg mirror (200 × 151 μm; Fresnel lens (500-μm focal length); rotating mirror (230 × 230 μm); fan mirror (230 × 497 μm); VCSEL (600 × 300 × 100 μm). Entire microscanner is less than 5 × 5 mm. Scanner performance: lateral scanning angle is 5.7 deg; vertical scanning angle is 4.4 deg.

Hinged devices published to date are typically assembled and positioned by hand, a time-consuming and delicate process. However, thermally actuated stepper motors have sufficient force to raise and position hinged plates, and to remotely adjust micromachined systems with low voltages. This eliminates the need for manual assembly or adjustment, thus making batch fabrication and automated assembly feasible. The automated assembly (also known as self-assembly) of MEMS may be particularly important for space-based applications. Automated assembly allows deployment and remote assembly of MEMS in the field of operation or readjustment of components to align a system for better performance.

Automated assembly of flip-up structures has been demonstrated.[14] Figure 12.45 shows the automated assembly system connected to a scanning micromirror.[14] The system consists of three separate parts: linear assembly motor (Fig. 12.45 a-c), vertical actuator (Fig. 12.45 d), and

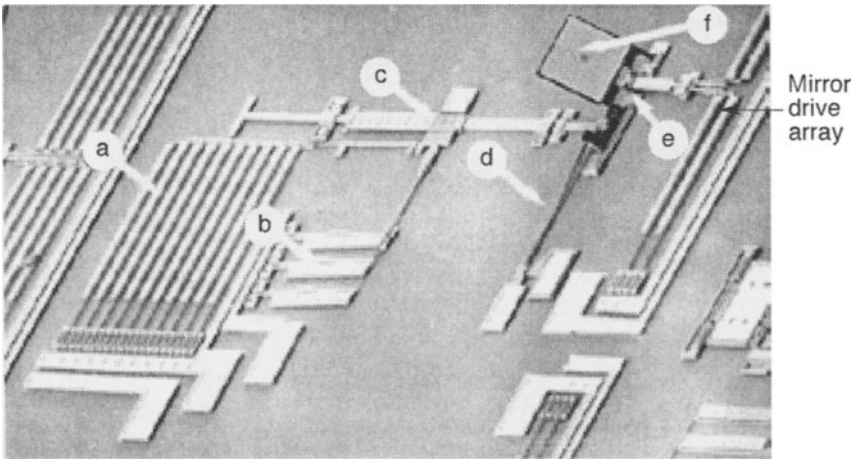


Fig. 12.45. Automated assembly system for scanning micromirror.[14] The various components are: (a) assembly motor drive array, (b) assembly motor push array, (c) linear drive arm, (d) vertical thermal actuator, (e) self-engaging locking mechanism, and (f) scanning micromirror. The scanning micromirror (f) is 75 × 75 μm.

microlatch (Fig. 12.45 e). A lift arm is connected to the drive rod of the linear assembly motor with a hinge. The other end of the lift arm is connected to the flip-up mirrored plate. First, the vertical actuator is used to lift the free end of the flip-up plate off the substrate, forming a triangle with the substrate, the lift arm, and the flip-up plate. The linear assembly motor then drives the lift arm toward the base of the flip-up plate, thus rotating the flip-up plate around its substrate hinges. Finally, the microlatch engages to hold the flip-up mirrored plate in position. If necessary, the linear assembly motor can be reversed to pull the lift arm away from the assembled structure.

A computer-based control system was developed to automate the sequencing of the linear motor and drive the micromirror after assembly.[28] A block diagram of the system is shown in Fig. 12.46. The CMOS ASIC has a digital interface and custom-designed output drivers to handle the current requirements of the thermal actuator arrays in the linear motor (~35 mA) and the scanning micromirror (~15 mA). The computer accepts user inputs for driving the linear motor or micromirror and sends control inputs to the CMOS ASIC via a computer interface card. The ASIC then selects the proper output channel and drives the actuator array.

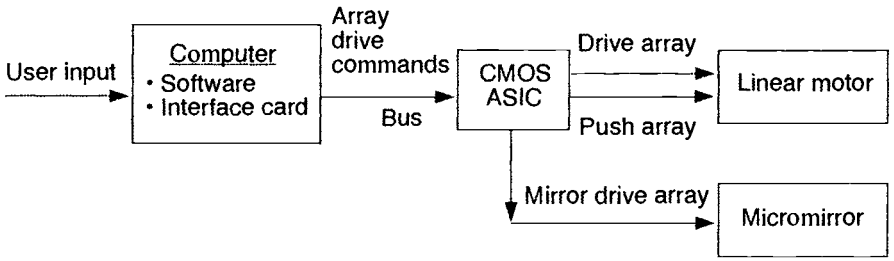


Fig. 12.46. Computer control system for automated assembly and operation of scanning micromirror.[28]

The computer-based system is highly flexible and has been used successfully to drive manually assembled scanning micromirrors as well as auto-assembled micromirrors. Positioning the micromirror plate can be controlled either by varying dc voltage delivered to the micromirror actuator array or by using a pulsed drive signal.[28] Use of one automated control system for assembling and operating the scanning micromirrors greatly simplifies the development and implementation of a practical microsystem. Incorporation of the automated assembly system does not degrade the performance of the scanning micromirror. The micromirrors erected with the automated assembly system were capable of scanning through the complete range of 20 deg (see Fig. 12.31).

The automated assembly system used for the scanning micromirror also has been used for assembling a CCR. Figure 12.47 shows an assembled CCR.[14] The vertical hinged mirrors are modulated by thermal actuator arrays. The gold layer on both vertical plates of the CCR is square, with 75 μm per side, and is extended down to the base of the polysilicon plate. The polysilicon plate is 160 μm wide, and has been extended laterally so that the locking mechanism will not interfere with the reflective surface of the mirror. Assembly of the CCR is essentially the same as that of the scanning micromirror, except the process is done twice to raise both of the vertical plates. Moreover, the same digital control system used for the scanning mirror can be used for assembling and operating the CCR.

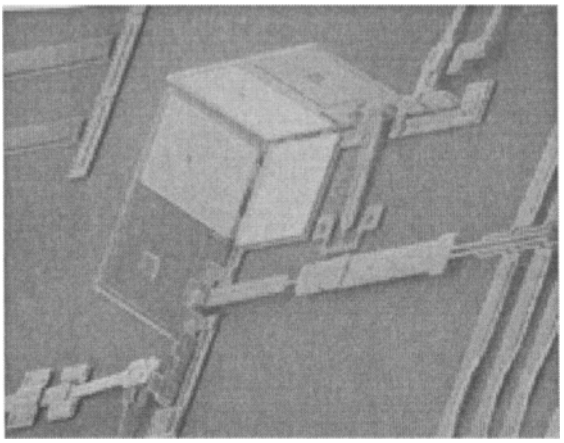


Fig. 12.47. A corner cube reflector assembled with the automated assembly system.[14]

12.11 Summary

Microprocessor-based control systems need to be developed to control optical MEMS. The digital interface would automate the assembly, positioning, and control of many microsystems. CMOS technology in particular is an attractive choice for integration with MEMS because of the wide availability of CMOS interface circuits and computer-aided design (CAD) tools. This implies that the actuators used in the micro-optical systems need to be compatible with CMOS voltage and current levels. All of the logic functions for the control of MEMS can be implemented in ASICs or programmable gate arrays. Monolithic integration of electronics and MEMS would be the best solution for aerospace applications requiring high levels of integration and performance. Currently, the microfabrication technology is not advanced enough to provide the best possible performance from monolithically integrated MEMS and electronics. However, a combination of ASIC die with MEMS die in a multichip module[61] would provide an excellent solution for many aerospace applications.

12.12 References

1. J. A. Cox, J. D. Zook, T. Ohnstein, and D. C. Dobson, "Optical Performance of High-Aspect LIGA Gratings," *Opt. Eng.* **36** (5), 1367–1373 (May 1997).
2. G. Vdovin, S. Middelhoek, and L. Sarro, "Deformable Mirror Display with Continuous Reflecting Surface Micromachined in Silicon," *IEEE Micro Electro Mechanical Systems*, (1995). pp. 61–64.
3. G. Vdovin and P. M. Sarro, "Flexible Mirror Micromachined in Silicon," *Appl.Opt.* **34**, 2968–2972 (1 June 1995).
4. G. Vdovin, S. Middelhoek, and P. M. Sarro, "Thin-film Free-Space Optical Components Micromachined in Silicon," *Digest IEEE/LEOS 1996 Summer Topical Meetings* (Keystone, CO, 5–9 August 1996), pp. 5–6.
5. D. Koester, R. Mahedevan, A. Shishkoff, and K. Marcus, *Multi-User MEMS Processes (MUMPS) Introduction and Design Rules*, Rev. 4, MCNC MEMS Technology Applications Center, Research Triangle Park, NC (15 July 1996).
6. D. M. Burns and V. M. Bright, "Designs to Improve Polysilicon Micromirror Surface Topology," *Proceedings SPIE* (1997), Vol. 3008, pp. 100–110.
7. J. H. Comtois. "Structures and Techniques for Implementing and Packaging Complex, Large Scale Microelectromechanical Systems Using Foundry Fabrication Processes," Ph.D. thesis, Air Force Institute of Technology, June 1996.

8. R. Nasby, J. Sneigowski, J. Smith, S. Montague, C. Barron, W. Eaton, and P. McWhorter, "Application of Chemical-Mechanical Polishing to Planarization of Surface-Micromachined Devices," *Technical Digest, Solid State Sensors and Actuators Workshop* (Hilton Head, SC, 3–6 June 1996), pp. 48–53.
9. J. R. Reid, "Microelectromechanical Isolation of Acoustic Wave Resonators," Ph.D. thesis, Air Force Institute of Technology, December1996.
10. K. S. J. Pister, M. W. Judy, S. R. Burgett, and R. S. Fearing, "Microfabricated Hinges," *Sensors and Actuators* **A-33**, 249–256 (1992).
11. S. R. Burgett, K. S. J. Pister, and R. S. Fearing, "Three Dimensional Structures Made with Microfabricated Hinges," *Proceedings ASME, Micromechanical Systems* (1992), pp. 1–11.
12. M. C. Wu, L. Y. Lin, and S. S. Lee, "Micromachined Free-Space Integrated Optics," *Proceedings SPIE* (1994), Vol. 2291, pp. 40–51.
13. J. H. Comtois and V. M. Bright, "Surface Micromachined Polysilicon Thermal Actuator Arrays and Applications," *Technical Digest, Solid State Sensors and Actuators Workshop* (Hilton Head, SC, 3–6 June 1996), pp. 174–177.
14. J. R. Reid, V. M. Bright, and J. H. Comtois, "Automated Assembly of Flip-up Micromirrors," *Digest of Technical Papers, 1997 International Conference on Solid-State Sensors and Actuators (Transducers '97)* (Chicago, IL, 16–19 June 1997), Vol. 1, pp. 347–350.
15. M. A. Michalicek, "Design, Fabrication, Modeling, and Testing of Surface-Micromachined Micromirror Devices," M. S. thesis, Air Force Institute of Technology, June 1995.
16. J. M. Younse, "Mirrors on a Chip," *IEEE Spectrum* **1993**, 27–31.
17. M. C. Roggemann, V. M. Bright, B. M. Welsh, S. R. Hick, P. C. Roberts, W. D. Cowan, and J. H. Comtois, "Use of Micro-Electro-Mechanical Deformable Mirrors to Control Aberrations in Optical Systems: Theoretical and Experimental Results," *Opt. Eng.* **36**, 1326–1338 (May 1997).
18. M. K. Lee, W. D. Cowan, B. M. Welsh, V. M. Bright, and M. C. Roggemann, "Aberration Correction Results From a Segmented Micro-Electro-Mechanical Deformable Mirror and a Refractive Lenslet Array" *Opt. Lett.* **23** (8), 645–647 (1998).
19. M. A. Michalicek, V. M. Bright, and J. H. Comtois, "Design, Fabrication, Modeling, and Testing of a Surface-Micromachined Micromirror Device," *Proceedings ASME Dynamic Systems and Control Division* (1995), DSC-Vol. 57-2, pp. 981–988.
20. T. H. Lin, "Implementation and Characterization of a Flexure Beam Micromechanical Spatial Light Modulator," *Opt.Engr.* **33** (11), 3643–3648 (November 1994).
21. P. M. Osterberg, R. K. Gupta, J. R. Gilbert, and S. D. Senturia, "Quantitative Models for the Measurement of Residual Stress, Poisson Ratio and Young's Modulus Using Electrostatic Pull-in of Beams and Diaphragms," *Proceedings Solid-State Sensor and Actuator Workshop* (Hilton Head Island, SC, 13–16 June, 1994), pp. 184–188.
22. P.B. Chu, P.R. Nelson, M.L. Tachiki, and K.S.J. Pister, "Dynamics of Polysilicon Parallel-Plate Electrostatic Actuators," *Sensors and Actuators* **A-52**, 216–220 (1996).
23. T. Akiyama and H. Fujita, "A Quantitative Analysis of Scratch Drive Actuator Using Buckling Motion," *Proceedings Eighth IEEE International MEMS Workshop* (1995), pp. 310–315.
24. W. C.-K. Tang, "Electronic Comb Drive for Resonant Sensor and Actuator Applications," Ph.D. thesis, University of California, Berkeley, CA, November 1990.
25. J. J. Sniegowski, S. L. Miller, G. F. LaVigne, M. S. Rodgers, and P. J. McWhorter, "Monolithic Geared-Mechanisms Driven by a Polysilicon Surface-Micromachined On-Chip Electrostatic Microengine," *Technical Digest, 1996 Solid-State Sensors and Actuators Workshop* (Hilton Head Island, SC, 2–6 June 1996), pp. 178–182.
26. M. J. Daneman, N. C. Tien, O. Solgaard, K. Y. Lau, and R. S. Muller, "Linear Vibromotor-Actuated Micromachined Microreflector for Integrated Optical Systems," *Technical Digest, 1996 Solid-State Sensors and Actuators Workshop* (Hilton Head Island, SC, 2–6 June 1996), pp. 109–112.
27. R. Kuhns, "Design and Fabrication of a Micro-Mechanical Gyroscope," M.S. thesis, Air Force Institute of Technology, December 1995.

28. J. T. Butler, V. M. Bright, and J. R. Reid, "Scanning and Rotating Micromirrors Using Thermal Actuators," *Proceedings SPIE: Optical Scanning Systems* (San Diego, CA, 30–31 July 1997), Vol. 3131, pp. 134–144.
29. J. H. Comtois and V. M. Bright, "Applications for Surface Micromachined Polysilicon Thermal Actuators and Arrays," *Sensors and Actuators* **A-58**, 19–25 (1977).
30. V. M. Bright, J. T. Butler, W. D. Cowan, D. M. Burns, and J. R. Reid, "Automated Assembly of Micro-Electro-Mechanical Systems" (to be published in the *Int. J. of Advanced Manufacturing Systems*).
31. J. R. Reid, V. M. Bright, and J. H. Comtois, "Force Measurements of Polysilicon Thermal Micro-Actuators," *Proceedings SPIE* (1996), Vol. 2882, pp. 296–306.
32. J. M. Beckers, "Adaptive Optics for Astronomy: Principles, Performance, and Applications," *Annu. Rev. Astron. Astrophys.* **31**, 13–62 (1993).
33. M. A. Ealey and J. A. Wellman, "Deformable Mirrors: Design Fundamentals, Key Performance Specifications, and Parametric trades," *Proceedings SPIE on Active and Adaptive Optical Components* (1991), Vol. 1543, pp. 36–51.
34. M. A. Ealey and J. F. Washeba, "Continuous Facesheet Low Voltage Deformable Mirrors," *Opt. Eng.* **29**, 1191–1198 (1990).
35. L. J. Hornbeck, "128 × 128 Deformable Mirror Device," *IEEE Transactions on Electron Devices* **ED-30**, 539–545 (1983).
36. L. Miller, M. L. Agronin, R. K. Bartman, W. J. Kaiser, T. W. Kenny, R. L. Norton, and E. C. Vote, "Fabrication and Characterization of a Micromachined Deformable Mirror for Adaptive Optics Applications," *Proceedings SPIE* (July 1993), Vol. 1945, pp. 421–430.
37. R. Krishnamoorthy, T. Bifano, and G. Sandri, "Statistical Performance Evaluation of Electrostatic Micro Actuators for a Deformable Mirror," *Proceedings SPIE* (1996), Vol. 2881, pp. 35–44.
38. T. Bifano, R. K. Mali, J. K. Dorton, J. Perreault, N. Vandelli, M. N. Horenstein, and D. A. Castanon, "Continuous-Membrane Surface-Micromachined Silicon Deformable Mirror," *Opt. Eng.* **36**, 1354–1360 (May 1997).
39. T. A. Rhoadarmer, V. M. Bright, B. M. Welsh, S. C. Gustafson, and T. H. Lin, "Interferometric Characterization of the Flexure Beam Micromirror Device," *Proceedings SPIE* (July 1994), Vol. 2291, pp. 13–23.
40. J. H. Comtois, V. M. Bright, S. C. Gustafson, and M. A. Michalicek, "Implementation of Hexagonal Micromirror Arrays as Phase Mostly Spatial Light Modulators," *Proceedings SPIE* (1995), Vol. 2641, pp. 76–87.
41. S. C. Gustafson, G. R. Little, V. M. Bright, J. H. Comtois, and E. A. Watson, "Micromirror Arrays for Coherent Beam Steering and Phase Control," *Proceedings SPIE* (1996), Vol. 2881, pp. 65–74.
42. W. D. Cowan, "Extension of Foundry Surface Micromachining Process for Fabrication of a Continuous Facesheet Deformable Mirror," Ph.D. thesis, Air Force Institute of Technology, June 1998.
43. R. L.Clark, J. R. Karpinsky, J. A. Hammer, R. Anderson, R. Lindsey, D. Brown, and P. Merrit, "Micro-Opto-Electro-Mechanical, (MOEM), Adaptive Optic System," *Proceedings SPIE* (1997), Vol. 3008, pp. 12–24.
44. W. D. Cowan and V. M. Bright, "Vertical Thermal Actuators for Micro-Opto-Electro-Mechanical Systems," *Proceedings SPIE: Microelectronic Structures and MEMS for Optical Processing III* (1997), Vol. 3226, pp. 137–146.
45. J. R. Reid, V. M. Bright, and J. H. Comtois, "A Surface Micromachined Rotating Micro-Mirror Normal to the Substrate," in *Digest IEEE/LEOS 1996 Summer Topical Meetings, Optical MEMS and Their Applications* (1996), pp. 39–40.
46. K. Brendley and R. Steeb, "Military Applications of Microelectromechanical Systems," RAND report to the Office of the Secretary of Defense, U.S. Air Force, U.S. Army, RAND, Santa Monica, CA (1993).
47. D. Gunawan, L. Lin, and K. Pister, "Micromachined Corner Cube Reflectors as a Communication Link," *Sensors and Actuators* **A:46–47**, 580–583 (1995).

48. J. H. Comtois and V. M. Bright, "Design Techniques for Surface-Micromachining Polysilicon Processes," *Proceedings SPIE* (1995), Vol. 2639, pp. 211–222.
49. J. G.Bouchard, V. M. Bright, and D. M. Burns, "Techniques and Applications for Integrating a Semiconductor Laser on a Surface Micromachined Die" (to be published in *Proceedings SPIE's Optoelectronics '98: Micro-Optics Integration and Assemblies*, San Jose, CA, 24–30 January 1998, Vol. 3289).
50. D. E. Sene, "Design, Fabrication, and Characterization of Micro Opto-Electro-Mechanical Systems," M.S. thesis, Air Force Institute of Technology, December 1995.
51. R. B. Apte, F. S. A. Sandejas, W. C. Banyai, and D. M. Bloom, "Deformable Grating Light Valves for High Resolution Displays," *Technical Digest, Solid State Sensor and Actuator Workshop* (Hilton Head, SC, 1994), pp. 1–6.
52. E. Hecht, *Optics*, 2nd ed. (Addison-Wesley Publishing, Reading, MA, 1990).
53. M. C. Hutley, *Diffraction Gratings* (Academic Press, NY, 1982).
54. D. E. Sene, V. M. Bright, J. H. Comtois, and J. W. Grantham, "Polysilicon Micromechanical Gratings for Optical Modulation," *Sensors and Actuators* **A-57**, 145–151 (1996).
55. D. M. Burns, V. M. Bright, S. C. Gustafson, and E.A. Watson, "Optical Beam Steering Using Surface Micromachined Gratings and Optical Phased Arrays," *Proceedings SPIE: Optical Scanning System* (San Diego, CA, 30–31 July, 1997), Vol. 3131, pp. 99–110.
56. D. M. Burns and V. M. Bright, "Micro-Electro-Mechanical Variable Blaze Gratings," *Proceedings IEEE MEMS-97 Workshop* (Nagoya, Japan, 1997), pp. 55–60.
57. D. M. Burns and V. M. Bright, "Development of Microelectromechanical Variable Blaze Gratings," *Sensors and Actuators* **A-64**, 7–15 (1998).
58. A. A. Oliner and G. H. Knittel, *Phased Array Antennas* (Artech House, Inc., Dedham, MA, 1972).
59. C. J. Christensen, V. M. Bright, J. W. Grantham, and J. H. Comtois, "Control of a Phase-Locked Laser Diode Array Using Piston Micromirrors," *Proceedings SPIE* (1996), Vol. 2881, pp. 26–34.
60. P. B. Catrysse, M. C. Bashaw, J. F. Heanue, and L. Hesselink, "Systems Issues in Digital Phase Code Multiplexed Holographic Data Storage," OSA Annual Meeting (Rochester, NY, 20–25 October 1996).
61. J. T. Butler, V. M. Bright, and J. H. Comtois, "Advanced Multichip Module Packaging of Microelectromechanical Systems," *Digest of Technical Papers, 1997 International Conference on Solid-State Sensors and Actuators (Transducers '97)* (Chicago, IL, June 16-19, 1997), Vol.1, pp. 261–264.

13

Micropackaging High-Density Radio-Frequency Microwave Circuits

L. P. B. Katehi* and R. F. Drayton†

13.1 Introduction

The more electromagnetic (em) waves are confined into two dimensions, the better they guide power along the direction perpendicular to the plane of confinement. This observation has been made repeatedly since the first attempts to use em waves for transmission of power. One method of wave confinement is the use of conducting walls, which results in high-efficiency metal waveguides. These waveguides allow the fields formed by em waves to propagate only above a certain frequency (cutoff frequency) and at the same time exhibit a frequency-varying phase velocity referred to as dispersion. The resulting method of wave guidance, therefore, is based on field shielding. Wave guidance based on far-field cancellation evolved to allow field propagation at lower frequencies, which inherently requires the use of relatively large metallic structures. While field shielding uses conducting walls, far-field cancellation uses wave-guiding structures called transmission lines. The propagation efficiency of these lines depends on two parameters: the geometry of the line conductors and the electrical properties of the material filling the space surrounding the conductors. The first transmission lines—the coaxial cable and the two-wire line—were used extensively in low-frequency systems and provided very effective signal guidance in first generation communication systems.

The invention of the transistor made apparent the necessity of very small transmission lines. These lines were needed for compatibility with the planar layout technology of the newly discovered solid-state devices and for effective coupling of power from these microscopic devices into the macroscopic systems. The effective coupling of power was achieved by implementation of integrated circuit (IC) technology, which allows the design of small radio frequency (RF) circuits that combine many functions and can be fabricated at low cost. IC technology advances such as very large scale integration (VLSI) greatly influenced spacecraft communication systems. Furthermore, there have been steady improvements during the past two decades in the scale of integration, the availability of new materials, batch-production yields, reliability of components, and raw performance of high-frequency (HF) and high-speed components. Because of these improvements, many frequency and speed requirements previously met by large volume and weight components are now achievable and reliable in miniature lightweight devices. Size reduction is provided by planar technology; however, in future systems, increased functionality, more power, and a further reduction in cost is desirable to optimize communication technology development. In satellite communication systems, where mass is a primary driver of cost, the advent of monolithic microwave integrated circuits (MMICs) has dramatically decreased the mass and cost of satellite communication systems. However, attempts to further reduce cost by using this miniaturization technique alone have met with diminishing returns, indicating that the application of conventional planar technology is reaching its limitations. Hence, alternative approaches are required to meet design and performance goals of future systems.

*Department of Electrical Engineering and Computer Science, University of Michigan, Ann Arbor, Mich.
†Department of Electrical and Computer Engineering, University of Minnesota, Minneapolis, Minnesota.

Future communication needs require increasing the functionality of communication systems to meet performance requirements of new highly integrated sensors and instruments that may also be found on satellites.[1] To maximize data transfer and to minimize ground operation cost, future communication systems must move to higher frequencies, such as Ka-band (25–40 GHz) rather than the traditional X-band (8–12 GHz). An example of a potential size reduction is shown in Fig. 13.1, where a traditional diplexer in a communication system is envisioned in a miniaturized version that uses a combination of standard MMIC and silicon (Si) micromachining technology.

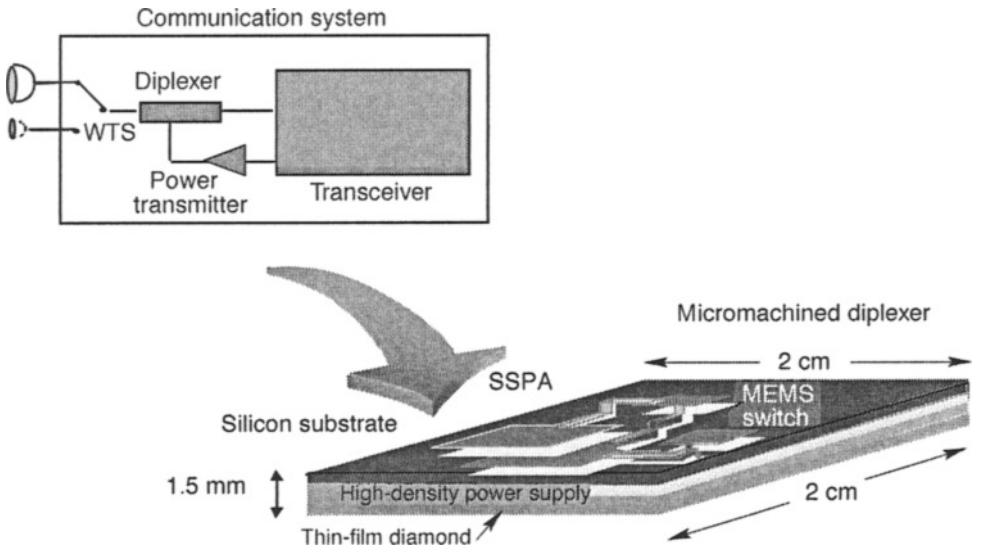


Fig. 13.1. Transformation of standard diplexer system with waveguide-based components into planar micromachined version. Some of the components in the diplexer are a solid-state power amplifier (SSPA), a microelectromechanical systems (MEMS) switch, and vertical interconnects that replace the waveguide-based version of these components, for example, a waveguide transfer switch (WTS).

Communication circuit miniaturization can be achieved by implementing three-dimensional (3D) packaging, where the circuits are arranged to be physically interconnected in all dimensions (see Fig. 13.2). This packaging approach has been effectively used to reduce the size and thereby increase the clock-rate speed of microprocessors. In digital circuits, multilevel integration uses a 3D packaging scheme whereby multiple power, ground, and signal lines are connected between the various levels through plated vertical holes called vias. This integration approach has been proven to be very effective for clock rates of the order of a few hundred megahertz, with higher harmonics reaching into the low-gigahertz range. A similar approach has recently been used to provide high-density integration (HDI) in HF radio and microwave circuits. The results offer an effective solution to circuit integration and packaging. Successful implementation of this approach at the lower gigahertz range, and in selective application at frequencies above the Ka band, form the basis of the study presented in this chapter.

Electronic packaging can account for up to 30% of the overall spacecraft mass, while the telecommunication subsystem can account for 15% or more of the dry mass. Based on this information, the key to reducing mass and improving performance is to address high-density integration and packaging in advanced HF microelectronics. The next step, therefore, is to move beyond the current state-of-the-art high-density packaging approaches of multichip modules (MCMs) into

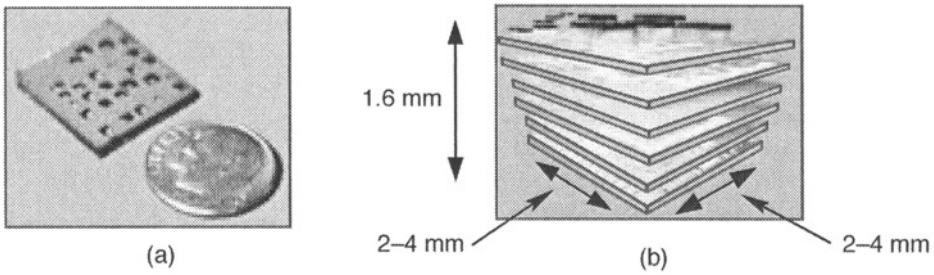


Fig. 13.2. State-of-the-art micromachined RF front-end system developed with high-density integration and micropackaging. (a) Photograph of a packaged microwave system with multiple layers; Si micropackaged RF front end. (b) Illustration of the compactness of the RF design with its various layers; miniature RF front end.

more advanced packaging approaches. These approaches integrate diverse technologies, such as HF electronics, Si-based (e.g., Si/germanium [Ge]) active circuits, advanced microelectromechanical systems (MEMS), and micromachined components (e.g., filter/multiplexers) to produce complex packaged systems in a single die. This advanced hybridized technology could be applied, for example, to a traditional communication transceiver shown in Fig. 13.1 that may use Si/Ge devices, MEMS switches, and micromachined filters and multiplexers. The result would be a significant reduction in mass (by a factor of 10) and in physical volume (by a factor of 75). The focus of this chapter is to illustrate the performance of various micropackaged components found in high-speed communication designs.

While the use of high-density HF electronics in space poses interesting and grand challenges, it also offers an opportunity to apply revolutionary concepts to circuit design, fabrication, and implementation. Moreover, since the specifications for compact system requirements cannot be satisfied by existing technologies alone, critical advancements are still needed that apply new concepts to the fundamental levels of circuit design and diagnostics.

13.2 Future Communication System Design Requirements

Advanced communication systems require the use of architectures that can handle mixed data signals (e.g., voice, video, still image) in systems that are wireless as well as portable. As data signal types proliferate, so do the requirements of the hardware needed to produce, manage, and distribute the signals. Furthermore, because of the complexity and diversity of these signals, traditional design methods are not capable of meeting the performance or design specification requirements for speed, size, or weight. In the context of a system, individual data types will continue to require unique circuit design rules and techniques to satisfy specific application needs. Once the data information has been encoded, transmission of the encoded information is typically done at higher frequencies, usually between RF waves at several gigahertz and millimeter waves up to 110 GHz. This transmission poses additional requirements and fabrication challenges.

In HF transmitter/receiver (T/R) technology, there has always been a demand to develop compact, lightweight systems for space applications. Other applications, however, have requirements for wireless and portable systems, which result in even greater demand for size and weight reduction. In addition to size and weight reduction, cost has also become a significant factor, since the customer base has expanded to include the individual user commercial market. Unfortunately, these cost requirements place even more challenging constraints on the development of advanced high-speed communication systems—primarily, that they be inexpensive to manufacture.

In response to these requirements, the question remains, "How can we cheaply develop compact high-speed systems that offer optimum performance and the flexibility to accommodate diverse complex data formats?" One approach is to implement high-density design techniques that use emerging technologies and materials. In T/R module design, two such design concepts are high-density packaging and high-density circuit design. Both techniques can significantly reduce the cost of HF systems. This chapter will present an advanced packaging approach for low-power, HF ICs and will introduce a novel high-density circuit design method that uses Si-based design approaches.

13.2.1 Advanced Micropackaging and High-Density HF Circuit Design

In HF applications where weight and size are critical, planar circuit and antenna technology offers conformal lightweight designs, low fabrication cost, and a large number of computer-aided design (CAD) tools, especially for microstrip transmission lines. Despite these features, loss mechanisms in the conductor, dielectric, and radiative planar lines can be large enough to have a substantial impact on HF circuit performance. As the operating frequency increases, the signal level strength of an active em device is reduced. In high-power applications, loss mechanisms can be tolerated; however, in low-power applications, high loss values result in substantial signal attenuation and distortion. The net result is a trade-off between size reduction gained by shifting to higher operating frequencies and the need to increase system circuitry (for amplification and reconditioning) to compensate for poor signal quality. The result is a more complicated circuit design that increases both the development and manufacturing cost.

The frequency of operation is an important factor in the selection of a circuit design approach. In low-frequency applications, design methods based on discrete lumped circuit components are used, because the size of individual elements are substantially smaller than the operating wavelength (λ). This wavelength is defined as f/v, where f is the operating frequency and v is the signal propagation velocity. In this case, the value of the individual electrical elements is nearly constant and can be used as a lumped element in discrete circuit design.

When the size of the component becomes comparable or larger than the operating wavelength, this approach is no longer valid. Then, the distributed nature of the element must be considered. This distributed nature is described by a distributed network of basic circuit element behavior: resistance, capacitance, inductance, and conductance. To account for the response of the component with respect to frequency, ideal transmission line theory is used, which is true for a variety of planar and nonplanar topologies. Examples of planar lines include microstrip, stripline, and coplanar waveguides and are illustrated in Figure 13.3. Once a specific type of transmission line has been chosen for a design implementation, analytic and numerical models can be used to determine the design parameters for the individual circuit elements and to analyze the response of the circuit.

In low-power HF applications, a distributed circuit can be developed using planar circuit technology. Here, the desired characteristics of R, L, C, and G values are obtained by appropriately

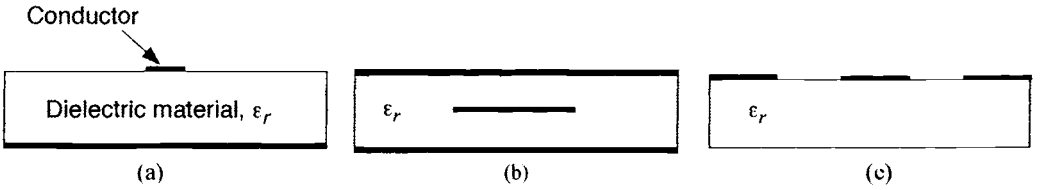


Fig. 13.3. Examples of planar transmission line geometries, where ε_r is the relative dielectric constant of the material. (a) Microstrip, (b) stripline, and (c) CPW.

choosing the conductor dimensions and dielectric constant of a planar material like a duroid, alumina, silicon, or gallium arsenide. As an example, a filter can be designed using filter theory for an L-C circuit. Once specific circuit element values are known, they can then be realized into a specific transmission line topology. This is done by determining the appropriate dimensions of each element in the corresponding transmission line to achieve the desired electrical response. In Figure 13.4, examples of microstrip-based lines are shown for an inductor and capacitor element. Above each microstrip cross section is a top view of the microstrip configuration denoted in black, and below each cross section is the corresponding circuit element. Note that the inductive section has a narrow microstrip line, while the capacitor has a wide microstrip line. Next, the individual element blocks are cascaded together to obtain the desired response of the design. For extensive discussion of transmission line theory and the specific types of lines, the reader is referred to Elliott, Ref. 2.

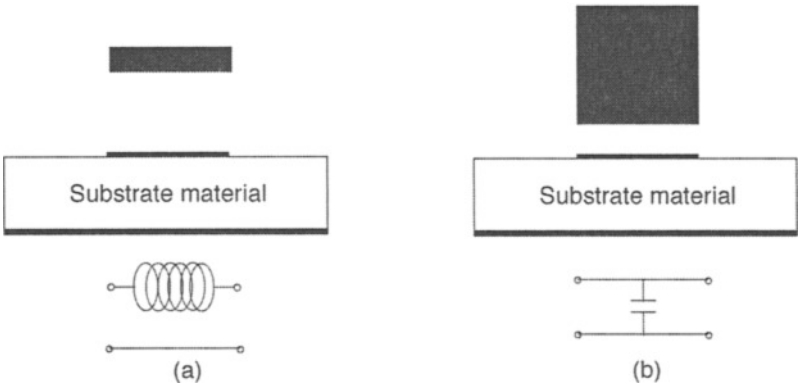


Fig. 13.4. Microstrip diagrams of inductive and capacitive sections of transmission lines. (a) Inductive section and (b) capacitive section.

One problem that arises when individual elements of different conductor widths are cascaded together is the formation of an interface discontinuity. This discontinuity, which is necessary for the design implementation, is also responsible for radiation at the interface junction. Radiation of this type can cause many performance problems in high-density circuit layouts because the energy radiated from the edges couple to any conductor located in the near vicinity. This coupling, also a form of em cross talk, results from unwanted energy from a section of the circuit actively carrying a signal to an unconnected nominally quiet section (i.e., nonsignal carrying)[3] of the design. A primary objective of the integrated package is to determine how to best contain and isolate the local radiation of individual circuits in order to reduce cross talk and circuit losses. With the use of micromachining, small integrated packages, hereafter referred to as micropackages, have been developed in an effort to address this problem.

The Si micropackage is a concept utilizing emerging fabrication technologies like micromachining[4] with design techniques for MMICs to produce individually packaged HF interconnects or components. The Si micropackage can also be used to build smaller MMIC subsystems such as low-noise amplifiers (LNA) that are essential elements in most low-power communication systems. While high-quality compound semiconductor materials like gallium arsenide (GaAs) or indium phosphide (InP) generally are used to develop high-performance active high-speed circuits, the development cost of the circuit can be expensive as a result of the costly material growth and fabrication of the III-V materials. An alternative approach to using a single compound exclusively in device and circuit development is to use a hybrid packaging design approach

that is commensurate with the Si-micropackage concept. For this approach, the active elements are batch fabricated. The known good die are then selected and attached to less expensive low-loss host substrates that contain the distribution and interconnect lines as well as passive components. Today, hybrid designs offer tremendous cost savings because of advanced bonding techniques, like flip-chip bonding. In addition, less-expensive high-quality materials, like Si, are available for processing in large, well-established microelectronics infrastructures. The development of the Si micropackage, therefore, and the inclusion of this micropackage in a number of communication circuits offer an excellent approach to achieve highly dense HF circuits that use Si and GaAs platforms in aerospace-based applications.

13.2.2 The Micropackaging Concept

In 1993, the first micropackage[5] was introduced with the coplanar waveguide (CPW) line (see Fig. 13.5) that offered a shielded environment above and below the signal line. The size of the micropackage, which can be chosen so that it is large enough not to interfere with the em field propagating in the planar transmission line design, is small enough to shift the package resonance above the frequency range of interest. The micropackage is defined by upper and lower shielding cavities that are placed in contact with each other to form a complete enclosure around the transmission line. The upper shielding is metallized and has a small air cavity region that is etched to a predetermined height using Si micromachining. The lower shielding is also metallized and has channels that are etched along both sides of the transmission line. When the two upper and lower shielding cavities are attached, the outer ground planes of the CPW line serves as the electrical conduit between the upper and lower cavities.

When the micropackage is implemented with different planar line geometries (e.g., microstrip, CPW), all the advantages of the design when it is not packaged are maintained. For example, the CPW offers uniplanar processing and ease of integration with two- and three-terminal active devices. The addition of shielding to the planar line offers lower radiation losses and reduced parasitic coupling to the fully or half-shielded planar line geometry. These losses and coupling would ordinarily exist in unshielded designs and affect neighboring components. A variety of planar components have been studied using this integrated micropackaging approach. For example, a 20-dB directional coupler that is printed on a dielectric membrane* has an air dielectric substrate and an integrated micropackage. The electrical performance shows an insertion loss of less than 0.4 dB over a 40-GHz bandwidth.[6] A variety of filter designs have also been implemented for operation up to 94 GHz. The shielded microwave low-pass filter design in Ref. 5 shows a −25 dB insertion loss in the cutoff region at 35 GHz with a response nearly identical to that of a theoretical model. This results from the significant reduction in radiation loss occurring because of the inclusion of the integrated package. At W-band (94–110 GHz), a dielectric membrane bandpass filter exhibits excellent filter performance with 8.5% bandwidth for a five-element design centered at 94.7 GHz. In this case, the losses observed are primarily a result of conductor loss, since the dielectric material has been removed and radiation losses have been suppressed.[7]

Because planar circuit designs often include conducting lines that bend and intersect, the micropackage concept was also investigated as a conformal design that follows individual conducting line paths. The first conformal micropackage was demonstrated in a detector design in 1995,[8] where discrete diodes were mounted onto an Si passive circuit and package. Extension of the micropackage to a conformal design involves the shaping of the shielding cavity so that it follows

*A dielectric membrane is a thin dielectric membrane consisting of a $SiO_2/Si_3N_4/SiO_2$ composite structure that is approximately 1.5 μm thick.

the conducting line. This shaping is easily accomplished because fabrication of the cavity is performed using lithographic and micromachining techniques common to Si processing with standard MMIC-based fabrication. This conformal design approach has benefits such as increased packing density of individually shielded circuits, which allows the circuits to be placed much closer together than is ordinarily possible in open structures.

In the following sections, the micropackage concept will be illustrated in a number of circuit designs to show the merits and flexibility that emerging technologies can offer HF design approaches. While other approaches are available for multichip module designs, these sections will concentrate primarily on advanced packaging based on Si micromachining. The general fabrication approach that is common to each of the demonstrated designs is described initially. This description is followed by the design and performance of several of the designs, including micropackages for high-isolation lines, conformal micropackages in MMIC designs, and a discrete micropackage design. The final section introduces a novel high-density filter design that uses micromachining to produce filters that are more compact than traditional design approaches for similar performance requirements.

13.3 Micropackage Fabrication and Testing Method

The implementation of the micropackage with HF circuits is complementary to the fabrication of microwave ICs. The issues pertinent to circuit fabrication are equally important to package development and address thin-film deposition, circuit patterning, metal-film deposition, and etching. The following sections focus on the general fabrication of an integrated package with an arbitrary planar circuit design that is either half or fully shielded (see Fig. 13.5). The micropackage is divided into an upper and lower shielding cavity that can be combined to form a fully packaged design, referred to as self-packaging. Any alterations to this basic package concept will be described in sections on specific examples.

The processing steps in the fabrication of the micropackage and high-density circuit designs can encompass a variety of VLSI Si fabrication techniques. Both the circuit and the package require the use of dielectric films to provide isolation, as well as masking material for the various etching steps. In the micropackage designs presented, dielectric films with silicon dioxide (SiO_2)

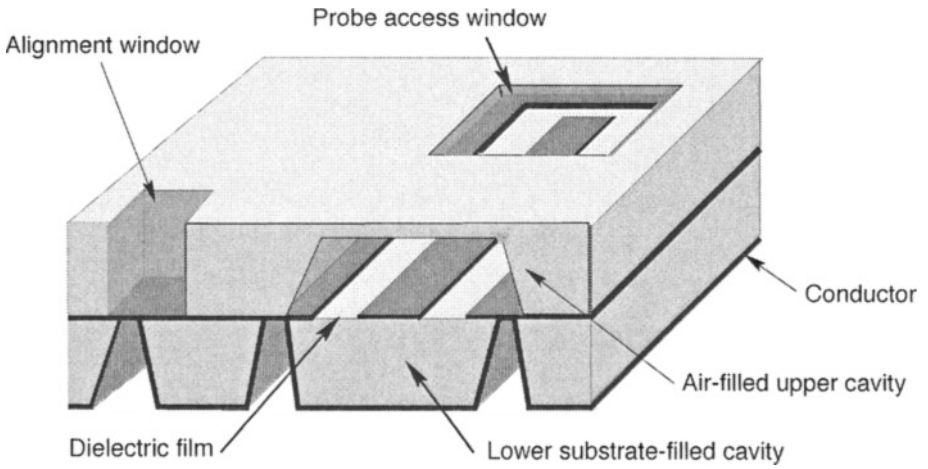


Fig. 13.5. Completely shielded micropackaged circuit with lower and upper wafer alignment. The figure illustrates the attachment of two wafers: one that supports the planar microstrip line and the other that is the air region cavity.

or silicon nitride (Si_3N_4) can be used in a single layer (7500 Å of SiO_2) or in combined layers (7500/3500/4500 Å of $SiO_2/Si_3N_4/SiO_2$, respectively) for processing flexibility. The Si wafers may be polished on one or two sides, referred to as single-sided polished (ssp) or double-sided polished (dsp), with a crystal plane orientation of <100>. The use of anisotropic wet etches such as potassium hydroxide (KOH) produces sidewall angles of 54.74 deg off the surface perpendicular to the wafer. Although dsp wafers may be slightly more expensive, they offer the most flexibility and ease in fabrication for portions of the micropackage requiring dual-sided processing. In addition, with dsp wafers, the fabrication of circuit designs requiring dual-sided processing show noticeably better etching quality and alignment between both sides compared with processes done on ssp wafers. If electrical circuits are to be printed on the wafer, the Si must have a resistivity of at least 1500 Ω-cm in order to reduce the losses caused by RF energy leakage into the substrate. If the wafer is to serve primarily as a shielding wafer that supports ground conductors, low-resistivity wafers (≤10 Ω-cm) based on the Czochralski method[9] are acceptable and produce good electrical results.

The examples of micropackage designs demonstrated in the following discussion reflect a small number of possible micropackage arrangements. Fabrication methods described in this section have proven successful even though new emerging technologies may exist to date that overcome some of the limitations indicated. For a more thorough review of Si processing and micromachining, see Wolf and Tauber[10] and Rai-Choudhury.[11]

13.3.1 The Micropackage

13.3.1.1 Upper Shielding Cavity

The cavity dimensions for the upper wafer [Fig. 13.6 (a)] in the micropackage can be determined using full-wave CAD tools based on quasi-static and full-wave analysis techniques. A suite of design and analysis tools has been developed by Hewlett Packard (HP). The dimensions of the different planar transmission line geometries can be determined using the tool, LineCalc.™ Both the HP EESof High-Frequency Design Solutions Series IV (i.e., electronic design automation toolset) and HP Microwave Design System (MDS) were available for the simulation and analysis of the transmission line-based planar circuits. In cases where complicated designs include conductors and dielectrics with arbitrary shapes, finite element method (FEM) or finite difference time domain (FDTD) based CAD tools, are used, where HP has a FEM tool, called High-Frequency Structure Simulator (HFSS).[12] In this work, in-house FDTD programs were used primarily to model the 3D structures of the packaged planar lines. Because the package designs in this work were used as shielding, the cavity heights in the upper and lower regions were designed to be far enough away from the conducting lines to not interfere with the em fields. For the work described here, the diagnostics have been performed via microwave on-wafer probing methods. These methods require the inclusion of steep-angle probe windows for access to the feed point of the circuit from the opening in the upper shielding cavity. The fabrication of this wafer involves thin-film deposition, photolithography for pattern definition, metal deposition, metal etching, and Si anisotropic etching.

To form the upper cavity wafer, a thin oxide or nitride film (SiO_2 or Si_3N_4) should be deposited onto a low-resistivity wafer. The oxide film can be based on thermal SiO_2 deposition procedures,[10] with typical oxidation temperatures of 1100°C for high-quality films that are 7500 Å thick. An alternative to the oxide thin film is the use of low-stress nitrides that can be formed with low-pressure chemical vapor deposition (LPCVD). If the thin-film layers used as masking material are also used as electrical insulation layers, it is important to maintain very clean deposition conditions. Otherwise, the result may be cross-contamination with elements such as boron (B),

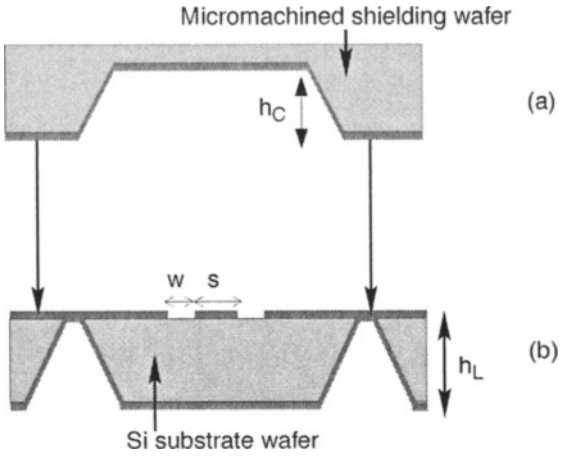


Fig. 13.6. (a) Upper shielded cavity with height (h_C), and (b) lower shielded cavity with height (h_L), conductor width (s), and slot width (w).

maintained during anisotropic etching to ensure that the final cavity dimensions of the wafer are consistent with the design. The order of the fabrication steps can vary depending on the dielectric film combination. One example of a general process for wafer fabrication is given in Table 13.1.

Table 13.1. Upper Cavity Development

General Process Outline
1. Clean wafer with a "piranha" etch solution (1:1 ratio of H_2SO_4:H_2O_2) and dehydrate bake the wafer.
2. Evaporate a thin metal film (e.g., chromium [Cr]/gold [Au]:500/1500 Å) on the dielectric film of the polished side of the wafer (Side 1).
3. Apply a protective layer of photoresist on the other side of the wafer (Side 2).
4. Apply and develop photoresist to define the probe access windows and cavity region to be etched on Side 1.
5. Remove the metal and dielectric film of the patterned region using a dry or wet process.
6. After performing a solvent clean (e.g., with acetone and isopropyl alcohol) of the wafer, partially etch the cavity and probe access regions to a predetermined height using an anisotropic etchant (KOH, ethylene diamine pyracatechol [EDP], or tetra ammonium hydroxide [TMAH] solutions).
7. Repeat Step 3 on Side 2.
8. Repeat Step 4 on Side 2 to open the probe access windows on the back side using infrared (IR) alignment.
9. Remove the dielectric film in this defined area, return the sample to the etch, and completely etch the remainder of the cavity and the opening to the access windows.
10. Strip the metal and dielectric films on the cavity region side of the wafer using the appropriate etchants.
11. Evaporate or electroplate a thick metal film for the shielding cavity, e.g., titanium (Ti)/aluminum (Al)/Ti/Au (500/15 K/500/4 kÅ) or Cr/Au (500/20 kÅ). This film may also be sputtered for good step coverage along the sloping sidewalls.

In this example, it is assumed that the films have been deposited using one of the dielectric film deposition methods mentioned previously.

13.3.1.2 Lower Shielding Cavity Formation

The lower shielding cavity [see Fig. 13.6(b)] can be developed in a similar manner to the upper shielding cavity, except that the circuit design must be taken into consideration. This substrate typically has high resistivity, and the process steps are specific to each circuit design. The process details will be given in a separate section, but the logistics of circuit fabrication will be described here. The main issue to consider is the requirement to maintain adequate ground plane continuity between the outer lower shield and the circuit. Since these circuit designs are tested using on-wafer probing techniques, a CPW-based probe point is utilized in all designs (see Fig. 13.7).

The lower shielding cavity essentially provides an outer shield to the planar transmission geometry that completely encases the transmission line design. This shield may or may not contain a dielectric substrate, depending on the desired substrate dielectric constant in the design. Assuming the cavity is filled with Si substrate, the aim is to maintain a constant ground signal between the outer shield of the package and the measurement test system. This aim can be realized by including ground planes on the upper surface of the circuit wafer and by ensuring electrical continuity through vias that have been extended to form channels along either side of the line. The ground plane on the upper surface of the wafer can then be put in contact with the metallization on the lower surface of the wafer, thereby making a continuous ground plane. To accomplish this

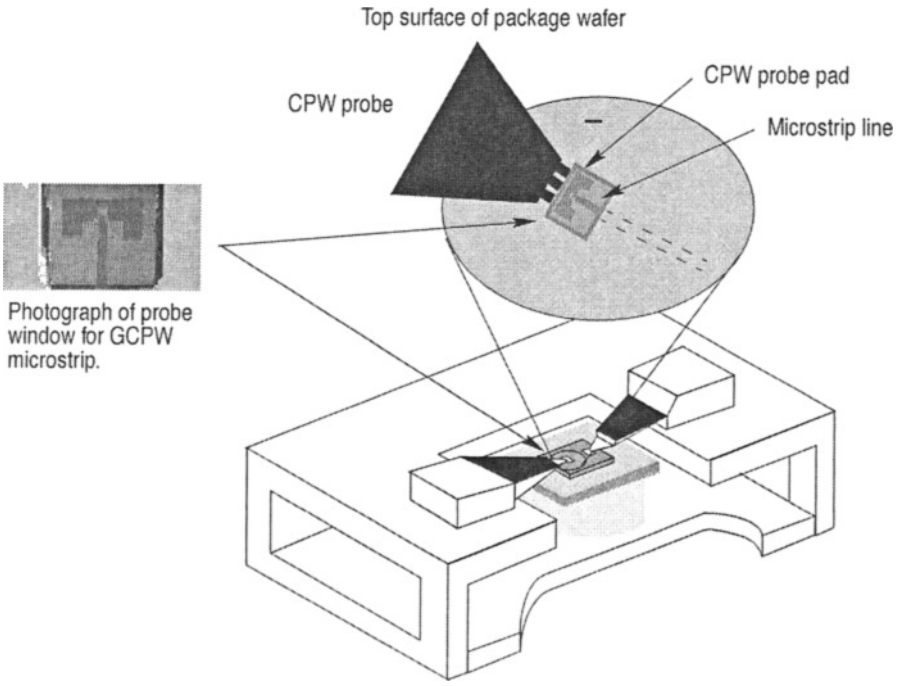


Fig. 13.7. Probe station measurement of the integrated package design. Each package design consists of two wafers. CPW probes are used to excite the circuits, where each circuit has a CPW probe pad accessible through a probe window on the top surface of the lower wafer. The pad allows for em field transitions from the CPW probes into the respective transmission line geometry (e.g., microstrip, CPW).

continuous ground plane, narrow regions of dielectric film must be removed on the top surface of the wafer in the ground plane areas prior to printing the circuits. This step guarantees that the metallized etched vias will be able to make contact to the upper ground. When microstrip transmission lines are used, ground planes must be added on the upper surface of the bottom wafer. These ground planes should be sufficiently far from the center conducting lines so that they do not disturb the microstrip mode of propagation (see Fig. 13.8). Table 13.2 outlines the general process for lower cavity development.

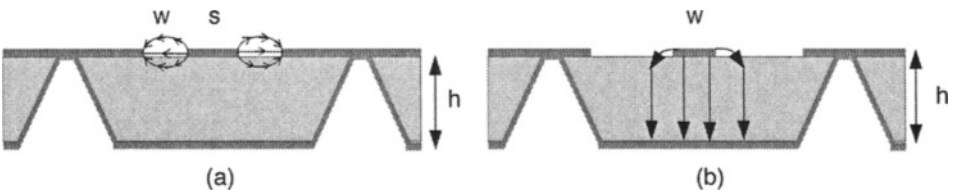


Fig. 13.8. Micropackaged transmission lines. (a) CPW line and (b) microstrip line, where w is a sufficiently large support microstrip mode.

13.3.1.3 Package Assembly and Testing

The micropackage is fabricated using one or more of the fabrication processes just described. If a fully packaged design (Fig. 13.5) is desired, the upper and lower cavities are aligned and secured using conducting epoxy[13] or wafer-bonding techniques. "On-wafer" testing methods are used to accurately characterize the packaged design. These methods are reliable for frequencies up to 110 GHz.

The experimental setup used in this work consisted of CPW probes, manufactured by GGB PicoProbes,[14] which have three fingers that are separated with a pitch spacing (or center-to-center spacing) of 150 μm. A Cascade or Alessi HF probe station is used with an HP 8510C network analyzer that can test from 2 to 110 GHz. The test probe effects are de-embedded (i.e., removed) from the measurement with calibration methods such as through-reflect-line (TRL), short-open-load-through (SOLT), or line-reflect-match (LRM) calibration. The test results shown in the following sections were collected using the TRL method implemented by "De-embed: Multical" software provided by the National Institute of Standards and Technology (NIST).[15,16]

The measurement data for the HF circuits herein will be presented in terms of scattering parameters. Because current and voltage cannot be measured directly on distributed circuits, a circuit response is commonly evaluated by measuring the reflected power at the input and transmitted power to the output of a circuit. Scattering parameters, [S], describe a relationship between the amount of energy that is reflected at one port to the amount of energy that propagates through the circuit when all other ports are matched [see Eq. (13.1)].

$$[S]_{ij} = \left.\frac{V_i^{reflected}}{V_j^{incident}}\right|_{V_k^{incident} for (k \neq j)} \tag{13.1}$$

where j represents the incident port, i represents the output port, and k represents all other ports. A requirement for the S-matrix is that all other ports are matched and that only one port has an incoming signal.[17]

Table 13.2. Lower Cavity Development

General Process Outline
1. Clean the wafer with a "piranha" etch solution (1:1 ratio of H_2SO_4:H_2O_2) and dehydrate bake the wafer.
2. Apply a protective layer of photoresist on nonpolished side of the wafer (Side 1).
3. Using photolithography, apply and develop the photoresist to define the narrow via channels to be etched in the dielectric film on the circuit side of the wafer (Side 2).
4. Perform a solvent clean on the wafer (e.g., with acetone and isopropyl alcohol) and dehydrate bake the wafer.
5. Evaporate a thin metal film (e.g., Cr/Au:500/1500 Å) on the dielectric surface of Side 2.
6. Etch the alignment openings through the metal on Side 2 for IR alignment of the back side of the wafer to the front side of the wafer.
7. Fabricate the specific circuit design using the appropriate process steps. The circuit can be printed and electroplated at this time.
8. Reapply a protective layer of photoresist on the metallic side of the wafer (Side 2).
9. Using photolithography, apply and develop the photoresist to define the cavity region on Side 2 of the circuit wafer using IR alignment.
10. Etch away the dielectric film of the patterned design using a dry or wet process from Side 1.
11. Reclean the wafer with solvents (e.g., with acetone and isopropyl alcohol) and etch the lower cavity and vias in an anisotropic etchant (KOH, EDP, or TMAH solution) through the entire wafer.
12. Repattern and etch the dielectric films out of the cavity region of Side 1, and remove the metal seed-layer from Side 2.
13. Evaporate or electroplate a thick metal film on the lower shielding cavity, e.g., Ti/Al/Ti/Au (500/15 K/500/4 kÅ) or Cr/Au (500/20 kÅ). This film may also be sputtered for good step coverage.

In an ideal system with no losses, the sum total of the reflected energy and the transmitted energy would equal the total power available at the input of the circuit. In real systems, which include losses, the amount of energy that is available to propagate through the circuit experience attenuation due to the losses in the conductor and the dielectric material, and because of radiation. In the work shown here, reduction or elimination of radiation losses are the primary objective.

Scattering parameters (s-parameters) can be determined according the number of ports of a given circuit system. If the system has an one input port and one output port, we have a two-port circuit. The scattering parameters essentially represent the ratio of power reflected at a given port compared with the power incident to a given port. When the circuits of interest are evaluated in terms of two ports, the respective reflection parameters are denoted as S_{11} and S_{22} and the transmitted or insertion parameters are denoted by S_{12} and S_{21}, where the subscripts are S_{ij} represent the port numbers.

13.3.2 The Conformal Micropackage

Conformal micropackages can be used to optimize space utilization in the development of high-density circuits. Several challenges must be addressed when implementing a conformal package around transmission line circuits that exhibit curves and bends. In Fig. 13.9, top and bottom views of a circuit and cavity geometries are shown for the planar distribution line shaped like a "U." In this arrangement of circuit and cavity, the corners of the package are potential candidates for severe rounding during the wet-etch process, as a result of undercutting. Since wet anisotropic etchants selectively attack the various crystal planes at different rates, it is difficult to achieve a good etch stop in the absence of orthogonal planes without the use of some compensation mechanism. This mechanism is described in more detail in Section 13.4.2.2.

Bean,[18] and Abu-Zeid-Corner and Abu-Zeid,[19] show the anisotropic etching of corners that are classified as concave* or convex† (see Fig. 13.9). The results show that each corner type is bounded by crystal planes that etch at different rates. Wu and Ko found that convex surfaces are bounded by the fastest etching planes, while concave corners are bounded by the slowest etching planes.[20] As a result, concave corners are easily formed without undercutting, and convex corners typically exhibit undercutting that can only be reduced with the incorporation of appropriately sized compensation geometries or shapes located at the respective corners. In the individual circuit examples, specific corner compensations have been incorporated to provide the best model for the conformal package employed.

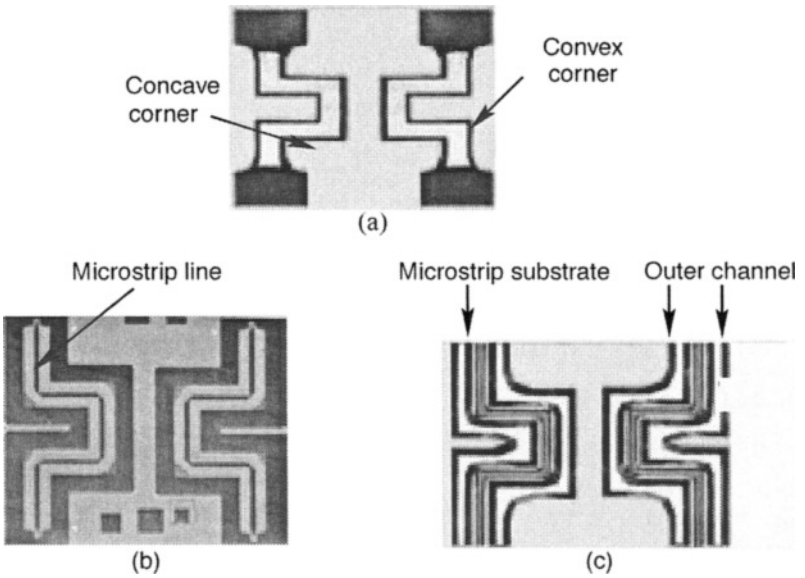


Fig. 13.9. Micropackaged interconnect. (a) Upper cavity package with probe windows (shown in black) for testing; (b) microstrip interconnect lines, where the conductors appear as the thin dark colored lines; (c) lower cavity package, showing reduced-thickness region in the center of the U-shaped line and lower package channels in the outer white regions.

*Concave (inside) corners are formed when two <110> crystal planes intersect to produce an interface that points inward.

†Convex (outside) corners are formed when two <110> crystal planes intersect to produce an interface that points outward.

13.4 Micropackaged High-Isolation Interconnects

As microstrip lines are placed in closer proximity to each other, em signals from individual lines can interact with each other, causing cross talk. Cross-talk effects are studied by incorporating the micropackages with the microstrip line and analyzing the em interactions between adjacent lines. The objective is to understand how the micropackage can be used to reduce the interactions between lines that tend to radiate. Coupling can occur either through direct signal interactions or from the radiative excitation of substrate modes associated with discontinuities. Coupling between two lines in an open environment can result in poor signal clarity and distortion. The basic circuit configuration, shown in Fig. 13.10, exhibits a microstrip line that has a micropackage around it. In this figure, the microstrip resides on a reduced thickness substrate, which will be discussed in more detail in Sec. 13.4.2.1. The effects of cross talk are studied for interactions between U-shaped or bending microstrip lines and straight ones, in open and shielded environments.

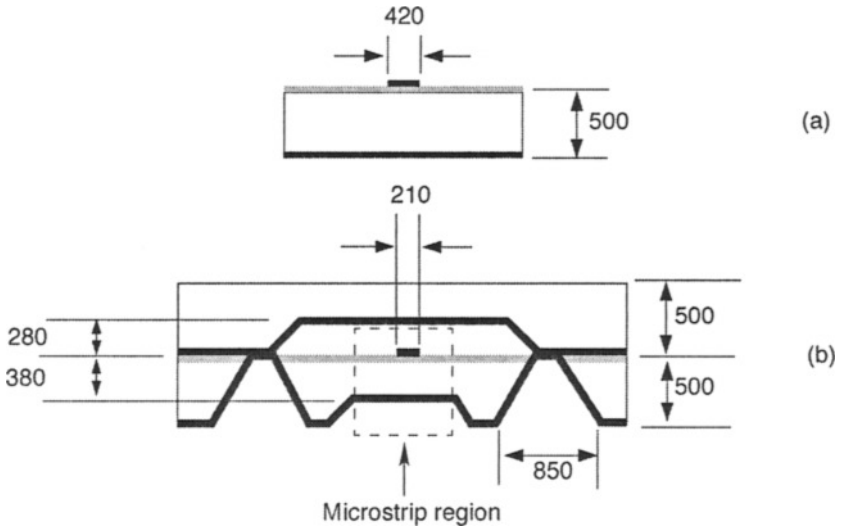


Fig. 13.10. Circuit topology (numbers given in microns). (a) Open microstrip on full thickness wafer and (b) packaged microstrip on reduced thickness wafer. The microstrip circuits are printed on a membrane dielectric that is located on the reduced thickness region of the lower wafer surface.

13.4.1 Design Considerations

Shielding effects and proximity coupling can be evaluated by comparing the electrical performance of interconnects in open and shielded environments, and by evaluating the interactions between various layouts of straight and U-shaped microstrip lines. In Fig. 13.11, the layout of straight and meandering lines is used to study the proximity effects between the 50-Ω lines that were developed with HP's Line-Calc and Libra Series 4 (CAD tools).[21] For the packaged version of the layout, in-house full-wave CAD tools were used to create a design with microstrip lines that did not interfere with signal propagation. Analysis of the structure was accomplished with a point-matching-method algorithm. The algorithm included both upper and lower shielding cavity effects, which resulted in characteristic impedance and effective permittivity data.[22] Similar modeling can be done using commercially available tools, such as HP's HFSS, to simulate the response of the circuit with the package. The dimensions for the planar circuit can be obtained with HP's LineCalc. Figure 13.11(a) (Design Layout A) consists of a single-section meander of

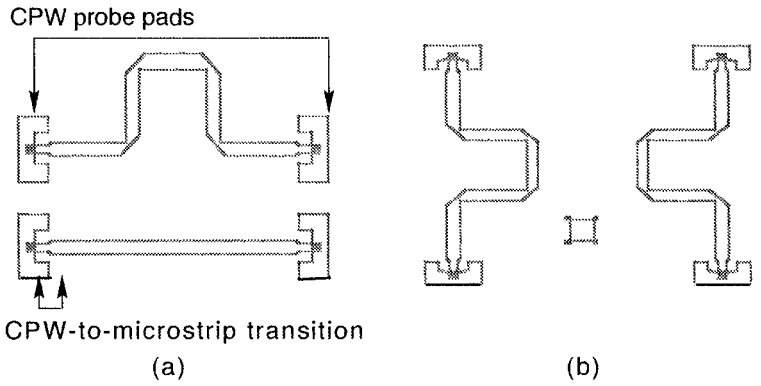


Fig. 13.11. Layout configurations. (a) Design Layout A: through-line and U-shaped (back-to-back right angle bend) line. (b) Design Layout B: two U-shaped lines.

a back-to-back right-angle bend (U-shaped line) and a through-line, and Fig. 13.11(b) (Design Layout B) consists of two U-shaped lines. These layouts have been chosen to evaluate the amount of coupling from the bends to an adjacent line resulting from radiation. In Fig.13.11(a), the coupling is assessed from the bend onto the through-line, while in Fig. 13.11(b), the coupling is evaluated between two lines of similar shape.

The microstrip lines are packaged using the micromachining approach developed in earlier work[5] for self-packaged CPW structures. In this case, the microstrip lines include ground planes on the conductor surface that are placed far enough (380 μm) from the signal conductor to maintain a microstrip mode of propagation and also to offer an electrical continuity between the upper and lower package for ground plane equalization. Without this continuity, the potential between the upper cavity and the lower one is not common and cannot suppress the RF leakage that occurs when direct contact is not made between the two conductors.[5]

13.4.2 Fabrication Challenges

13.4.2.1 Substrate Thickness Reduction

The micropackaged interconnects shown in Fig. 13.10 employ both a package and reduced-thickness region under the microstrip line. Interconnect lines should propagate the dominant mode only; therefore, care must be taken to reduce the potential for dispersion and higher order mode excitation. A 500-μm-thick substrate has been reduced locally to 380 μm to ensure that the dominant microstrip mode propagates in the K-band frequency range. The wafer thickness reduction requires a two-step etch process to accommodate the deep etch requirement of the vias and package channels as well as the shallow etch requirement for the reduced-thickness region.

In the first step of this process, vias and channels are etched several hours to remove approximately 400 μm of material while the reduced-thickness regions remain protected. In the second step, the protected regions are opened and etched an additional 2 h to remove 170 μm of material while the via and lower package channels are etched entirely through. In Fig. 13.12, a scanning electron microscope (SEM) picture is shown of an inverted lower cavity package with a cross-section view of a via and a lower package cavity.

13.4.2.2 Conformal Micropackaging with Reduced Thickness Regions

Conformal packages can be used to optimize space utilization in the development of high-density circuits. The fabrication procedure is similar to the one in Sec. 13.3, except the substrate that holds the microstrip conductor is reduced in thickness. This reduction requires additional

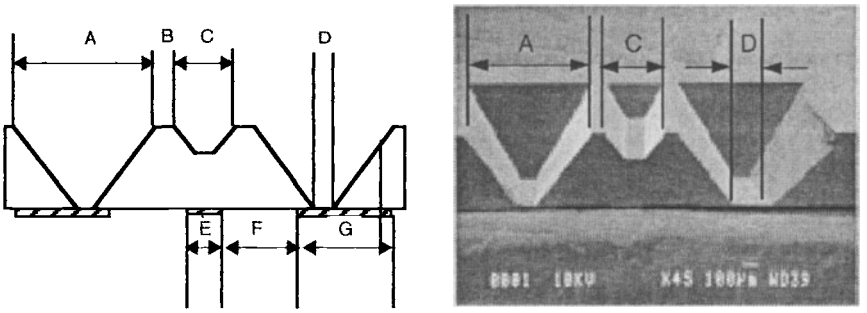


Fig. 13.12. (a) Front view drawing of a microstrip printed on an inverted lower package with dimensions of A = 850, B = 100, C = 400, D = 150, E = 210, F = 380, G = 750, and conductors that are indicated as hashed lines. All dimensions are in microns. (b) SEM photo of the inverted lower package developed during the two-step etch procedure.

design techniques to account for the conformal package formation around the reduced-thickness region. The etched channel that forms the lower package follows along the center conductor, while a partially etched channel is placed underneath the center conductor to reduce the wafer thickness. This type of lower package requires two kinds of compensating geometries—one set for the lower package channel and one for the reduced-thickness region. The geometries compensate for corner definition problems that arise from a variation in the lines in the two etched sections. From the processing standpoint, the compensation geometries are incorporated in the mask layout of the etch channels. At each convex corner of the original design, a square of the appropriate size is centered. The addition of this square retards the rate of the chemical etching in the formation of the corners so it is similar to the etch rate of the preferred <111> plane that forms the channel sidewalls. As an example, Fig. 13.13 shows an illustration of the bottom surface of the U-shaped lower wafer layout. Step 1 shows the original layout without compensation, where the channels to be etched are shown in grey. When this layout is etched, the convex corners in the cavity experience extreme overetching, because the etch rate for the corners is faster than the rate to define the channel depth. Step 2 is the original layout with compensation squares to protect the convex corners (shown in black) of the channels to be etched (shown in grey). Once the channels are etched, the corners will etch more slowly, as shown in Step 3 of Fig. 13.13. The lines that appear to be cracks are reflections of light off the acute corner surfaces. When the desired channel depth has been reached, as seen in Step 4, there is still some rounding along the corner surfaces. However, these corners do not interfere with the reduced thickness region located underneath the microstrip line shown in Fig. 13.9. Through an iterative process, the best square dimensions were found to be approximately 1.4 times the desired etched depth that produces corners similar to those shown in the upper cavity package of Fig. 13.6. The two different etch depths require two sets of convex corner compensation squares to produce a desirable corner shape. Hence, each square is designed to accommodate the via depth of 550 μm and the shallow region depth of 170 μm in the conformal package.[23]

13.4.3 Electrical Characterization

In characterizing the package response, a comparison is made between the packaged and unpackaged U-shaped line. Because substrate mode excitation does occur when there is radiation into the substrate, the package material may exhibit increased electrical noise if it is not completely isolated. Therefore, this package noise is evaluated independently to identify the potential sources

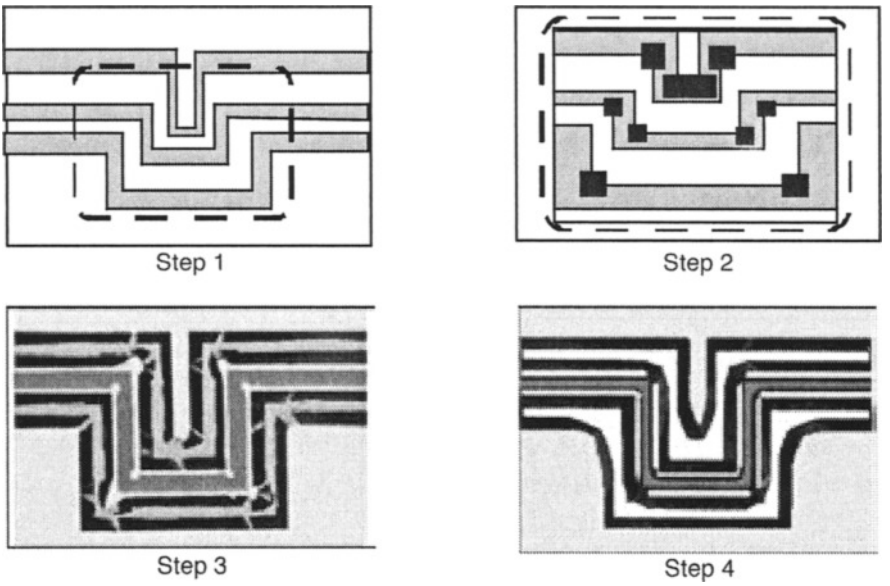


Fig. 13.13. Process steps for the development of the lower package wafer. Step 1: mask layout for the channels (shown in grey) to be etched on the lower surface of the circuit wafer. Step 2: close-up of mask layout with the square compensation geometries (shown in black) added to protect the convex corners of the channels. (Note: The white regions will not be etched. The squares are added to protect the corner sections of the nonetched region [shown in white].) Step 3: photograph of the bottom of a partially etched lower package wafer. (Note: the outer channels have been etched, while the center portion, which shows square patterns in a dielectric film in this region, has not been etched. The corners of this partially etched outer channel are acute and show high reflection of light that appear as be cracks in the corners regions.) Step 4: the lower package after the etch is complete. The outer package channels are shown in white, the reduced thickness region are shown in dark grey, and the sloping sidewalls of the channels are shown in black.

of leakage currents that may exist in the package material and to assess the impact of these sources on packaging design and background noise. Therefore, noise coupling is also measured in the two layouts shown in Fig. 13.11 for both the packaged and unpackaged arrangement.

13.4.3.1 Radiation Loss Reduction

A SOLT calibration technique is used to evaluate a straight microstrip delay line of 13.392 mm and a U-shaped bend of 13.392 mm in the open and packaged environments. The experiment reveals the comparative radiation loss between an open U-shaped bend and a straight microstrip line (Fig. 13.14). Typically, radiation loss can become a large problem for planar circuits above 20 GHz. The experimental results show that above 20 GHz, radiation from the corners of the U-shaped lines excite substrate modes. The effect is an increase in insertion loss that causes the signal to exhibit an oscillation in the insertion loss data because of multimode propagation. Note that this effect becomes magnified in the calculation of the total loss.

When an on-wafer package is included with the bend structure, the effect of radiation is greatly suppressed, as shown in Fig. 13.15 with oscillations absent. Furthermore, when the straight microstrip line and U-shaped bends are packaged, the total loss becomes similar, indicating that radiation effects have been minimized. Total loss is defined as the normalized input power to the line minus reflected power from the line minus transmitted power along the line to the output $(1 - |S_{11}|^2 - |S_{12}|^2)$. From the calculation, the total loss observed in the open U-shaped line can be

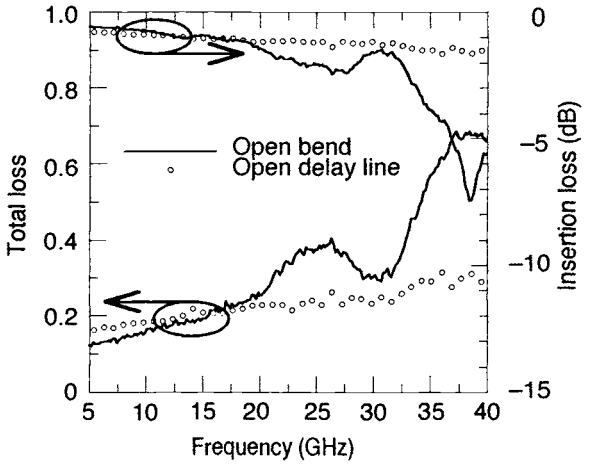


Fig. 13.14. Electrical response of open microstrip circuits in for a delay line and a U-shaped bend of similar length.

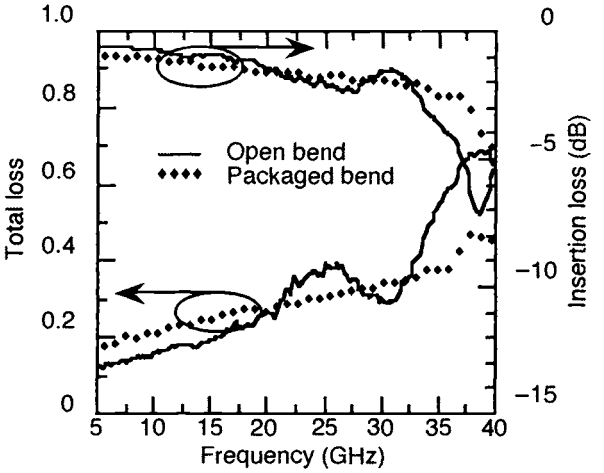


Fig. 13.15. Electrical response of an open and packaged U-shaped microstrip circuit.

as high as 65%, which is more than double the loss associated with the straight microstrip line. In low-power applications, this amount of loss can have a substantial effect on power consumption and would require additional components for signal amplification to overcome these losses.

A comparison is also made between the open and packaged bend design. Figure 13.15 shows that as frequency increases, the open bend performance degrades rapidly. This degradation occurs because of parasitic radiation that affects the insertion loss and the total loss associated with the open bend. In comparison, the packaged bend has a much flatter response in the insertion loss and total loss data. Even though these values are slightly higher than those observed in the straight microstrip line, the overall performance of the packaged bend is much better compared with that of the open U-shaped circuit. The higher loss in this case is associated with the additional ohmic loss introduced by the top metallization layer of the package. Radiation from the bend, on the other hand, has been eliminated in the packaged U-shaped design and shows a total loss calculation similar to that of a straight microstrip line of similar length (see Fig. 13.16).

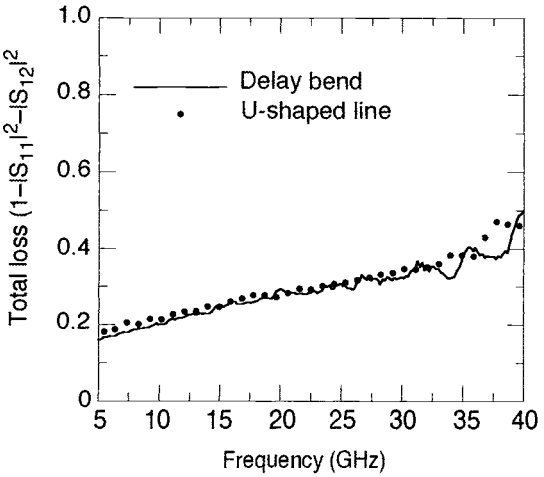


Fig. 13.16. Electrical response of a packaged U-shaped bend and microstrip delay line circuit of similar length.

13.4.3.2 Cross-Coupling Evaluation

In open environment signals, cross-coupling is strongly dependent on the interconnect layout. In Fig. 13.17, cross coupling is observed to be as high as 20 dB[13.11(a)] in the midfrequency range for the U-shaped bend and the straight microstrip line used in Layout A. In Layout B, this cross coupling is as high as –10 dB in the upper frequency range of the two U-shaped lines. With the inclusion of the integrated package, Fig. 13.18 shows a reduction in the coupling by –20 dB in Layout B with half-shielding and a reduction in the coupling by 10 dB when the line is completely packaged. These results indicate that the radiation from the corners not only excites leakage into the substrate but into the air as well. By incorporating the micromachined package, each microstrip line is isolated, and the leakage currents are shorted to ground such that the maximum coupling above 33 GHz is –45 dB compared with the –10 dB for the open environment U-shaped line.

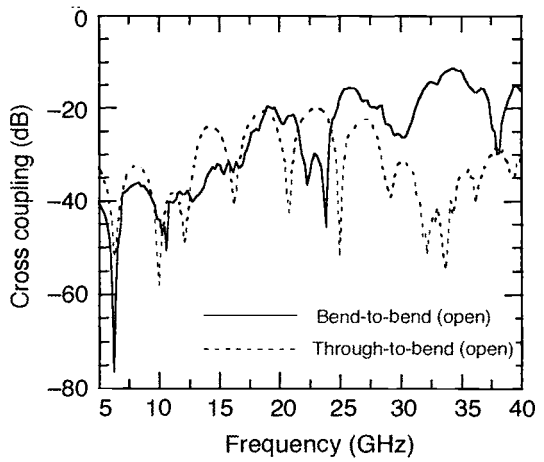


Fig. 13.17. Comparison of cross-coupling effects in open microstrip structures for Design Layout A between the through- and U-shaped bend and Design Layout B between the two U-shaped bends.

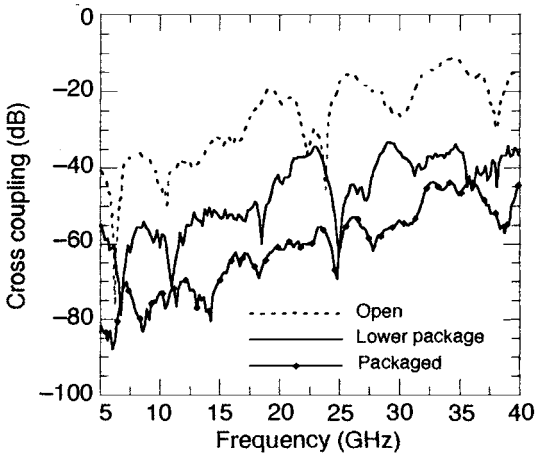


Fig. 13.18. Cross-coupling effects in Design Layout B for two U-shaped bends in open, lower half-packaged, and full-packaged designs.

13.4.3.3 Noise Characterization

Coupling mechanisms associated with different circuit layouts were evaluated and then compared to the minimum reference noise value of the packaged system. Because several circuits were evaluated, the data presented reflects an average value.

Two types of noise measurements are made with the two U-shaped meanders shown in Layout B (see 13.11). One is a "contact" measurement: the measurement probe is placed in physical contact with the top of the package. The other is a "noncontact" measurement: the probes are elevated 15 mm above the circuit surface, with an identical separation distance of 8.38 mm between the two U-shaped circuits in Layout B. The contact noise is the excitation of the line at one port and the detection of residual transmitted signals in the surrounding package structure. These signals are measured at different locations atop the circuit package (see Fig. 13.19). The measurement

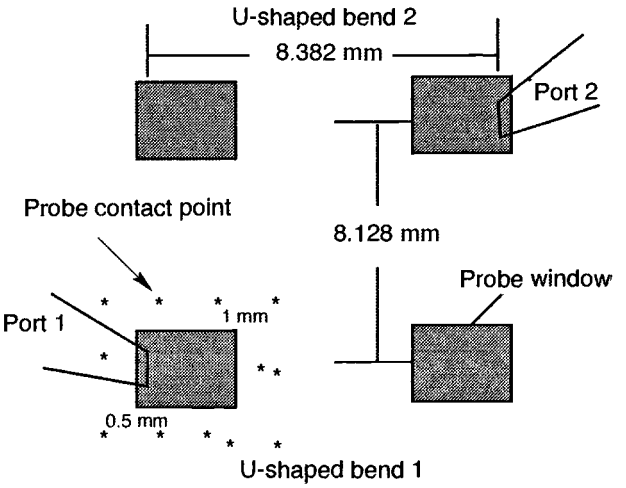


Fig. 13.19. Probe window layouts (shown by grey blocks) in design layout B for the two U-shaped bends. Star points represent where the probe makes contact with the upper wafer surface. Probing points in the contact noise measurement are indicated by stars at Port 1 for an input signal onto the conducting line at Port 2.

reflects an average value of the data from the different locations. By placing the probe near the opening of the secondary cavity port, random propagating signals are detected on the upper side of the package and used to determine the maximum noise contribution from the package. Figure 13.20 shows that the contact and noncontact noise measurements are very similar, even though between 20 and 30 GHz, the contact noise tends to be slightly higher than 10 dB. When the coupling between the two on-wafer packaged U-shaped lines is compared, the results are nearly identical until about 30 GHz. While the package does provide shielding, it is not a hermetic seal. Therefore, at higher frequencies, leakage from the entrance and exit ports of the package may be responsible for the observed higher radiation leakage and the resultant higher coupling.

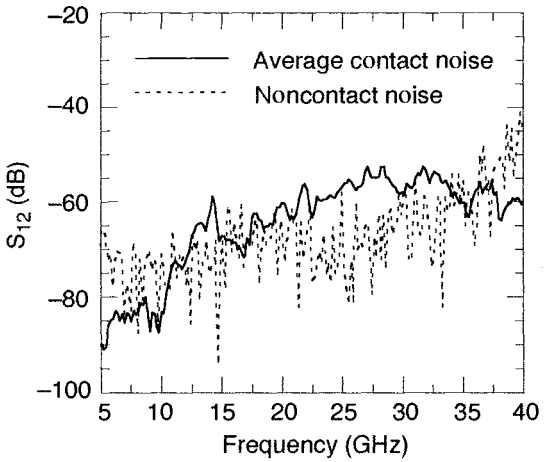


Fig. 13.20. Noise data from contact and noncontact measurements.

13.5 Conformally Micropackaged LNA

MMIC design and development cost can be reduced significantly by decreasing the number of design iterations and by limiting the use of expensive materials such as GaAs and InP. To illustrate this concept, a conformal Si micropackage is used in the design of a low-noise K-band amplifier (see Fig.13.21).

At the MMIC level, it is difficult to predict the final performance of a design at the packaged level. As with most planar circuits, parasitic effects—in this case, between neighboring MMICs—may require additional design iterations to recover lost performance as a result of chips placed in close proximity. With the integration of the micropackage into the MMIC design, parasitic effects are sharply reduced.

In this demonstration, flip-chip bonding[24] is incorporated into the design of the micropackaged LNA that is composed of discrete devices. The entire MMIC performance is based on optimization of individual elements, where the MMIC represents a combination of an Si motherboard for the passive components and distribution lines, a flip-chip bonded 20 GHz high-electron mobility transistor (HEMT) device, and a micromachined integrated package. The flip-chip bond of the device to the MMIC uses very reliable solder bump techniques that provide repeatable, low-inductance connections. In addition, the bond allows the circuit designs to be fabricated faster using rapid prototype services. With this approach, a device can even be replaced if failure occurs at a later time. As a result, expensive materials, like GaAs and InP can now be used primarily for active device fabrication rather than for both active device and circuit fabrication.

13.5.1 Amplifier Design

The amplifier uses an InP-based HEMT with a gate length of 0.15 mm and a total gate width of 300 mm. This HEMT was developed by Hughes Research Laboratories for a low-noise performance application.[25] The chip transistor is roughly 3 mm in size and is mounted on the amplifier circuit via tin (Sn)/lead (Pb) solder bumps that have been electroplated on the transistor contact pads. Figure 13.21 shows a close-up of the HEMT flip-chip device, with solder bumps that are 25 μm high and 50 μm in diameter on the gate, drain, and both of the source contact pads. Although

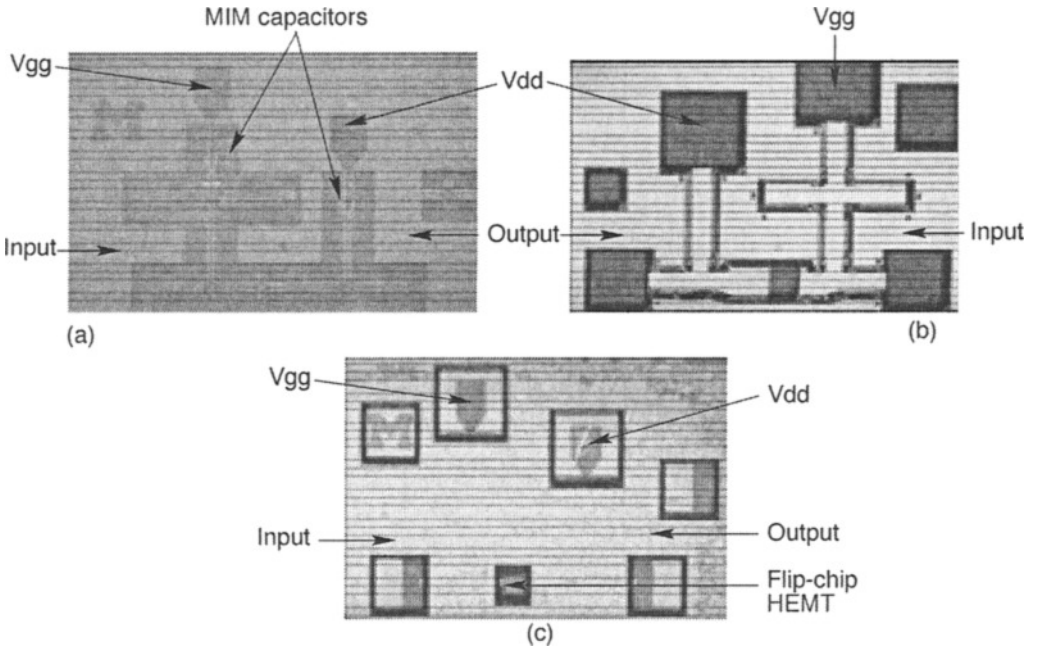


Fig. 13.21. (a) Layout of the 20-GHz LNA circuit (without InP flip-chip HEMT), (b) view from the bottom of the micromachined cavity after final metallization has been completed, (c) view of the LNA after enclosure with the micromachined shielding cavity. RF and dc bias probe pads, as well as the inverted flip-chip HEMT, are visible.

the original design (in a microstrip configuration on a 250-μm alumina substrate) of the amplifier circuit was optimized for low-noise performance, these design parameters were converted to a shielded CPW environment by maintaining equivalent impedances and electrical lengths for all transmission line sections. The Point Matching Method (PMM) algorithm[26] was used to generate the line dimensions for the shielded CPW lines. CPW impedance parameters are related to the dimensions of the conductor width (s) and the slot width (w). These dimensions are also chosen to provide a convenient interface to the gate and drain pad dimensions of the HEMT (Table 13.3).

Table 13.3. Circuit Dimensions[a] for the Shielded CPW

Shielded CPW Impedance	Conductor Width (s)	Slot Width (w)	s + 2w
50 Ω	94	53	200
75 Ω	34	83	200

[a]In microns.

The overall width of the cavity is chosen as 1 mm, which allows the incorporation of the flip-chip HEMT package into the conformal package.

The amplifier layout of the shielded CPW implementation is illustrated in Fig. 13.22. The circuit includes source feedback stubs to improve the low-noise impedance-matching condition, and an open circuit stub on the drain side of the transistor to improve output impedance matching. Lumped elements are integrated as thin-film resistors and metal-insulator-metal (MIM) capacitors, and the bias networks are designed to resonate at approximately 18.5 GHz. In the packaged amplifier, the gate bias network uses a balanced open stub configuration in place of the original radial stub design, and the drain bias stub is replaced by a 20-pF MIM capacitor.

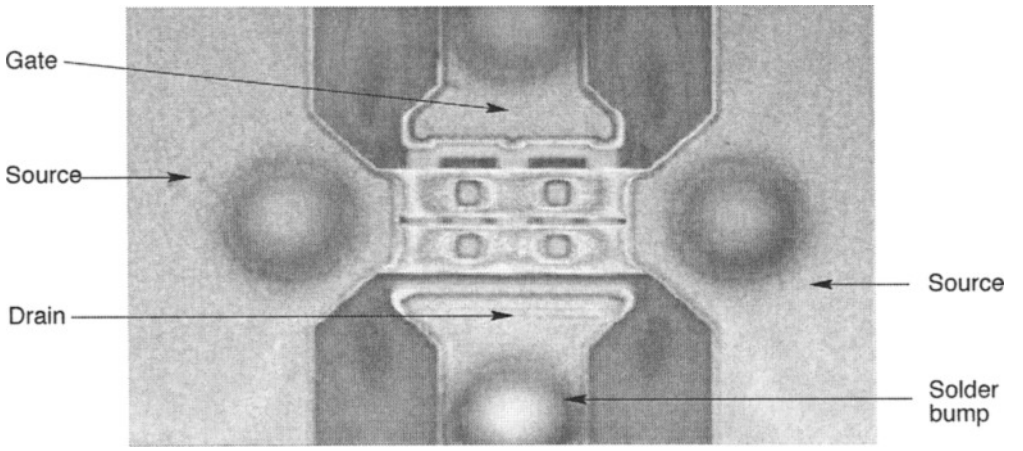


Fig. 13.22. Photograph of the flip-chip HEMT, showing the Sn/Pb solder bumps. In the packaged amplifier, the gate bias network uses a balanced open stub configuration in place of the original radial stub design, and the drain bias stub is replaced by a 20-pF MIM capacitor.

13.5.2 Fabrication Considerations

The fabrication process for the package follows these steps:

- The packaged amplifier is fabricated on a high-resistivity Si substrate with an 8000-Å-thick layer of thermal SiO_2.
- The amplifier circuit patterns are created with 3-mm-thick electroplated gold. Thin-film resistors are realized with a 400-Å Nichrome thin film with a sheet resistance of 33 W/square.
- MIM capacitors are constructed with an evaporated 1500-Å-thick alumina (Al_2O_3) film with a dielectric constant of approximately 8–10.
- Air bridges are formed with 2-mm-thick electroplated gold.

Access to the RF probe pads, the direct current (dc) bias contacts, and the flip-chip mounting locations is provided with etched through-windows. These windows are formed by etching from the back side of the cavity wafer during the conformal package creation. The depth of the shielding cavity is selected to be 275 μm, so that when etching is from both sides of the 550-μm-thick wafer, the access windows will open as soon as the cavities reach the proper depth. The final step in the package fabrication is the deposition of a 1.4-μm-thick Ti/Al/Ti/Au metallization layer.

Assembly of the LNAs occurred in two steps. First, the package wafer was aligned and bonded to the circuit wafer using silver epoxy.[13] Second, the flip-chip HEMT devices were placed on the circuit wafer through the access windows in the shielding wafer, and soldered to the amplifier circuit. Figures 13.21(a) and 13.21(b) show the fabricated LNA with the package removed and with

the conformal package. Finally, Fig. 13.21(c) shows the amplifier in packaged form, with only the dc/RF probe pads and the flip-chip HEMT visible.

13.5.3 Amplifier Performance

Figure 13.23 exhibits calibrated measurements of the S-parameter data with the simulated results of the original microstrip amplifier design from the Libra Series 4 software.[21] The TRL method is used for calibration, and the dc biasing conditions of the HEMT are achieved through the on-wafer bias network incorporated in the design with bias point conditions of $V_{DS} = 1.0$ V, $I_{DS} = 43$ mA, and $V_G = -0.7$ V. The good agreement observed between simulated and measured data for the return loss at port 1 (S_{11}), the trend in the insertion loss (S_{21}), and return loss at port 2 (S_{22}), show that fairly accurate modeling of the conformal package was accomplished by simply including the upper shielding cavity in the design of the CPW components. Performance of the shielded CPW balanced stubs, used for RF isolation in the gate bias network, was measured independently and is shown in Fig.13.24. The stub resonates at approximately 20 GHz, instead of the desired 18.5 GHz, but has a broad response with a resonance of –30 dB in $|S_{21}|$. The amplifier circuit was expected to achieve a noise figure of 0.9 dB at 20 GHz.

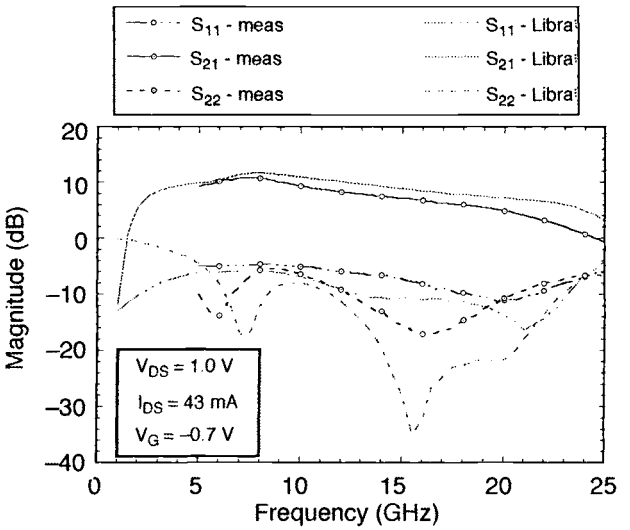


Fig. 13.23. Measured S-parameters of the LNA compared to simulated results for an unshielded microstrip amplifier on an alumina substrate.

13.6 Discrete Micromachined Packages

For many years, electronic packages for microwave components and systems have been a low priority compared with the development of HF active devices and circuit design techniques. The electronic package is employed to provide electronic shielding, physical protection, and thermal isolation. When the packaged performance of these components and systems is compared with unpackaged performance, the packaged design exhibits significant degradation to the electrical response of the circuits, especially those that operate above microwave frequencies (low gigahertz).

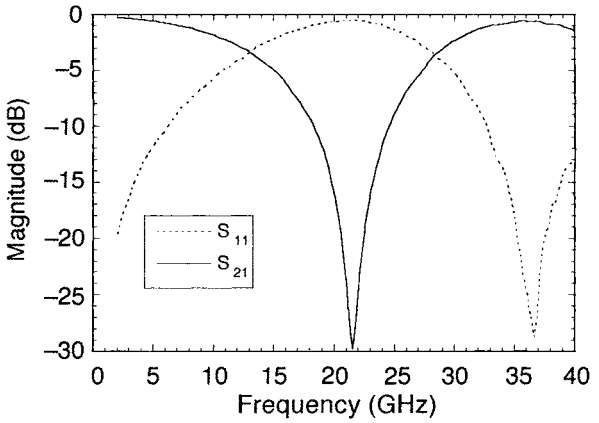


Fig. 13.24. Measured S-parameters of the shielded CPW-balanced stubs used in the gate bias network.

Recent trends, however, indicate a shift in priority toward the understanding, improvement, and development of packaging approaches that can offer low cost, high quality, and electrical compatibility to such designs. MCMs, for example, have been developed in cofired ceramic substrates with printed wiring boards that use thick and thin film technologies. For analysis, CAD tools are being developed to predict how the electromagnetic behavior of the package affects the performance of a given circuit.[27] Although some advanced quality packaging approaches are currently available, the high cost and large weight/volume associated with them limit their use in generic applications.[28] To address the aforementioned issues, a novel approach has been taken to replicate the thick-film electronic microwave package shown in Fig. 13.25(a) that was designed for a planar phase shifter. The new package, shown in Fig. 13.25(b), is based on Si as substrate and uses Si micromachining techniques to realize the appropriate vias and package design configurations.

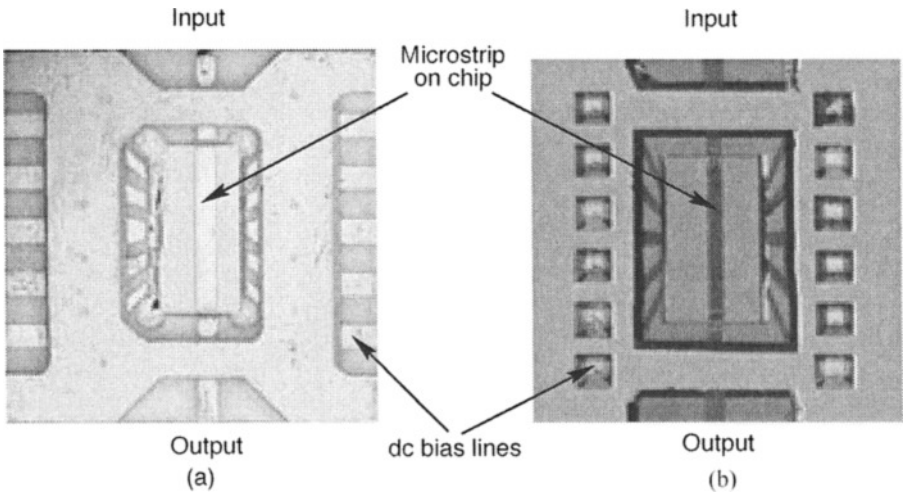


Fig. 13.25. (a) HTCC hermetic package designed by NASA Lewis Research Center and (b) micromachined silicon-based package.

13.6.1 Silicon as a Discrete HF Package

Si as a substrate offers several advantages to HF packaging design. First, the mechanical properties of Si, as described in 1982 by Petersen,[4] are excellent and compatible with those of a variety of metals. Second, enabling technologies, like Si micromachining, combined with standard IC processing techniques, offer flexible options to control HF parasitics associated with vias and the signal line conductor. With the fabrication precision available with these enabling technologies, via hole diameters and conductor line width dimensions can be produced reliably. Reducing parasitic inductance and capacitance requires precision fabrication and minimal variability. Third, as shown in Table 13.4, the variation in thermal conductivities between semiconductors (e.g., GaAs, InP, and Ge semiconductors) and ceramics (e.g., 92% alumina) can cause major problems with heat dissipation. In contrast, Si with a thermal conductivity of 135 W/(m•K) is much more similar to the other compound materials and behaves as an excellent heat sink.[29] Because of the combined electrical and mechanical properties of Si, it has potential use in low-cost batch-fabricated components and packages, and because these components and packages can also be developed in an on-wafer manner, higher circuit densities can be realized without significant increases in the spatial volume and weight of the package.

Table 13.4. Thermal Conductivities of Various Materials

Material	Thermal Conductivity [W/(m•K)]
GaAs	80
InP	65
Ge	68
92% alumina	18
Si	135

13.6.2 The Package Layout

The phase-shifter package in Fig. 13.25(a) is a hermetically sealed HTCC design that was processed on 92% pure alumina ($\varepsilon_r = 9.5$) by Hughes Aircraft for the National Aeronautics and Space Administration (NASA) Lewis Research Center.[30] A micromachined prototype (Fig. 13.26) of the same package was developed in Si and consists of five layers that are combined to provide a package for an IC chip. To characterize the package, a 50-Ω microstrip line is printed on a substrate chip. The layers of the prototype follow:

- Layer 1 is the metal base plate to the package.
- Layer 2 is the signal line substrate, which contains the RF input/output and dc bias lines that are printed on the high-resistivity Si. A hole has been etched through the center of the wafer to define a portion of the compartment that holds the IC chip. Six micromachined vias are etched along the left and right sides of the package housing to provide em shielding and to provide continuous electrical connection between the various layers.
- Layer 3 is the seal frame that contains the vias and IC chip opening, which are similar to the vias and opening on Layer 2, with additional probe access windows for on-wafer testing.
- Layer 4 is the upper metallization for the seal frame.
- Layer 5 is the top metal lid used to enclose the IC chip in the package.

The overall Si package dimensions are $7.112 \times 7.112 \times 1.27$ mm^3. This design varies from the alumina design by having smaller bias line widths, three instead of four ultrasonic wire bonds to

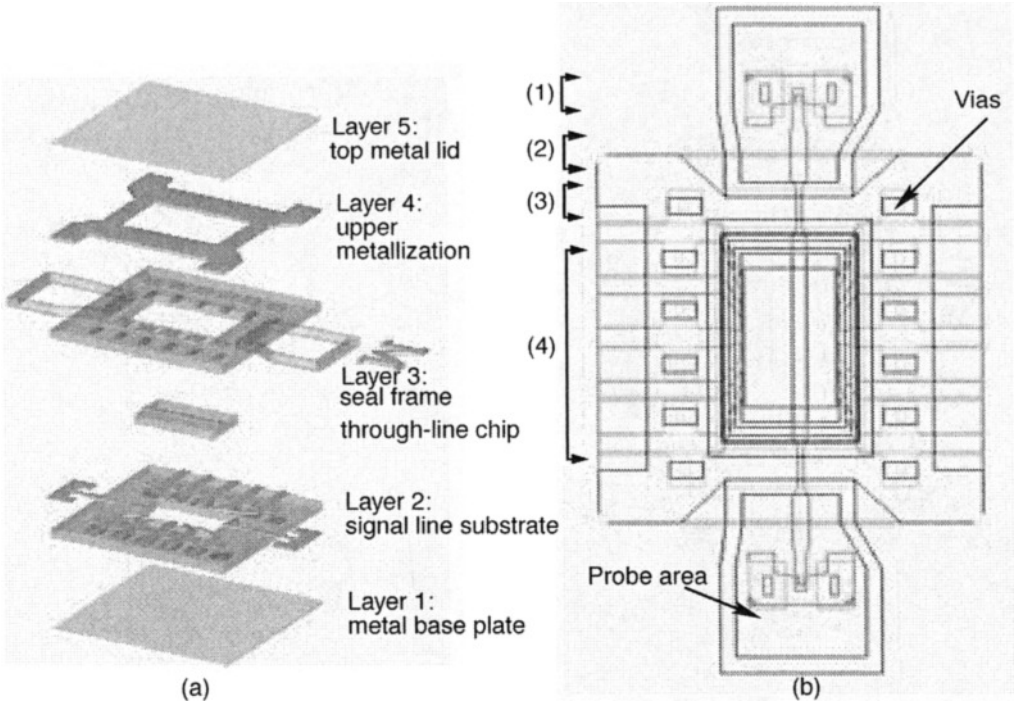


Fig. 13.26. (a) Three-dimensional view of package layers. (b) Actual CAD drawing with all layers superimposed with line transitions: (1) GCPW, (2) microstrip, (3) stripline, and (4) shielded microstrip.

connect the microstrip to the package, and symmetry between the square vias and the feed line to the IC chip, as opposed to a slight offset between the vias and the feed line in the alumina design.

13.6.3 Package Fabrication

Conventional thin-film processing technology and Si micromachining are used to fabricate the discrete package. The metallization consists of 3 μm of electroplated Au on the package and for the microstrip line [Fig. 13.26(a)]. To launch a signal into the package, a 50-Ω GCPW-to-microstrip transition is used at the input/output of the chip. Next, the IC chip has a 50-Ω through-line that changes from the GCPW to microstrip, stripline, and shielded microstrip sections, as shown in Fig. 13.26(b). Finally, conducting silver epoxy is used to attach the layers and provide the connection between the upper and lower metallization surfaces of Layers 2 and 3.

13.6.4 Electrical Response

The insertion loss and return loss for the IC chip in the alumina package and Si-based discrete package are shown in Fig. 13.27. The insertion loss of the Si package is consistently lower than that of the alumina package by 0.5 dB over the entire frequency range, while the return loss of the Si package remains similar to that of the alumina design.

Note that the irregularities in the return loss curve for the alumina package reflect the presence of parasitic effects more than impedance mismatch. If the line impedance (Z_0) is mismatched to the testing probe (Z_L), the return loss would be much higher than that observed based on the expression for reflection, ($|\Gamma|$), given in Eq. 13.2.

$$20\log(|\Gamma|) = 20\log\left(\frac{Z_L - Z_0}{Z_L + Z_0}\right), \text{ in dB} \tag{13.2}$$

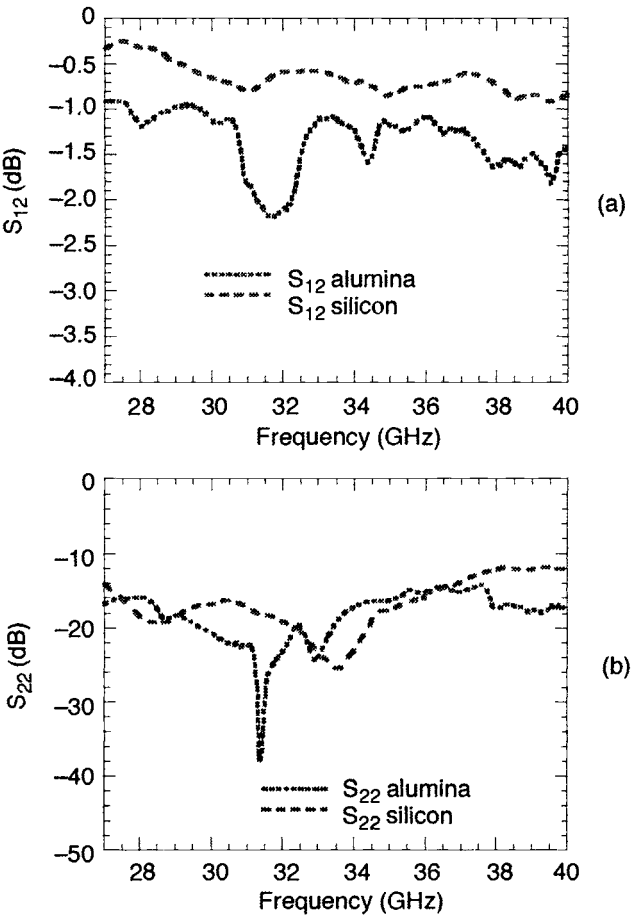


Fig. 13.27. (a) Insertion loss and (b) return loss.

In an evaluation of the ohmic loss, Fig. 13.28 shows that the signal attenuation is strongly related to ohmic loss, as shown by the reduction of the attenuation value when metal thickness increases from 3 to 8 μm.

13.7 Micromachined Filters for High-Density Integration

Many communication-related applications (e.g., wireless systems, collision avoidance radars) require efficient usage of the frequency spectrum. This usage requires the development of high-performance components such as filters and switches to optimize system behavior. Optimum planar filter performance is achieved by having a large ratio of high- and low-characteristic impedance values. These values are typically achieved in low-dielectric-constant (or low-index) materials such as laminants and ceramics. Even though filter designs on other materials can produce sharp cutoffs between pass and stop bands as well as high stop-band attenuation, the materials are incompatible with active device fabrication. On the other hand, high-dielectric-constant materials like semiconductors are compatible with active device fabrication, but they suffer from reduced high-/low-impedance value, which limits the cutoff slopes between frequency bands and produces a higher insertion loss in the stop band. Improving the range of impedances for high-index materials will result in substantial improvements to HF planar circuit components like filters.

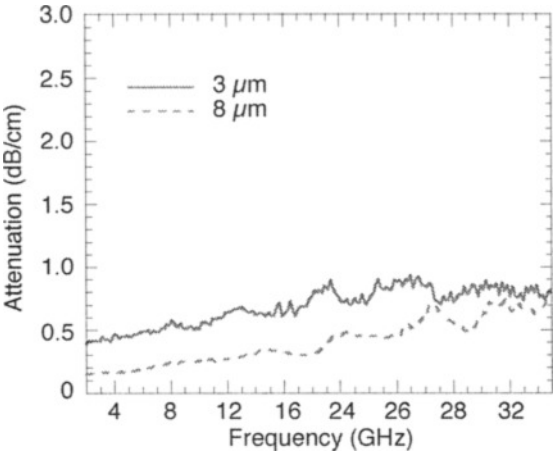


Fig. 13.28. Attenuation of 9-mm through-line in a package.

In many filter designs, performance is greatly affected by the ability to accurately realize prototype filter elements, such as a capacitor or an inductor, in the equivalent transmission line component. These capacitive and inductive values are highly dependent on the ability to design appropriate high- and low-characteristic impedance values in a specific type of transmission line. For each circuit value, a distributed line is chosen whose characteristic impedance and length closely approximate the lumped-circuit value. Impedance values that are very low make excellent RF capacitors; those that are very high make excellent inductors. The dielectric constant of the material affects the value of the characteristic impedance (Eq.13.3). Planar transmission lines on high-dielectric-constant (ε_r) substrates such as Si with $\varepsilon_r = 11.7$ exhibit a reduction in the high-characteristic impedance (Z_{high}) value and an increase in the low-characteristic impedance (Z_{low}) value when compared with a similar type of line on low-dielectric-constant materials. Equation 13.3 shows the calculation for the characteristic impedance, where μ_r and μ_o are the relative and free-space permeability, and ε_r and ε_o are the relative and free-space permittivity.

$$Zo = \sqrt{\frac{\mu_r \mu_o}{\varepsilon_r \varepsilon_o}} \tag{13.3}$$

With the use of micromachining to selectively etch the semiconductor material, the optimum substrate properties, based either on thickness or dielectric constant, can be synthesized to produce very high- or very low-characteristic impedance values (see Fig. 13.29). This selective etching allows individual elements to be preferentially designed, and when these are printed on high-index substrate environments, circuit design requirements are optimized, such as those found in filters.[31]

13.7.1 Low-Impedance Design

Microstrip lines on electrically thin substrates have lower impedance values than those on electrically thick substrates. Si micromachining is used to selectively reduce the thickness of a substrate. This processing affects the capacitance because the value depends on the substrate thickness (d), dielectric constant (ε), and plate area (A), as described in Eq. (13.4).

$$C = \frac{\varepsilon A}{d} \tag{13.4}$$

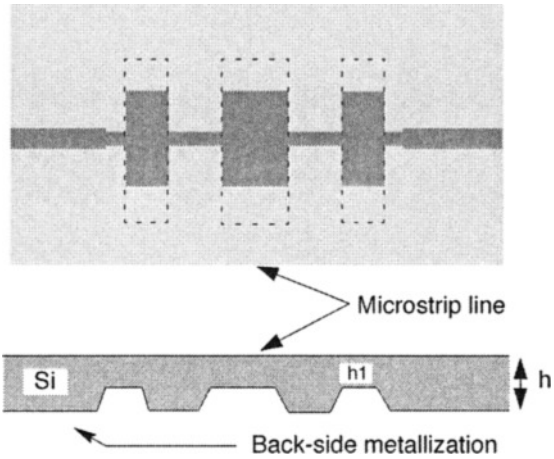


Fig. 13.29. Circuit layout for micromachined filter with a synthesized low-impedance section (i.e., capacitive regions). The substrate height is labeled h, and the reduced thickness air region is labeled h1.

There are important implications for components such as the planar microstrip filter. In these components, to synthesize the desired filter response, the required filter response is determined by using lumped-circuit elements that are converted to appropriate impedance values and electrical lengths, in the transmission line. Moreover, in high-index designs, the upper and lower impedance values are limited, which makes an optimized response function difficult to realize.

For a step-impedance filter design, commercial CAD tools, like HP's LineCalc and Microwave Design System software, can be employed to design and evaluate the microstrip response of a filter on full-thickness substrates. To implement such a design process on a synthesized substrate, the optimized design parameters are converted to the appropriate transmission line characteristic impedance. Then using HP's LineCalc, the impedance dimensions and line lengths are determined for a given substrate parameter (e.g., thickness, dielectric constant). The design can then be simulated in HP's MDS software. The advantages of this design are more compactness because of the reduction in the capacitive region width (on the thinner substrate regions) compared with similar capacitance sections on thicker substrate material. The reduction in substrate thickness makes it possible to reduce the conductor area and allows for the realization of similar or lower impedance values (Table 13.5).

Table 13.5. Dimensions for the Step Impedance Filter Design Shown in Fig. 13.29

Section (Ω)	Length	Width (h =100)	Width (h = 50)
1 (100)	135	10	10
2 (20)	270	380	190
3 (100)	684	10	10
4 (20)	480	380	190
5 (100)	684	10	10
6 (20)	270	380	190
7 (100)	135	10	10

Using the popular Butterworth filter with seven sections, a conventional design and micromachined design are implemented, based on characteristic impedance values of 20 and 100 Ω, respectively. The synthesized microstrip substrate with reduced-thickness regions can offer lower capacitive values (i.e., less than the typical 20 Ω), thereby resulting in an improved design because of the increase in the high and low realizable impedance values. The filter design and dimensions given in Fig. 13.29 and Table 13.5 have inductive regions that are identical to those of the conventional design. These regions are printed on full-thickness material—100 μm, in this case—of the host wafer, while the capacitive sections are printed on 50-μm-thick regions to produce a capacitance equivalent to the reference filter design.

13.7.2 Fabrication Considerations

The fabrication procedures are as follows.

- The circuits are printed on high-resistivity Si with a thickness of 100 μm.
- The low-impedance sections are developed on 50-μm-thick regions that have been etched using KOH anisotropic etchant.
- The etched cavity regions produce a sloped sidewall angle of 54.7 deg that results in a gradient of 35.4 μm at each low-impedance edge.

Each design has a 50-Ω microstrip line and is fed by a CPW probe pad that converts the on-wafer probe excitation to a microstrip mode. The circuits have 3.4 μm of electroplated Au and have a 2.5-μm ground plane metallization of Cr/Al/Cr/Au that has been evaporated to cover the entire wafer surface. Finally, the circuit wafer is attached to a secondary wafer for additional support with similar metal composition.

13.7.3 Characterization

For the seven-section Butterworth design, the low-impedance sections have line widths that decrease from 380 μm on the 100-μm-thick substrate to 190 μm on the 50-μm-thick region. The response, as shown in Fig. 13.30(a), produces near similar results between the synthesized design and the conventional design.

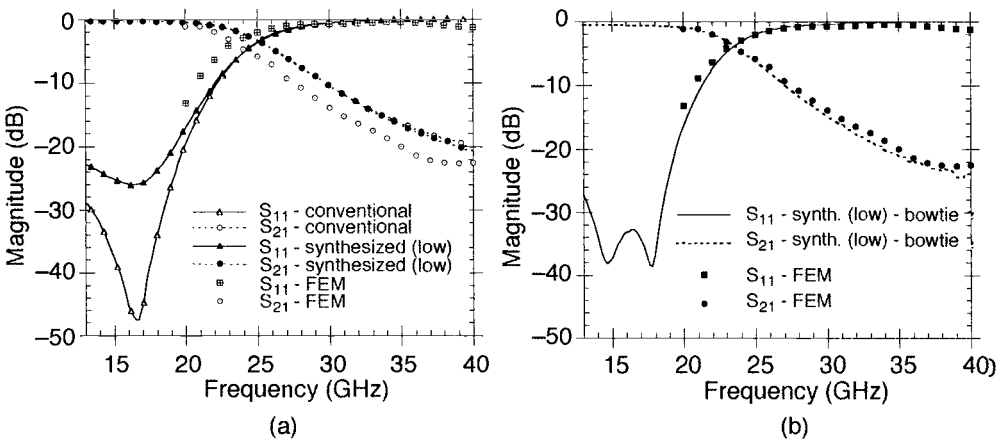


Fig. 13.30. Synthesized low-impedance filter design. (a) Filter response with measurement and modeled results and (b) comparison of FEM to bowtie response. The FEM follows the lines on either S11 or S11 synthesized (low) for (a) and S11 synthesized (low)-bowtie for (b). The FEM calculations were not made below 20 GHz.

A FEM simulation[32] of the design on the synthesized substrate with vertical side walls indicates that the filter 3-dB point would be shifted 1 GHz lower than the point on the rectangular conductor design. In this case, bowtie tapers were introduced to offer a transition along the etched angle profile in the low-impedance section. The measured and modeled data in Fig. 13.30(b) produced nearly identical results, which indicates that the etch angle does affect the phase delay of the various sections. The response curves of this design are very similar to those of the regular design, with better attenuation in the stop band as well as a better cutoff frequency.

In Fig. 13.31, the radiation loss of the conventional and bowtie design is compared with the synthesized design with rectangular conductors. Losses are slightly higher in the bowtie design, but these losses occur because the sharp corners of the bowtie cause current crowding. Overall, this approach has shown merit, and the synthesized and regular designs yield similar losses.

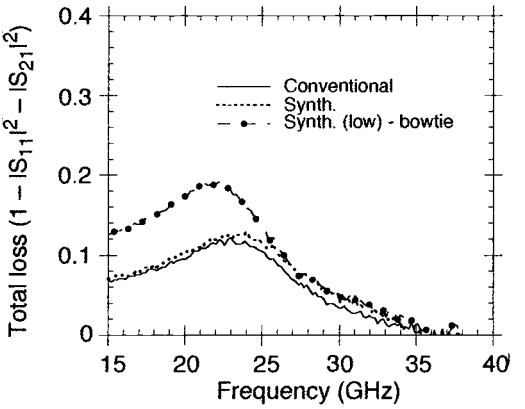


Fig. 13.31. Total loss calculations between two designs.

13.8 Conclusions

This chapter has presented demonstrations showing how Si-micromachined on-wafer packages can be integrated into a variety of communication systems interconnects and components to produce enhanced performance. A significant benefit of using this type of technology is the extreme light weight and compactness offered by this type of micropackage, as well as improved performance for a number of designs above 20 GHz. In addition, on-wafer packaging developed with Si micromachining can be easily used with other HF design and fabrication processes, such as MMIC, commonly used at millimeter waves.

Micropackaged interconnects provide extremely low electromagnetic coupling and parasitic radiation in microstrip-based interconnects. The package can be implemented with standard IC processing techniques and is amenable to integration with more complex high-density designs. Furthermore, these packaged interconnects reduce circuit interactions between near-neighbor interconnects, offering coupling levels comparable to the overall background noise found in the packaged system, and also provide performance improvements to those circuits that exhibit high radiation.

Conformal micropackages in the LNA designs at 20 GHz offer RF shielding and isolation without degrading the performance of the amplifier. In addition, the conformal packaging approach is compatible with flip-chip techniques, which further reduce fabrication costs and improve yield. A discrete package has also been shown, which verifies that Si can offer the

advantages of heat sinking found in alumina HTCC packages with the advantage of improved operation at higher frequencies and lower development cost.

The micromachined filters on synthesized substrates have demonstrated more compact designs for improved low- and high-impedance sections. The filter design is much more compact compared with designs on full-thickness high-index substrates.The potential for much higher ratios of high- and low-impedance sections is particularly important for filter applications. This potential can be improved by a factor of 1.5 to 2 by reducing the low-impedance section by half or increasing the high impedance sections by 1.5. Finally, Si-micromachining synthesized substrates are easily realizable in high-index semiconductor materials and have been used with simple filter designs, like the step-impedance filter, using commercial CAD tools based on quasistatic and full-wave models.

In the future, it is not unreasonable to expect that advanced packaging technology, such as HF micropackaging, will provide avenues for developing unprecedented lightweight communication systems that offer high performance. For this advanced technology to be developed, the investigation and application of enabling technologies—such as Si micromachining—with HF circuit design techniques are necessary to demonstrate proof of concept. Examples of Si micropackaging have been illustrated on a number of communication building blocks, such as interconnect lines, amplifiers, and filters. When these building blocks are integrated into a single system, future advanced communications systems can appear, as shown in Fig. 13.32.

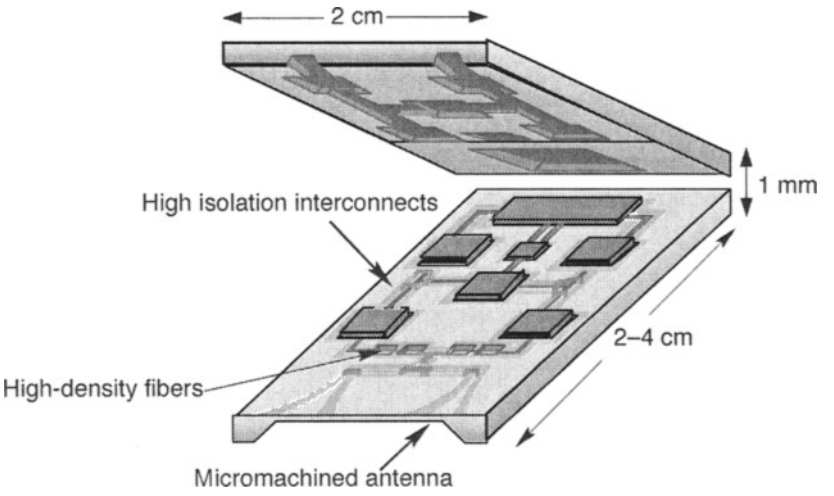


Fig. 13.32. Advanced high-frequency micropackaged communication system.

13.9 References

1. Jet Propulsion Laboratory, "System on a Chip," *CISM Workshop, 1998* (Pasadena, CA, 1998).
2. R. S. Elliott, *An Introduction to Guided Waves and Microwave Circuits* (Prentice Hall, NY, 1993).
3. D. Doane and P. D. Frazon, *Multichip Module Technologies and Alternatives—The Basics* (Van Nostrand Reinhold, NY, 1993), p. 74
4. K. E. Petersen, "Silicon as a Mechanical Material," *Proceedings of the IEEE* (May 1982), Vol. 70, No. 5, pp. 420–457.
5. R. F. Drayton and L. P. B. Katehi, "Development of Self-Packaged High Frequency Circuits Using Micromachining Techniques," *IEEE Trans. Microwave Theory Tech.* **43** (9), 2073–2080 (September 1995).

6. S. V. Robertson, L. P. B. Katehi, and G. M. Rebeiz, "A 10-50 GHz Micromachined Directional Coupler," *1996 IEEE MTT-S International Microwave Symposium Digest* **2**, 797–800.
7. S. V. Robertson, L. P. B. Katehi, and G. M. Rebeiz, "Micromachined Self-Packaged W-Band Bandpass Filters," *1995 IEEE MTT-S International Microwave Symposium Digest* **3**, 1543–1546.
8. R. F. Drayton and L. P. B. Katehi, "Micromachined Conformal Packages for Microwave and Millimeter-wave Applications," 1995 *IEEE International Microwave Symposium Digest* **2**, 1387–1390.
9. R. A. Laudise, *The Growth of Single Crystals* (Prentice Hall, Englewood Cliffs, NJ, 1970).
10. S. Wolf and R.N. Tauber, *Silicon Processing for VLSI Era, Volume 1: Process Technology* (Lattice Press, Sunset Beach, CA, 1986).
11. P. Rai-Choudhury, *Volume 2: Micromachining and Fabrication* (SPIE Publishing, 1997).
12. Hewlett-Packard Company, Santa Clara, CA.
13. Epoxy Technology, Inc., Billerica, MA.
14. GGB Industries, Naples, FL.
15. R. B. Marks and D. F. Williams, Program MultiCal, rev. 1.00, NIST, August, 1995.
16. R. B. Marks, "A Multiline Method of Network Analyzer Calibration," *IEEE Trans. Microwave Theory Tech.* **39**, 1205–1215 (July 1991).
17. D. M. Pozar, *Microwave Engineering*, 2nd Ed.,(John Wiley & Sons, NY, 1998), pp. 196–206.
18. K. E. Bean, "Anisotropic Etching of Silicon," *IEEE Trans. Electron Devices* **ED-25** (10), 185–1193 (October 1978).
19. Abu-Zeid-Corner and M. M. Abu-Zeid, "Corner Undercutting in Anisotropically Etched Isolation Contours," J. Electrochem. Society **131** (9), 2138–2142 (September 1984).
20. Xian-Ping Wu and W. H. Ko, "Compensating Corner Undercutting in Anisotropic Etching of (100) Silicon," *Sensors and Actuators* **18**, 207–215 (1989).
21. Hewlett Packard, EESOF Libra 4 Software.
22. N. Dib and L. Katehi, "Modeling of Shielded CPW Discontinuities Using the Space Domain Integral Equation Method (SDIE)," *Journal of Electromagn. Wave Appl.* **5** (4/5), 503–523 (1991).
23. R. F. Drayton, R. M. Henderson, and L. P. B. Katehi, "High Frequency Circuit Components on Micromachined Variable Thickness Substrates," *IEEE Electronics Letters* **33** (4), 303–304 (February 1997).
24. M. Matloubian, "Flip-Chip Components for Realization of Low-Cost Millimeter Wave Systems," to be presented at the *1996 Asia-Pacific Microwave Conference* (New Delhi, India, December 1996).
25. R. Isobe, C. Wong, A. Potter, L. Tran, M. Delaney, R. Rhodes, D. Jang, L. Nguyen, and M. Le, "Q- and V-Band MMIC Chip Set Using 0.1 mm Millimeter-Wave Low Noise InP HEMTs," in *IEEE MTT-S International Microwave Symposium Digest* (Piscataway, NJ, May 1995), Vol. 3, pp. 133–1136.
26. N. I. Dib and L. P. B. Katehi, "Impedance Calculation for the Microshield Line," *IEEE Microwave and Guided Wave Letters* **2** (10), 406–408 (October 1992).
27. J. G. Yook, L.P.B. Katehi, R. N. Simons, and K. A. Shalkhauser, "Experimental and Theoretical Study of Parasitic Leakage/Resonance in a K/Ka-Band MMIC Package," *IEEE Trans. Microwave Theory Tech.* **44** (12), 2403–2410 (December 1996).
28. M. I. Herman, K.A. Lee, E. A. Kolawa, L. E. Lowry, and A. N. Tulintseff, "Novel Techniques for Millimeter-Wave Packages," *IEEE Trans. Microwave Theory Tech.* **43** (7), 1516–1523 (July 1995).
29. J. E. Licari, *Multichip Module Design, Fabrication and Testing* (Mc-Graw Hill, Inc., NY, 1995).
30. R. M. Henderson and L. P. B. Katehi, "Silicon Based Micromachined Packages for Discrete Components," *1997 IEEE MTT-S International Microwave Symposium Digest* **2**, 521–524.
31. R. F. Drayton, S. Pacheco, J. Yook, and L. P. B. Katehi, "Micromachined Filters on Synthesized Substrates, *1998 IEEE MTT-S International Microwave Symposium Digest* **3**, 1615–1618.
32. J.-G. Yook, N. Dib, and L. Katehi, "Characterization of High Frequency Interconnects Using Finite Difference Time Domain and Finite Element Methods," *IEEE Trans. Microwave Theory Tech.* **42**, 1727–1736 (September 1994).

14

MEMS-Based Active Drag Reduction in Turbulent Boundary Layers

T. Tsao,* F. Jiang,* C. Liu,* R. Miller,* S. Tung,† J.-B. Huang,† B. Gupta,* D. Babcock,* C. Lee,† Y.-C. Tai,* C.-M. Ho,† J. Kim,† and R. Goodman*

14.1 Introduction

Drag reduction is a problem of great interest. From a practical point of view, any object moving in a fluid experiences drag. When the object is a hand-built vehicle, and the drag experienced by this vehicle results in increased fuel costs and decreased operating efficiencies, understanding and reducing drag become extremely worthwhile. Indeed, for as long as people have been moving in vehicles, they have been interested in moving faster with less effort.

As an example of how important the reduction of drag would be for the airline industry, one can look at work done by Walsh.[1] In 1985, Walsh estimated that if the viscous drag experienced on the fuselage of airplanes could be reduced by 10%, the cost savings would be on the order of $350 million. By extrapolating to 1998 using a 3% annual rate of inflation and a more than double rate of air travel,[2] those numbers would translate into a savings of over $1.1 billion.

Most of this chapter is dedicated to explaining an ongoing effort to reduce drag in turbulent boundary layers through the use of distributed microelectromechanical systems (MEMS) to enact active control. The result of the successful completion of this project will be a MEMS system that will combine microsensors, microactuators, and microelectronics (or M^3) all integrated on a single wafer. Such a system will be used in an active manner to reduce drag in a turbulent boundary layer. In addition, the technologies developed for this particular M^3 system will be applicable to other MEMS-based systems, not just to drag reduction. Before details are presented, some background is given regarding both the nature of drag and some examples of drag reduction in nature.

There are many sources of drag on any object moving through a fluid. In this chapter, we will only consider two types: pressure or form drag, due to flow separation, and viscous or skin-friction drag.

Pressure or form drag is a result of a negative pressure differential between the front and the back of the object. A fluid moving past an object will separate from the object and cause a wake to be produced behind the object. The downstream velocity in this wake is less than the upstream velocity. By applying the laws of conservation of momentum, the velocity lost by the flow is transferred to a retarding force (which can also be thought of as a pressure differential) on the body. Race car drivers often take advantage of these pressure differentials by "drafting" or closely following the driver ahead of them, who is "pushing" most of the air out of the way. To reduce this type of drag, it is necessary to delay the onset of separation for as long as possible.

Viscous or skin-friction drag exists within a very thin region adjacent to the object. The existence of this "skin," called the boundary layer, is a direct result of a boundary condition applied to the flow field, known as the no-slip boundary condition. Essentially, this condition requires that the tangential velocity of any flow next to a solid boundary be equal to the velocity of the boundary itself. For a nonmoving boundary, this implies a tangential velocity of zero. Within the

*California Institute of Technology, Pasadena, California.
†University of California, Los Angeles.

boundary layer, then, the velocity increases in a very short distance from zero to the free-stream velocity. This fast change in velocity, u, results in a shearing force at the surface that is directly proportional to the gradient of the velocity, du/dy, with respect to the distance away from the surface. The shear stress and the drag, which is the shear stress integrated over an area, can be expressed as follows:

$$\tau = \mu \frac{du}{dy}$$

$$F_d = \int_A \tau dA$$

where τ is the shear stress, μ is the fluid viscosity, and F_d is the drag force.

In an attempt to reduce drag on airplane wings, one question needs to be answered: is it more effective to reduce form drag or viscous drag? As previously mentioned, form drag is reduced by delaying separation. Airplane wings are already designed with this factor in mind. Therefore, it is unlikely that a large drag reduction could be achieved through an attempt to reduce form drag. Friction drag, however, is very high in a turbulent flow across a wing's surface. Therefore, it makes sense to attempt to reduce friction drag. In a turbulent flow across a surface, many random structures[3] form. These structures are within the boundary layer, directly next to the surface, and are composed of counter-rotating vortex pairs that travel downstream (Fig. 14.1). In between these vortex pairs, faster moving (or higher momentum) fluid is brought down closer to the surface. This downwash results in a higher velocity gradient, which results in higher shear stress and, ultimately, higher surface drag. By preventing the occurrence of such structures, or by reducing the effect of the structures, it is conceivable that drag reduction is feasible in turbulent boundary layers. Typical flow conditions often yield structures that have sizes on the order of millimeters and lifetimes on the order of milliseconds. Any drag-reduction system designed to interact with these vortex pairs must have sensors and actuators of the same size and with similar operating

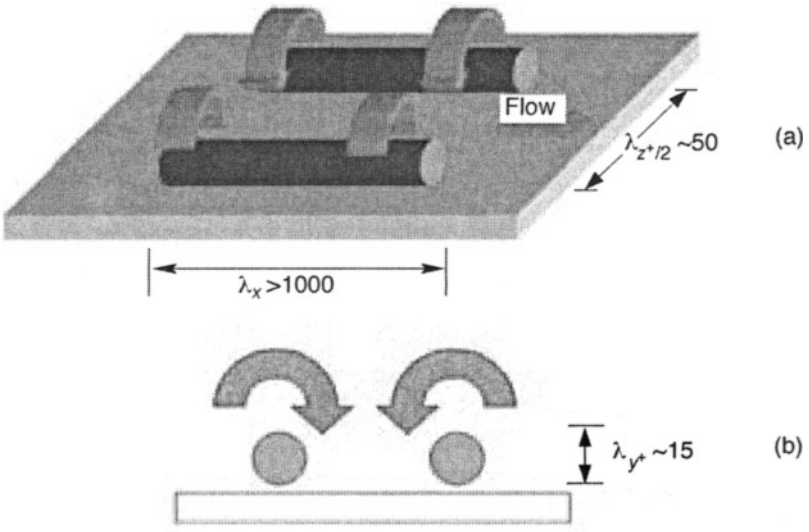


Fig. 14.1. Simplified (a) perspective and (b) side views of a counter-rotating vortex pair. In fluid mechanics, normalized length scales are often represented by λ and depend on characteristics such as Reynolds number. In the UCLA wind tunnel, for example, typical lengths range from 8 to 30 mm.

frequencies. Therefore, MEMS, which can be fabricated in the submillimeter dimension, seems a logical solution for such a system.

Passive boundary layer control has been the subject of numerous scientific works. Although passive control can be effective, the primary reason for studying passive, rather than active, control is one of simplicity: it is far easier to construct passive elements and test them in a flow than it is to devise an active control system.

One of the primary methods for passive control involves the use of small riblets on a surface.[4–9] At the expense of increasing the effective surface area (or "wetted" area), the riblets are used primarily to prevent the streamwise vortices from bringing high momentum fluid down to the surface. These riblets are generally placed in a direction parallel to the flow and can have various cross sections, with the most popular one being a V-groove. In the case of the V-groove, it is thought that the high momentum fluid being brought down cannot penetrate into the grooves, provided the groove size is on the order of the size of the vortices. If the integration of this lower shear stress over the larger wetted area is less than the integration of the standard shear stress over a flat plate, drag reduction is the result. In fact, drag reductions of 8% [1] have been achieved in a laboratory environment.

While passive drag reduction appears to be promising, it also has some drawbacks. Because the vortices occur randomly in space and time, any passive system with no sensors or feedback must necessarily be turned "on" at all times. As an example, riblets prevent high momentum fluid from being transferred down to the surface when counter-rotating vortices pass over them. However, riblets also result in a larger effective area over which drag can occur. The presence of riblets, then, actually increases drag when not reducing it. Also, should flow parameters, and hence the structure size, change, fixed-size riblets may not be effective. Such concerns lead one to consider an active control system. Two questions exist, then, about such a system: (1) how effective would such a system be and (2) what would be the requirements of the system?

The question of effectiveness is partially answered by Choi,[10] who has shown, through numerical simulation, the effectiveness of various methods of active control on the reduction of drag in turbulent boundary layers. Some of the techniques described in that work show drag reductions as large as 25%. Perhaps even more astonishing is the result from controlling only 5% of the surface, drag reductions of 15% can still be obtained. This incredible result is a direct consequence of using an active control scheme to determine the locations where actuation would be most effective and the subsequent control of the most significant structures.

One important point must be noted about Choi's work. Because the work was numerical simulation, the assumptions were that "sensors" could be placed anywhere in the flow, and actuation methods included the ability to exactly control all three velocity components (i.e., by blowing or suction at distinct locations). Most of these assumptions cannot be implemented experimentally. However, one practical set of simulations was done by placing sensors only at the wall. Regardless of the methods of simulation, the important point is that the structures contributing to high shear stress can be controlled in an active manner with positive results. A conceptual drawing of an M^3 unit cell that would need to be used to experimentally validate Choi's results is shown in Figure 14.2.

The idea of localizing the control becomes important because of the logistical nightmare that would result from folding all the data inputs into a single, centralized processing unit. In addition, localized control works because it probably isn't necessary for any decision-making unit to know what is happening globally in order to make localized decisions. An example can be seen in the human nervous system. When a person accidentally touches a hot object with his or her hand, it is not the brain that makes the decision to pull away. Local areas of the central nervous system

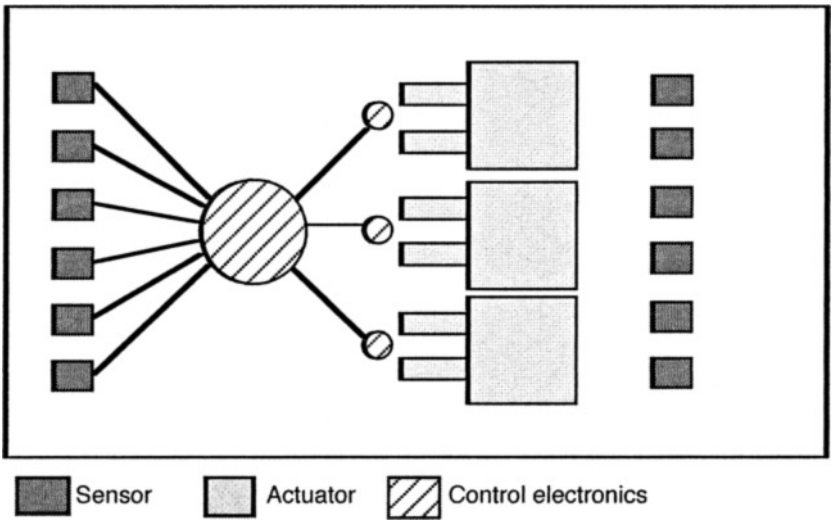


Fig. 14.2. Drawing that shows concept of a single sensor/actuator/electronic flow interaction unit.

can "make the decision," known as a reflex, before the brain is even aware of the situation. This is an example where no other information about other parts of the body is needed before the pull-back decision is made. The time saved in not sending the information to the brain minimizes damage from occurring.

The human nervous system is not the only example of an interesting nature-inspired solution to drag reduction. Much effort has been spent studying drag reduction in animals.[11] In particular, many examples of viscous drag reduction have been found in marine animals. Some of these examples involve the coverage of animal surfaces with various "slimes."[12] Many fast-swimming sharks[13] utilize a passive drag reduction scheme with small riblets. Passive in this sense means that the riblets are no more than simple attachments to the scales of the shark. No movement or control scheme is involved (see Figure 14.3).

Dolphins and porpoises have also been the focus of many drag reduction studies. In 1936, Gray[14] performed some simple observations and calculations that made him believe that the drag associated with the flow around a dolphin must be laminar, if not "better." He believed that the

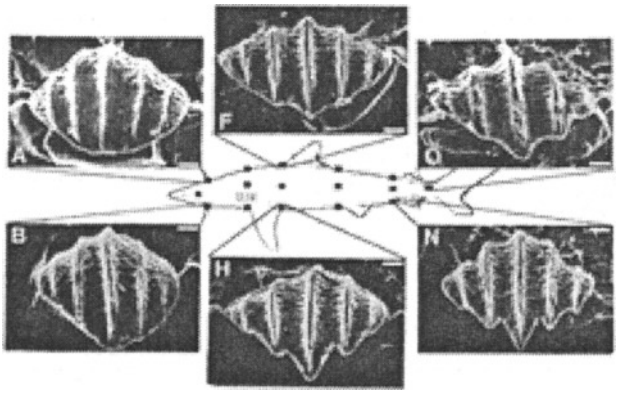


Fig. 14.3. Scanning electron microscopy (SEM) image of riblets on a fast-swimming shark. Typical riblet lengths are on the order of 50 μm.

flow could not be turbulent because the dolphins could not generate the power required to move through the water at the speeds he measured. Gray therefore postulated that dolphins have some means of converting turbulent flow to laminar flow for drag reduction. In later works, other researchers have speculated that this conversion may be the result of compliant skins or perhaps even active control. Others[15–17] postulated that many of Gray's assumptions and measurement techniques were in error and that dolphins do not have any such means of reducing drag. To this day, no overwhelming evidence has been obtained that proves any group of researchers correct. It is difficult to obtain accurate experimental evidence, as the very act of observing can often affect the dolphins (e.g., the very presence of a boat may spook the dolphins into going underwater, or the presence of a moving boat may generate wakes that affect the flow around the dolphins).

As a final note, there is a possibility that MEMS-based drag reduction could be used on aircraft for turbulent shear stress reduction. Whether or not this happens, however, the fundamental knowledge obtained from building an M^3 system for controlling drag in the turbulent boundary layer has broad applications for numerous engineering and scientific challenges. The effort to reduce drag in turbulent boundary layers requires the collaboration of engineers in many different fields. These fields include MEMS, which developed the shear stress imager and magnetic actuators; experimental fluid mechanics, which was responsible for making new measurements and determining actuator/flow interactions; control, which tested novel neural net algorithms; and VLSI electronics, which contributed several novel circuits.

14.2 Shear Stress Imager

A MEMS hot-film shear stress imager was developed for use in the drag-reduction system. The capability of this device for imaging surface shear stress distributions was demonstrated. The imager consists of multiple rows of vacuum-insulated shear stress sensors with a 300-μm pitch. Each sensor consists of a polysilicon resistor sitting atop a vacuum-insulated cavity. Current is passed through the resistor, and as fluid flow cools the hot film, an anemometry circuit attempts to maintain a constant temperature. The heat loss to the flow, as measured by the circuitry, is related to the shear stress. The small spacing between the sensors allows the imager to detect surface flow patterns that could not be directly measured before. The high frequency response (30 kHz) of the imager under constant temperature bias mode also allows it to be used in high Reynolds number turbulent flow studies. Measurements obtained by the imager in a fully developed turbulent flow agree well with the numerical and experimental results previously published.[18–20]

Because there is not one set size for the vortex pairs in a turbulent flow, their dimensions are best described statistically. One thing that can be said, however, is that the statistical size of a drag-inducing vortex-pair streak decreases as the Reynolds number of the flow increases. For a typical airflow of 15 m/s in our wind tunnel tests and a Reynolds number of about 10^4, the vortex streaks have a mean width of about 1 mm. The length of a typical vortex streak can be about 2 cm, giving the streaks a 20:1 aspect ratio. The frequency of appearance of the streaks is approximately100 Hz. The lifetime of the streaks is about 1 ms.[3]

Many methods can be used to measure the shear stress associated with these streaks.[21] One technique is the thermal method, which uses hot-film sensors to determine shear stress indirectly. This method has many advantages over other techniques, such as using a floating element, for real-time flow measurement and control. For example, it can achieve high sensitivity and high frequency response while keeping the sensor size small. Traditional hot-film sensors are electrically heated thin-metal film resistors on substrates. Since only heat convection responds to the shear stress change, it is desirable to minimize the conductive heat loss by thermally isolating the thin-film resistor from the substrate. By ensuring a maximum heat transfer through convection,

sensitivity is increased. In the past, the only way to reduce conductive heat loss was to use low thermal conductivity materials such as quartz for the substrate. Reasonably good sensitivity could be obtained only when such sensors were used to measure high thermal conductivity fluid such as water. However, they were not sensitive enough for the measurement in low thermal conductivity fluids such as air. Moreover, the size of traditional hot-film sensors is typically in the millimeter range.[21] This may be tolerable in measuring the mean shear stress value, but is definitely not acceptable in shear stress imaging with reasonable spatial resolution.

Thanks to the development of surface micromachining technology, we can optimize both the materials and the structure of the sensors for detecting turbulent structures. Figure 14.4 shows the

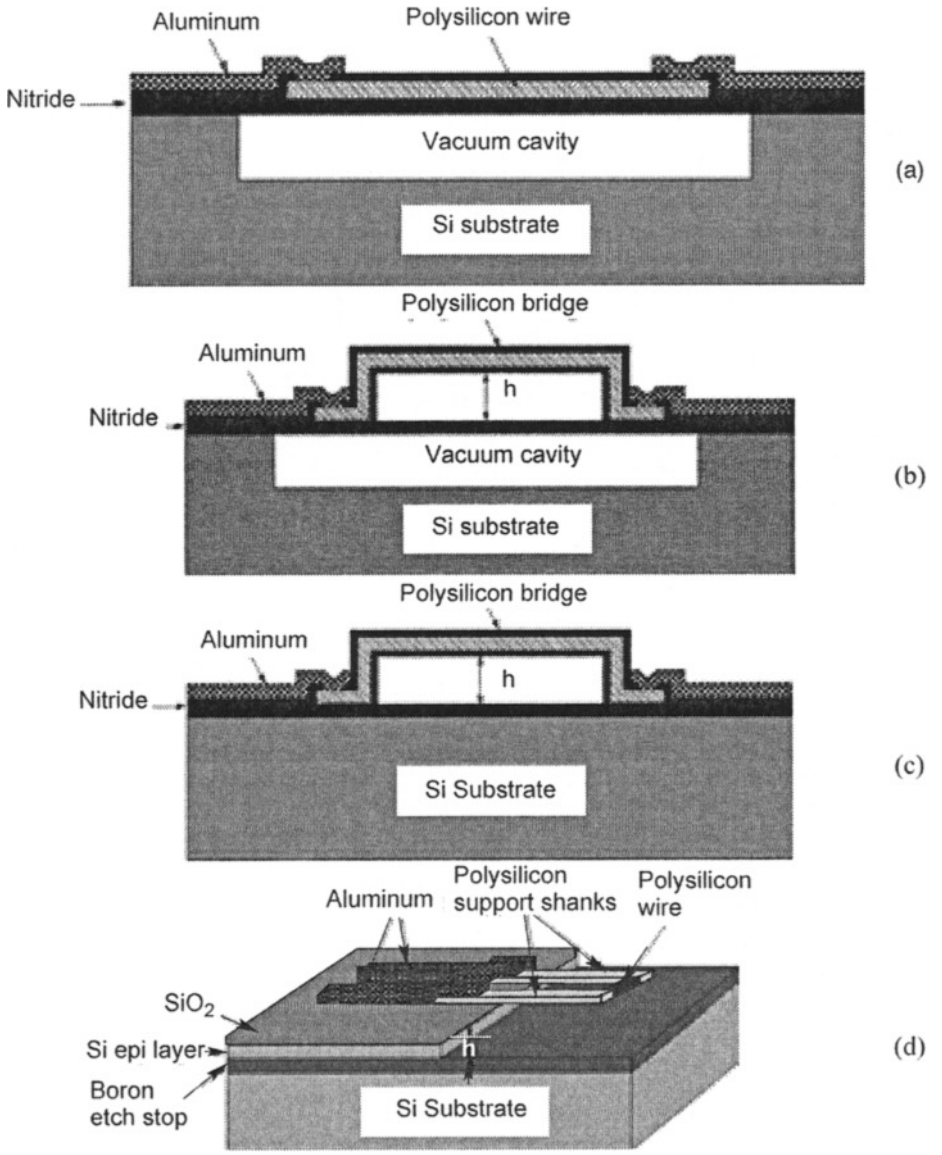


Fig. 14.4. Structures of the sensors. (a) Type I, (b) type II, (c) type III, (d) type IV.

cross-sectional structure of a few micromachined shear stress sensor types. Type I features a 2-μm-deep vacuum cavity with a 0.25-μm-thick polysilicon wire embedded in the nitride diaphragm. The vacuum cavity is designed to thermally isolate the diaphragm from the substrate.[22] Type II has a similar structure to that of type I, except that the polysilicon wire is lifted 4 μm above the diaphragm, thus achieving better thermal isolation. Type III is a conventional polysilicon bridge sitting on the solid substrate.[23] Type IV is basically a micromachined hot wire close to a wall, as has been previously reported. The wire is a few microns above the substrate surface and is in the linear velocity distribution region so that it measures the wall shear stress instead of velocity.[24] All four types of sensors were fabricated on a single chip to ensure identical thermal and electrical properties of the sensor materials. Figure 14.5 shows the calibration results of the four types in wind tunnel tests. The output changes are proportional to one-third the power of shear stress, which agrees with the heat transfer theory.[21] It is obvious that types I and II are the most sensitive sensors. Moreover, type I has a much simpler fabrication process than type II. Therefore, type I was chosen as the building block of this generation of shear stress imaging chips. These imaging chips will be described later.

After the structure of a sensor is decided, the geometry of each layer of material within the sensor is optimized to give maximum sensitivity. Sensitivity is higher for thinner diaphragms and for larger aspect ratio polysilicon resistors (i. e., larger L/W, where L is the length and W the width). However, the diaphragm cannot be too thin, or it would break during fabrication or operation. L is limited by the size of the whole device, which is at most 300 μm for the application in shear stress imaging. Therefore, L is chosen to be 150 μm. Because of directional sensitivity, the resistor should be straight. Therefore, the aspect ratio cannot be increased by using a serpentine structure. Also, W is limited by the photolithography and etching technology. Therefore, given our fabrication constraints, it is designed to be 3 μm to ensure good uniformity across many devices.

Figure 14.6(a) shows the photomicrograph of an individual type-I shear stress sensor, while Fig. 14.6(b) shows a picture of the 2.85 × 1.0 cm imaging chip. The chip is specifically designed for studies in turbulent flow with Reynolds numbers near 10^4. There are two identical sensor rows 5 mm apart, parallel to the broad side of the chip. Additional test rows are also on the chip and are

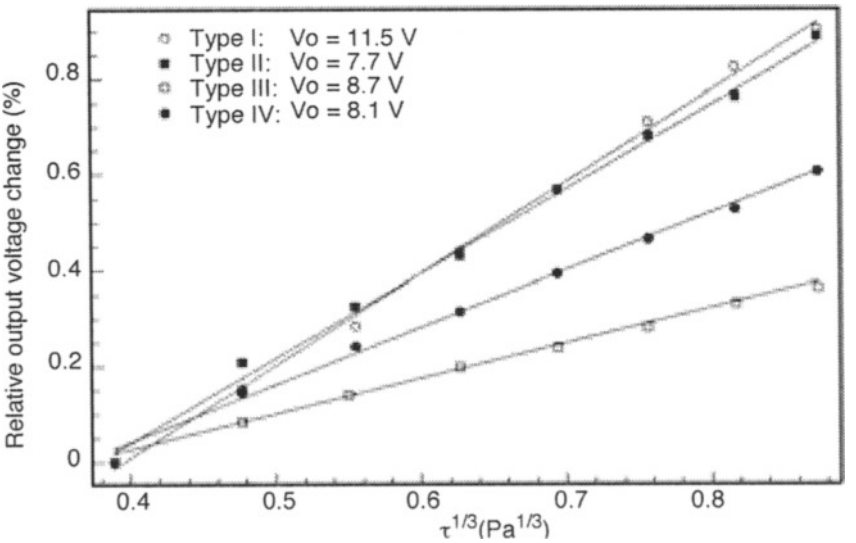


Fig. 14.5. Wind tunnel calibration of sensors.

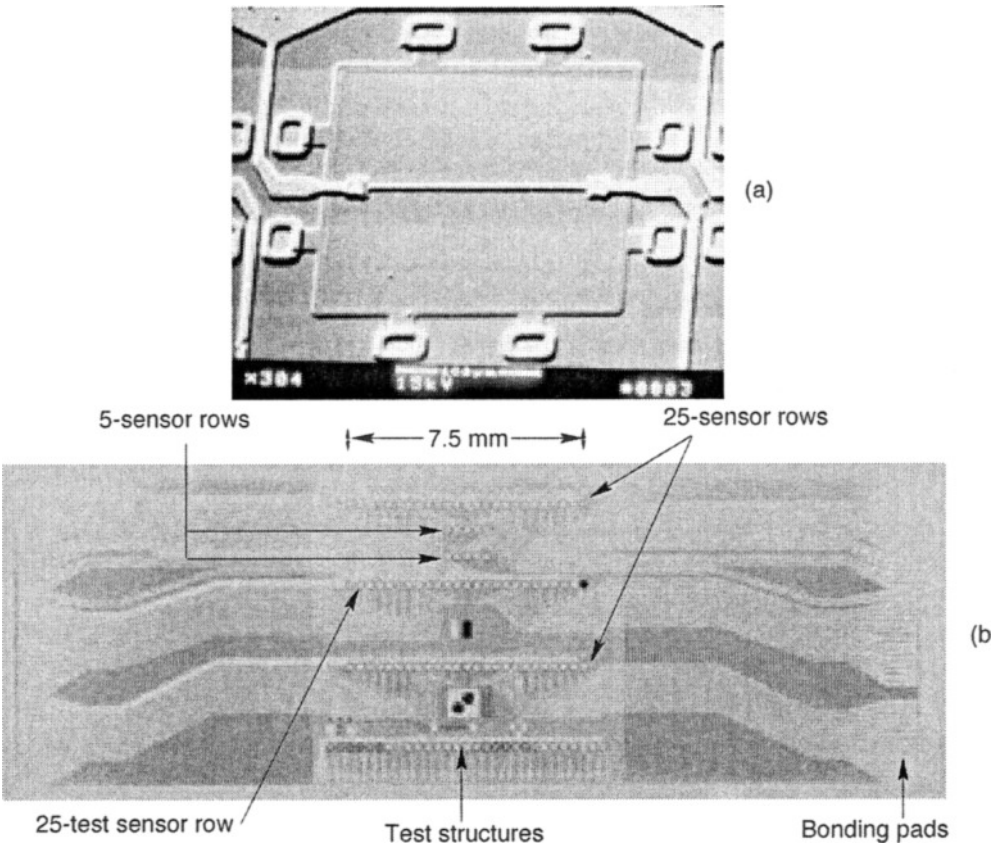


Fig. 14.6. (a) SEM image of individual sensor and (b) shear stress imaging chip.

used for process testing. The distance between rows is chosen such that at least four data points can be taken from a vortex-pair streak in the streamwise direction. Each sensor row has 25 sensors with 300-mm pitch, which is already the minimum for this type of sensor. The row should give at least three data points from a vortex-pair streak in the spanwise direction and should be able to image more than one streak. The 1 cm spacing between the sensors and the left and right edges of the chip is necessary to avoid the upstream bonding wires from interfering with the downstream sensors.

The fabrication process flow of the shear stress imager, depicted in Fig. 14.7, is as follows.

- Low-pressure chemical vapor deposition (LPCVD) is used to deposit 4000 Å low-stress silicon nitride (Si_xN_y).
- In patterned areas, the nitride is removed by plasma with a little over-etch to give a 7000–8000 Å cavity.
- The wafer is then put in an oxidation furnace to grow 1.7-μm-thick oxide in the cavity.
- After a short time etch in buffered hydrofluoric acid (BHF) to remove the oxidized nitride, 4000 Å of phosphosilicate glass (PSG) is deposited, patterned, and annealed to form the sacrificial layer etching channel.
- Next, 1.2 μm of LPCVD low-stress nitride is deposited as the diaphragm material.
- Etching holes are opened to expose the end of the PSG etching channel, and a 49% HF etching is done to completely remove the PSG and thermal oxide underneath the diaphragm.

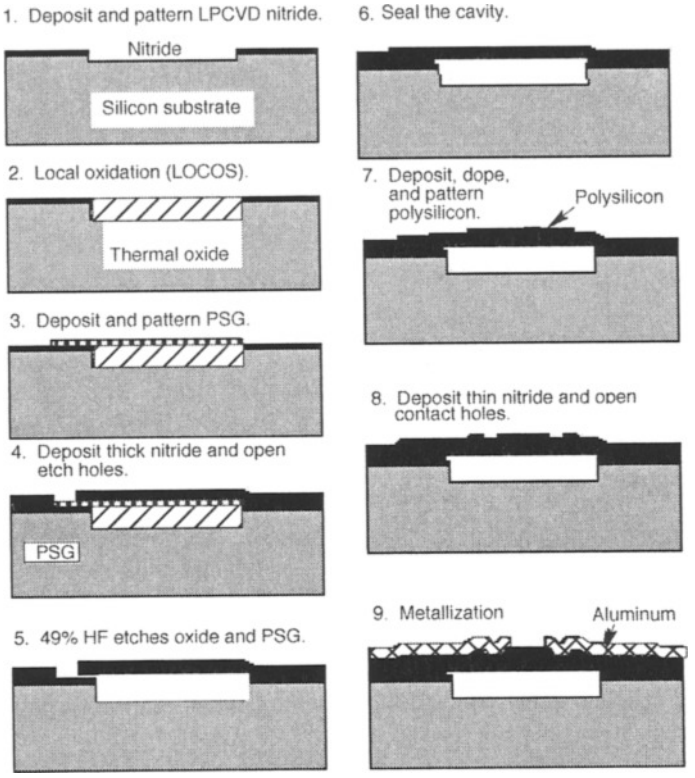


Fig. 14.7. Process flow for shear stress imager.

- The cavity is then sealed by LPCVD low-temperature oxide (LTO) and nitride deposition at a vacuum of 200 mtorr.
- The sealing materials on the diaphragm are removed by plasma and BHF etching to minimize the diaphragm thickness.
- A 4000-Å polysilicon layer is deposited, doped, annealed, and patterned to form the resistor on diaphragm.
- Another 2000 Å of low stress nitride is deposited to passivate the resistor.
- After the opening of the contact hole, metallization is done, and the wafer is diced to 2.85 × 1 cm chips.

The package for the imaging chip is a fine-line PC board with a recess in the center so that the imaging chip can be flush mounted. The chip and the PC board are electrically connected by wire bonding. The traces on the PCB then run to the edges, where through-holes electrically connect the front to the back of the PCB. Then, electrical leads are soldered on the back side of the PCB, which is flush mounted on a specially made plug that fits into the wall of the wind tunnel (Fig. 14.8), with the sensor row perpendicular to the flow direction.

The wind tunnel supplies a two-dimensional (2D) channel flow. The channel is 4.87 m long with a cross-sectional area of 60 × 2.5 cm. The walls of the channel are constructed of 2.5 cm thick Plexiglas and are supported by a steel frame. An axial blower powered by a dc source supplies the air flow in the channel. At the highest blower speed, the centerline velocity in the channel is about 25 m/s. Hot-wire velocity measurements at 10 m/s indicate that the channel consists of a laminar entrance flow region that gradually transforms into a fully developed turbulent flow in

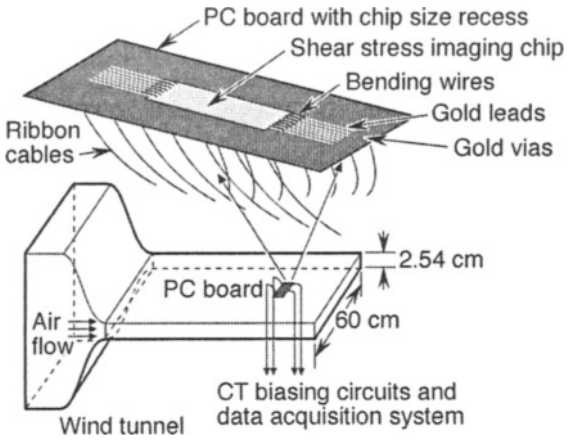
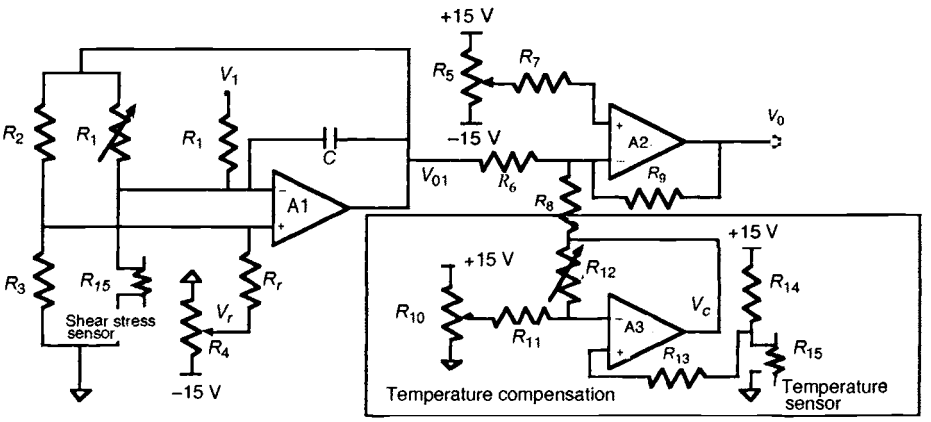


Fig. 14.8. Imaging chip packaging and experimental setup.

the downstream two-thirds portion of the channel. All calibration and testing of the imaging chip is carried out in this turbulent region of the channel.

In our experiments, the sensors are biased in constant temperature (CT) mode. Although the CT mode is more complicated than the constant current mode, it can achieve much higher frequency bandwidth, which is crucial in turbulence measurement. Therefore, arrays of CT circuits and gain stages have been made on PC boards using op-amps and discrete components (Fig. 14.9). The dc offset of the outputs can be adjusted individually, but the gain is fixed to 10. A computer-controlled data acquisition system is used to measure all the outputs simultaneously.



A1, A2, A3: OP-27
$R_3 = R_3/a_R$ (a_R: over-heat ratio)
C = 100–500 pF
R_1: 3.6 kΩ multiturn precision potentiometer
$R_2 = R_3 = R_6 = R_7 = R_8 = R_{11} = R_{13} = 5.0$ kΩ
R_{15}: temperature sensor - shear stress sensor operated at very low power
Temperature compensation stage can be shared by all channels

R_4, R_5, and R_{10}: 10 kΩ multiturn pots
$R_r = R_t = 1$ MΩ
$R_9 = 50$ kΩ
R_{12}: 470 kΩ multiturn precision pot
$R_{14} = 150$ kΩ

Fig. 14.9. CT biasing circuit, gain stage, and temperature compensation stage.

Before the sensors are used to measure the shear stress distribution, their dc outputs are calibrated against known wall shear stress levels, which are calculated from the centerline velocity by using the following experimentally derived relationship,[25]

$$\frac{u_\tau}{U_c} = 0.119 Re^{-0.089}$$

$$\tau_w = \rho_f u_\tau^2$$

$$\tau_w = 0.00427 U_c^{1.822}$$

where Re is the Reynolds number, U_c is the centerline velocity of the channel, u_τ is the friction velocity, τ_w is the wall shear stress, and ρ_f is the density of air.

Figure 14.10 shows the calibration results for 10 sensors in a row. Although each sensor has a different offset, the trend of all curves is almost the same. Polynomial fitting is performed on each curve to extract the fitting parameters for later use in real data processing.

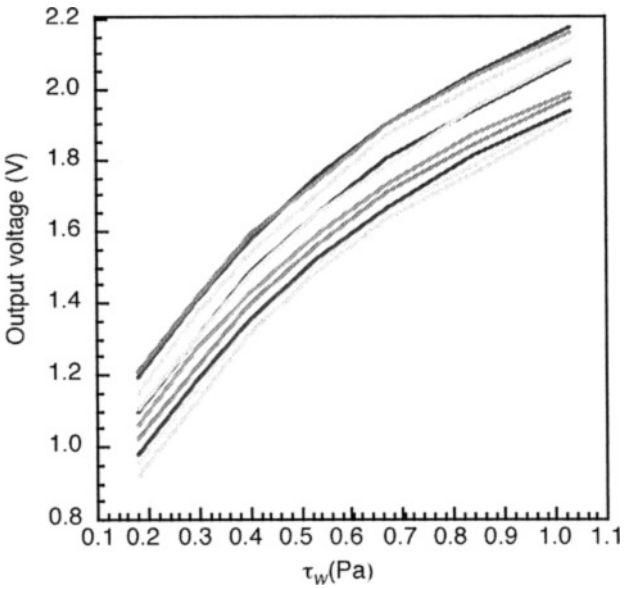


Fig. 14.10. Calibration curves of 10 sensors in a row

It is important to note that the dc outputs of the sensors are sensitive to the fluid temperature because the shear stress sensor is essentially a thermal sensor with a temperature coefficient of resistance around 0.1%/°C. This dc drift can be compensated by measuring the sensor temperature sensitivity and monitoring the fluid temperature change using a temperature sensor. Figure 14.11 shows that an order of magnitude of improvement on the thermal stability has been achieved by the compensation circuit given in Fig. 14.9. The temperature sensor used for compensation is just another shear stress sensor operated at very low power such that the self-heating is negligible.

To confirm that the sensors have enough bandwidth to pick up all the information in a turbulent flow investigation, the frequency response of a sensor in CT mode is measured using a

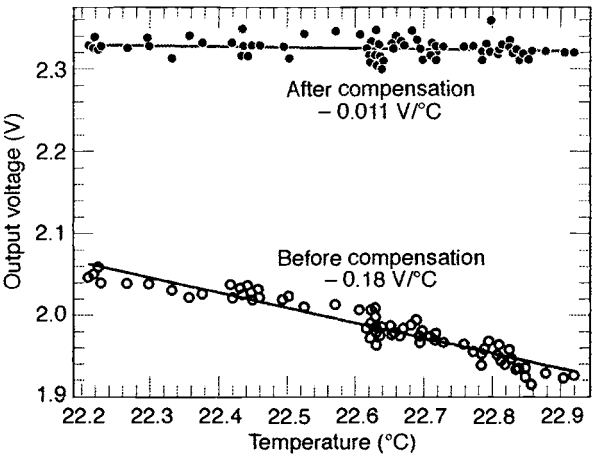


Fig. 14.11. Typical temperature sensitivities before and after temperature calibration.

sinusoidal electrical testing signal v_t. Figure 14.12 shows that this bandwidth reaches 30 kHz, which is sufficient for the turbulent flow under study. The deviation between the experimental data and the fitted curve results because a real sensor has multiple thermal time constants.

14.3 Shear Stress Measurements and Fluid/Actuator Interaction

With the successful fabrication of the shear stress sensor and shear stress sensor arrays, new fluid mechanical experiments could be run. First, individual sensor performance was validated in a turbulent flow using single-point measurements. Next, the array was used to image the vortices. In addition to imaging the vortices, the sensor arrays were used in conjunction with a neural net processor to determine the edges of the vortices. Finally, experiments were conducted to determine the effect of magnetic actuators (not described here but in Refs. 26 and 27) interacting with the flow.

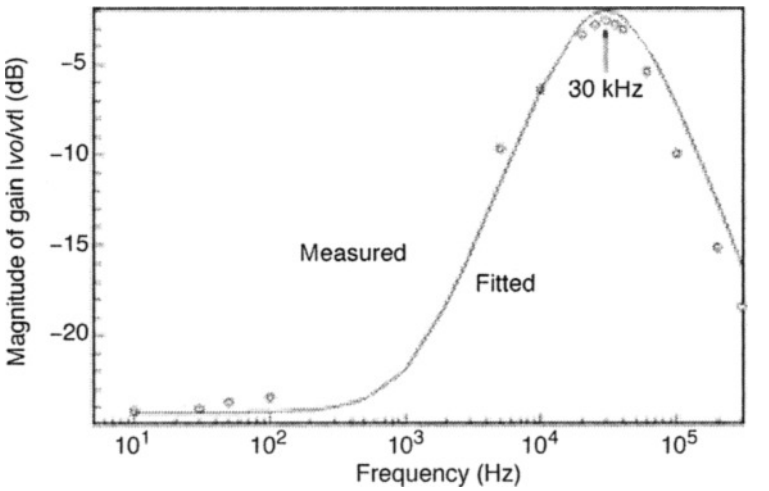


Fig. 14.12. Frequency response of a CT sensor. Derivation of the theoretical curve is shown in Ref. 28.

14.3.1 Single-Point Surface Shear Stress Measurement

For real-time shear stress imaging, the output voltage is sampled at 10 kHz and converted to a shear stress signal based on the calibration performed previously. In order to establish the credibility of the imaging chip, the turbulence statistics calculated from the shear stress fluctuations recorded by a single shear stress sensor are compared to previously established results. Figure 14.13 shows the comparison in terms of the normalized root-mean-square (rms) level, the skewness factor, and the flatness factor.[29] It is obvious that the current results agree very well with previous results in all three areas.[18–20] In addition, the current statistics appear to be independent of the Reynolds number, which is predicted by turbulence theory.

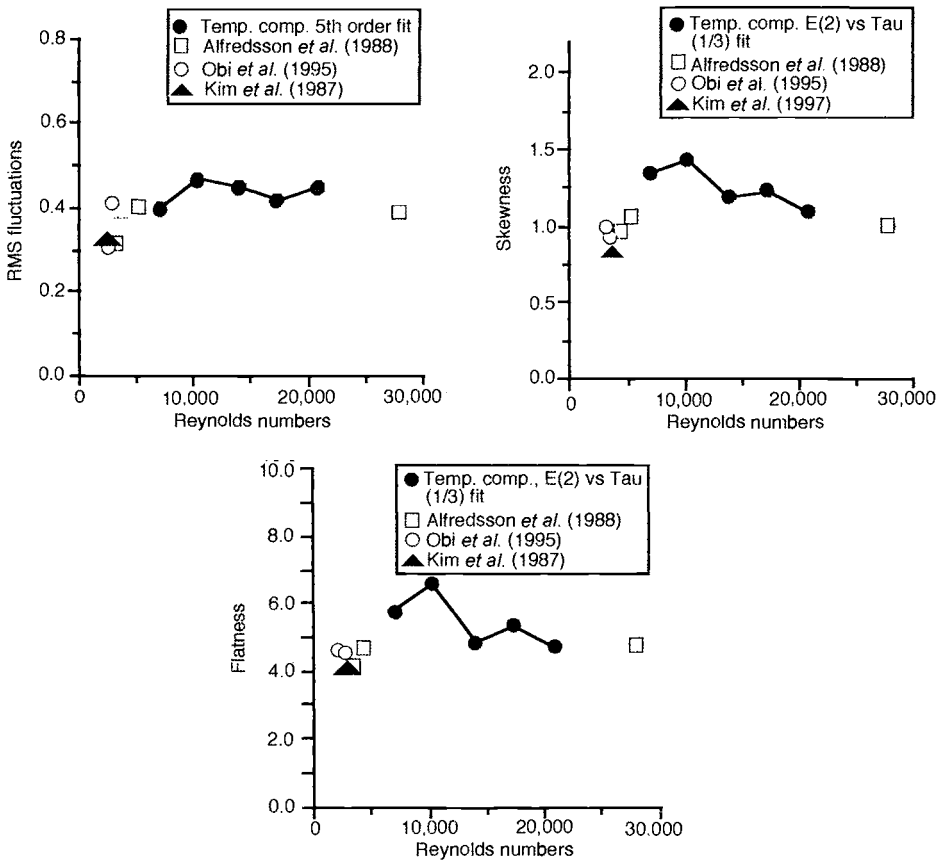


Fig. 14.13. Turbulence statistics measured by a single micro shear-stress sensor.

14.3.2 Distributed Measurement

Next, the instantaneous turbulent shear stress distributions of the channel were recorded by using one of the sensor rows on the imaging chip. Figure 14.14 shows the contour plots of the instantaneous shear stress distributions at two different centerline velocities. The white streaky structures in the plots represent regions of high shear stress on the wall of the channel where the imaging chip is located. They are caused by the presence of near-wall streamwise vortices, which bring high momentum fluid from the free stream to the wall. Because of the small-scale nature of these structures, previous experiments in turbulent boundary layers have only succeeded in

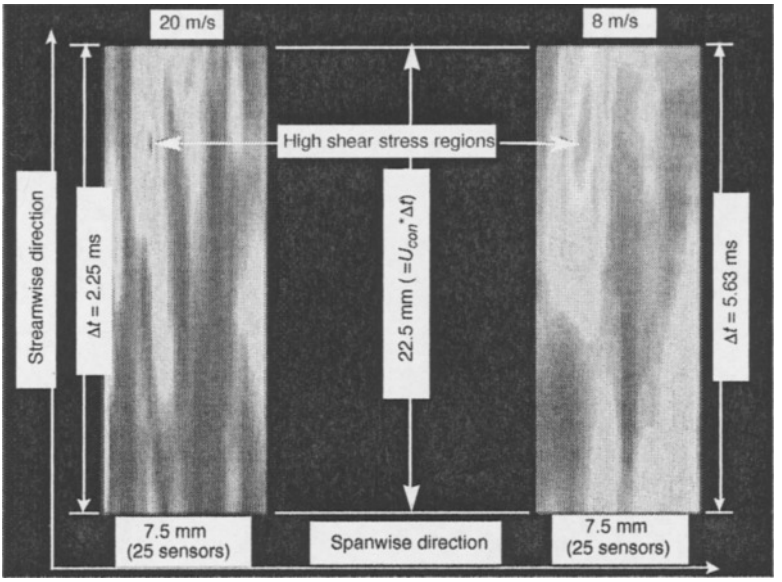


Fig. 14.14. Contour plots of the pseudo-2D shear stress distributions. Note high shear stress regions indicated by horizontal arrows.

qualitatively demonstrating their existence without obtaining any quantitative information. This is the first time that the instantaneous shear stress levels associated with the near-wall structures are recorded.

The contour plots in Fig. 14.14 indicate that the scales of the streaks are different at different centerline velocities. The streaks in the high-speed (20 m/s) case appear to be thinner and more densely packed than those in the low-speed (8 m/s) case. Similar phenomena have also been observed in previous experiments. The average streamwise length, the average spanwise scale, and the average spanwise spacing of the streaks at three different centerline velocities are estimated and shown in Fig. 14.15. The information shown is based on real-time movies generated from the contour plots similar to the ones in Fig. 14.14. Once again, the current results agree well with previous studies. The average length of the current streaks falls within the upper and lower bounds of the established results.[18–20] The average spanwise scale and spacing of the current streaks are either on, or slightly higher than, the upper bound.

14.3.3 Edge Detector for Identifying High Shear Stress Regions

From our experimental results, it appears that the high shear stress streaks are randomly distributed. A real-time detection scheme is needed to properly identify the streaks among the background fluctuations so that the downstream actuator can be activated. A neural network-based detection circuit (described subsequently) was developed for this purpose. In this circuit, the signal output of a sensor is compared with that of its neighboring sensors through a filter-threshold combination. Whenever a sudden change is detected, the boundary of the high shear stress region is marked. This circuit has been used in conjunction with the imaging chip. A typical result is shown in Fig. 14.16. The light region in the control output of Fig. 14.16 (b) represents "positive" identification of high shear stress streaks. The good matching between the light region and the high shear stress area (light gray region) measured by the imaging chip indicates the effectiveness of the detection circuit.

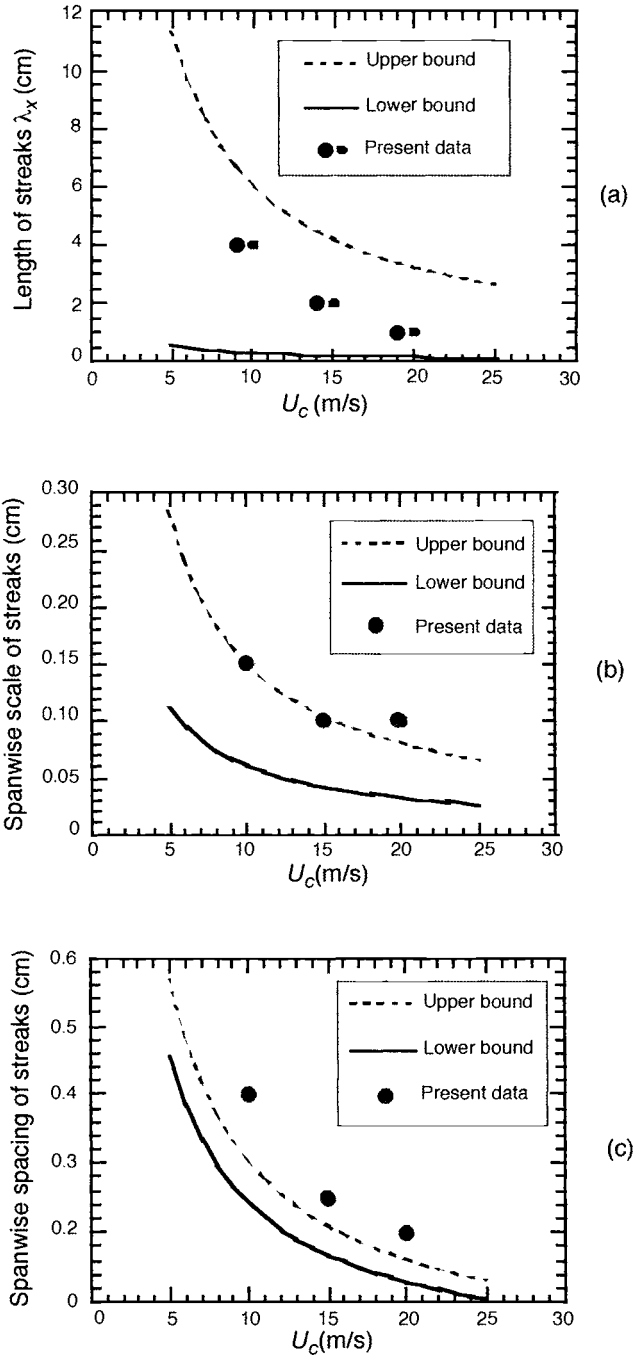
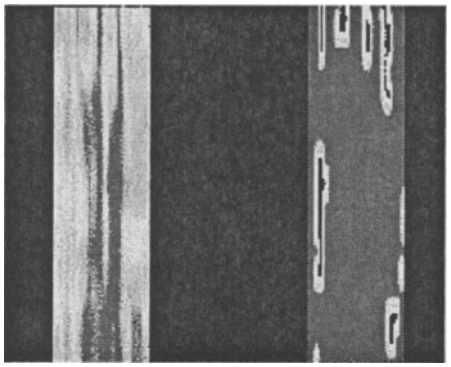


Fig. 14.15. Scales of the near-wall streaky structures at different Reynolds numbers. (a) Length of streaks, (b) spanwise scale of streaks, (c) spanwise spacing of streaks. (From Refs. 18 to 20.)



(a) (b)

Fig. 14.16. Contours of (a) the instantaneous surface shear stress and (b) the output of the control circuits.

14.3.4 Interaction between Micromachined Actuator and Streamwise Vortices

To properly control the high shear stress streaks with microactuators,[26,27] an in-depth understanding of the interaction between the actuators and the flow structures is required. Experiments have been carried out to investigate the interaction between a single high shear stress streak and a micromachined actuator. In this study, a 1.3-mm-thick vortex generator is used to generate a longitudinal vortex pair in a laminar boundary layer. The micromachined actuator used is a silicon (Si) flap with 30 turns of copper coil. The flap matches the physical size of the longitudinal vortex pair. Oscillation of the flap is achieved by the combination of an external permanent magnet and an ac coil current. The maximum deflection of the flap is about 30 deg, which corresponds to a height of 2 mm from the tip of the actuator flap to the wall. The maximum oscillation frequency is about 100 Hz. The actuator is placed downstream from the vortex generator, where the longitudinal vortex pair is generated. Experiments are carried out for different combinations of actuator frequency (ω) and maximum tip height (d). For each combination, a two-component (streamwise and spanwise) hot-wire anemometer is used to measure the velocity distributions downstream from the actuator. From the velocity distributions, the vorticity and surface shear stress distributions are calculated and phase averaged.

The longitudinal vortices convect high-speed fluid to the wall, inducing a local high shear stress streak (Fig. 14.17) on the surface of the wall. As the actuator oscillates, it generates perturbations in the flow, which change the shear stress intensity of the streak. As indicated by the ensemble-average result shown in Fig. 14.17, the intensity of the streak decreases as the flap actuator deflects away from the surface, reaching a minimum at the maximum actuator tip height (phase = π). As the actuator moves back to the surface, the streak intensity reverses to its original level. To evaluate the net effect of the actuator oscillation on the shear stress distribution, the net shear stress coefficient, C_{DN}, is evaluated for different combinations of ω and d and is shown in Fig. 14.18. C_{DN} is a measure of whether the drag is higher or lower in the area behind an actuator and is defined as

$$C_{DN} = \int_0^{2\pi} [C_D(\theta) - C_{DVG}]d\theta \qquad \text{and} \qquad C_D(\theta) = \frac{1}{0.5\rho U^2}\int_{-z}^{z} \mu \frac{\partial u}{\partial y}$$

where $C_D(\theta)$ is the coefficient of friction (a normalized skin friction drag) as a function of the angle (θ) of the oscillating actuator, U is the velocity, and ρ is the density. C_{DVG} is the coefficient of friction in the area beyond the vortex generator. More specifically, the C_{DVG} in the equation

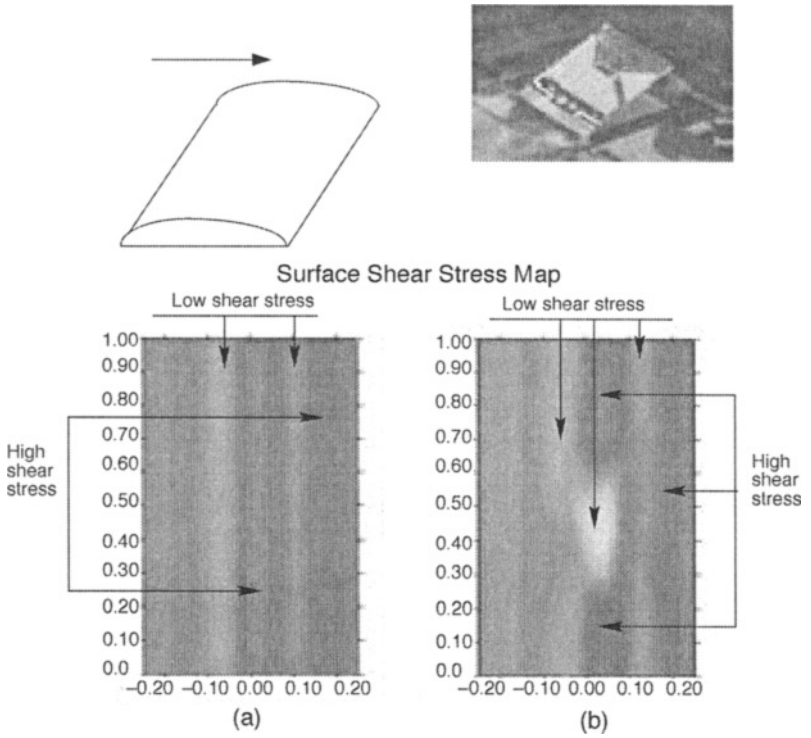


Fig. 14.17. Contours of ensemble-average dU/dy. (a) Vortex generator only; (b) vortex generator and an oscillating microactuator.

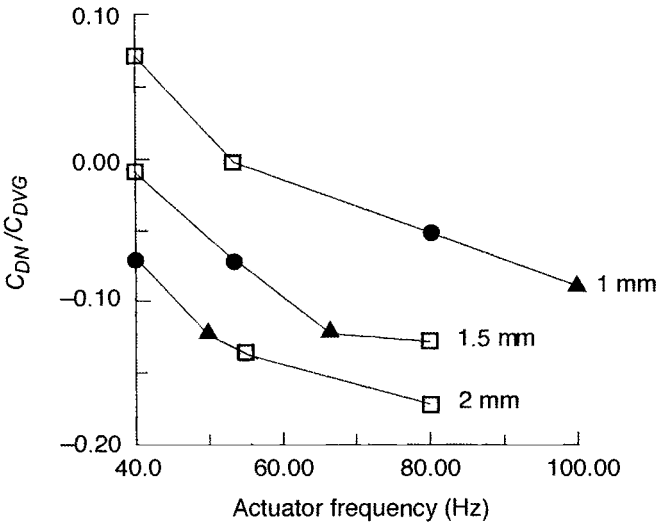


Fig. 14.18. Variation of the normalized shear stress coefficient C_{DN}/C_{DVG} with actuator frequency ω and maximum actuator tip height d. Each line is the result of one d, as indicated at the end of the line. The solid circles correspond to an ωd of 80, and the solid triangles correspond to an ωd of 100.

is the time-invariant shear stress coefficient associated with the stationary high shear stress streak induced by the vortex generator. Defined this way, C_{DN} indicates a shear stress increase, if positive, and a drag decrease, if negative. As shown in Fig. 14.18, higher drag reduction is achieved at higher ω and higher d. In addition, similar shear stress reduction results if the product of ω and d is constant. Since the product of ω and d is a measurement of the transverse velocity of the actuator flap, this result indicates that the amount of shear stress reduction is directly related to the transport of high-speed fluid away from the surface by the vertical motion actuator.

14.4 Integration

As shown in the preceding section, the results of sensor/flow and actuator/flow interaction are very promising. The next step, then, is to integrate sensors, actuators, and electronics on the same substrate to form the M^3 system. Toward this goal, an integrated M^3 chip was fabricated but has not yet been extensively tested.[30] There are many generic concerns in MEMS-electronics integration that have been addressed in M^3 chip fabrication. Elements of this MEMS-electronics integration effort can be applied to other, more generic, integration efforts.

The primary concern of all integration efforts revolves around the choice of when the MEMS processing steps are completed relative to the integrated chip (IC) processing steps. There are three options: (1) an interweaved process, (2) electronics first, followed by MEMS, and (3) MEMS first, followed by electronics. Each option has its advantages and disadvantages. In short, the interweaved process is typically the best from a technical point of view, while the electronics-first process is often the most practical process. The MEMS-first process is only attempted by very limited research groups and will not be discussed here.

One assumption made while comparing these different options is that the MEMS designer does not have full and high-priority access to an IC-capable fabrication facility. This is a reasonable assumption for an academic environment where few laboratories even have basic MEMS capabilities, let alone IC fabrication facilities. In the few U.S. institutions that do have complementary metal oxide semiconductor (CMOS) capabilities (e.g., University of California, Berkeley, and Stanford), the electronics yield is usually unacceptably low compared with that of industry. The lack of a dedicated access to a fabrication line is not limited to academia. Even in a corporation with large fabrication facilities, internal MEMS research facilities often are completely separate from very large-scale integration (VLSI) lines, and MEMS researchers often are not allowed full access to such lines. If full access were available to MEMS researchers, the interweaved process would almost definitely be the choice, for reasons that will be discussed subsequently.

From a device robustness point of view, the interweaved process is ideal. The order in which steps are completed reflects a process flow designed with optimum device performance in mind. This is in stark contrast to the other options, which often require steps that are not needed for the eventual devices (or electronics) per se, but rather are crucial for the overall survival of the process. For example, in an electronics-first process, during the MEMS processing, layers are often deposited, patterned, and later removed simply to protect the aluminum (Al) metallization on the electronics portion of the wafer. In an interweaved process, the metallization steps could occur after any harsh MEMS processing steps have been completed. From this, it can be seen that the interweaved process also is usually the shortest process (both in time and in number of processing steps). General processing wisdom holds that the fewer the steps, the more robust the process. For commercial applications, fewer steps also reduce the cost significantly while increasing the yield.

Extremely few IC facilities will allow preprocessed wafers to enter their fabrication lines because of fear of contamination, especially the fear of unknown (by VLSI standards) MEMS materials. Therefore, finishing the IC processing first has one primary advantage: many choices of

IC foundries can provide wafer-level electronics. After the foundry fabricates the IC electronics, the wafers are then subject to MEMS processing, which can be completed in a facility separate from the VLSI fabrication line.

Often, the choice of which option to use for integration depends on metallization concerns. Typically, VLSI fabrication uses Al metallization. Al is considered a low-temperature material and cannot withstand processing temperatures above 450°C. Therefore, certain steps such as diffusion and oxidation cannot be attempted after Al deposition. If electronics, including metallization, is completed before MEMS processing, only low-temperature steps can be used to finish the process. Other considerations regarding metallization are (1) ideally, only one deposition/pattern/etch step should be used for each metal layer and (2) fine line widths are often required for the electronics portion of the chip. These two concerns may imply that the intelligent approach is to complete all the metallization (for a given layer of metal) at one time and to do it at the VLSI fabrication facility, where fine-line metal patterning and etching are well-characterized steps. There may exist some facilities that will take preprocessed wafers on a one-time basis. For these facilities, MEMS processing, up to but not including metallization, can be done before submitting the wafers to the foundry. Complete CMOS processing is then completed, with the metallization connecting both electronic and MEMS devices.

Finally, although less interesting from a research point of view, economic reasons can often dictate which option is chosen. When choosing an outside foundry as a vendor, the least expensive option involves requesting as standard a package as possible. This usually implies that, of the three above-mentioned options, the second (electronics first, followed by MEMS) option may be the most viable. Cost, however, is often not the only economic concern. The idea of interweaving a process is often the most appealing option from a technological point of view. From a foundry's point of view, it is also the most costly option. While most foundries will not turn down business that requires standard technology, they do require financial justification before attempting an expensive custom run, which could divert their labor and resources from other high-volume projects. Foundries such as Standard Microsystems Corporation (SMC), which have dedicated MEMS-electronics fabrication facilities, often also have long lines of corporate customers buying high volumes of devices. They often are not interested in low-volume applications (such as university research or small-company products).

Therefore, small-volume applications are limited to pursuing the electronics first, followed by the MEMS option. If high-temperature MEMS steps are required, it is possible to switch to an interweaved approach, where the electronics processing up to, but not including, metallization is completed at the VLSI foundry. Next come the MEMS steps, and the process is completed with metallization at the MEMS facility (or another facility willing to take preprocessed wafers). This option requires the ability to dry-etch contact holes, to pattern fine lines, and to dry-etch Al in the post-IC facility. This ability may or may not be a limitation. This is the option we chose to complete our first attempt at integration. A picture of a completed chip is shown in Fig. 14.19, and an abbreviated process flow is shown in Fig. 14.20.

14.5 Control

A new adaptive controller based on a neural network was constructed and applied to our system for drag reduction. A simple control network is employed that directs blowing and suction at the surface, based only on the wall shear stresses in the spanwise direction. Such a network was shown to reduce the skin friction by as much as 20% in direct numerical simulations of a low-Reynolds-number turbulent channel flow. Also, a stable pattern was observed in the distribution of weights associated with the neural network. This allowed us to derive a simple control scheme

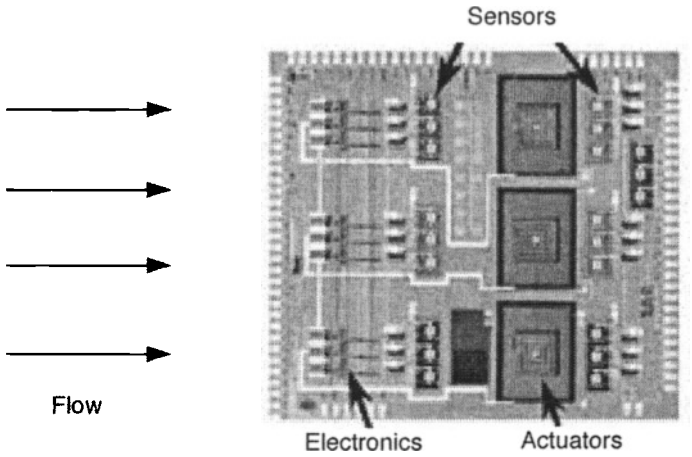


Fig. 14.19. Picture of a fabricated integration chip.

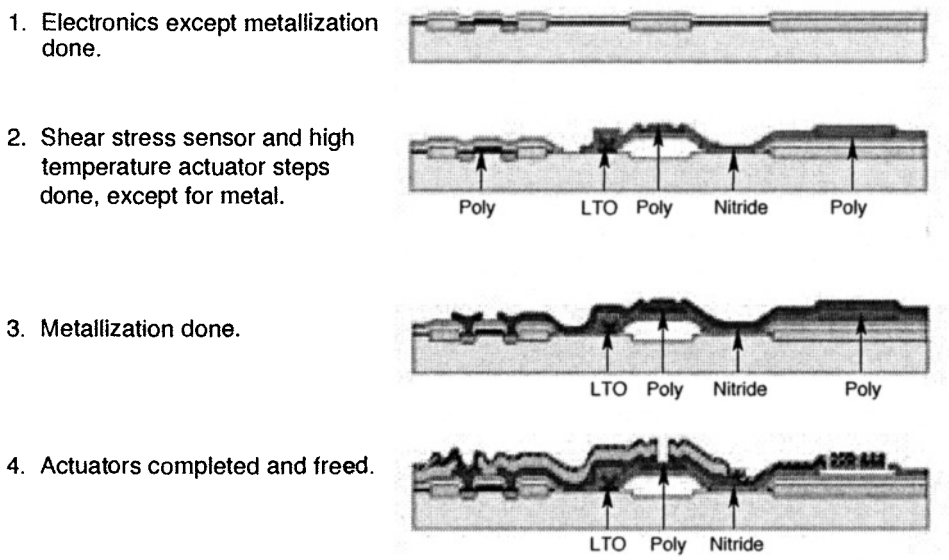


Fig. 14.20. Abbreviated integration process flow.

that produced the same amount of drag reduction. This simple control scheme generated optimum wall blowing and suction proportional to a local sum of the wall shear stress in the spanwise direction. The distribution of corresponding weights was simple and localized, thus making real implementation relatively easy.

Although construction of a neural network generally requires no prior knowledge of the system, knowledge about the near-wall turbulence structures provides a guideline for the design of the network architecture. Initially, $\partial u/\partial y$ and $\partial w/\partial y$ (w is the velocity in the spanwise direction and is perpendicular to u, which is in the streamwise direction) at the wall at several instances of time were used as input data fields. The actuation at the wall was used for the output data of the network. Experimentally, we found that only $\partial w/\partial y$ at the wall from the current instance of time was necessary for sufficient network performance. Because we wanted the output to be based

only on a local input area, we designed our network using shared weights. The network had a single set of weights (a template) that was convolved over the entire input space to generate output values; that is, we used the same set of weights for each data point, and the training involved iterating over all data points. The template extracted spatially invariant correlations between input and output data. The size of the template was initially chosen to include information about a single streak and streamwise vortex, and then was varied to find an optimal size.

We used a standard two-layer feed-forward network with hyperbolic-tangent hidden units and a linear output unit (see Fig. 4.21). The functional form of our final neural network was:

$$v_{jk} = W_a \tanh\left(\sum_{i=-\frac{N-1}{2}}^{\frac{N-1}{2}} W_i \frac{\partial w}{\partial y}\bigg|_{j,\,k+li} - W_b \right) - W_c, \quad 1 \le j \le N_x \text{ and } 1 \le k \le N_z \tag{14.1}$$

where the W's denote the weights, N is the total number of input weights, and the subscripts j and k denote the numerical grid point at the wall in the streamwise and spanwise directions, respectively. N_x and N_z are the number of computational domain grid points in each direction. The summation was done over the spanwise direction. Seven neighboring points (N = 7), as well as the point of interest, were included in the spanwise direction (corresponding to approximately 90 wall units with our numerical resolution). These points were found to provide enough information to adequately train and control the near-wall structures responsible for the high skin friction. Note that the blowing and suction are applied at each grid location according to the above equation as a numerical approximation of distributed blowing suction on the surface. A scaled conjugate gradient learning algorithm[31] was used to produce rapid training. For given pairs of $\left(v_{jk}^{des}, \frac{\partial w}{\partial y}\Big|_{jk}\right)$, the network was trained to minimize the sum of a weighted-squared error given

$$\text{by } Error = \frac{1}{2}\sum_j \sum_k e^{\lambda\left|v_{jk}^{des}\right|}\left(v_{jk}^{des} - v_{jk}^{net}\right)^2 \tag{14.2}$$

where v^{des} is the desired output value and v^{net} is the network output value given by Eq. (14.1).

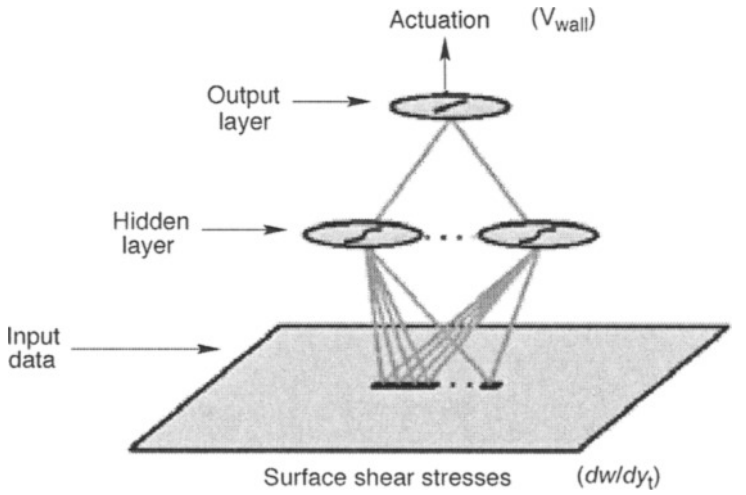


Fig. 14.21. Neural network architecture.

The weights were initialized with a set of random numbers. Note that the error defined in Eq. (14.2) exponentially emphasizes (proportional to λ) large actuations. This error scaling was chosen based on the observation by Choi *et al.*[10] that large actuations are more important for drag reduction. Usually, within 100 training epochs, the error reached its asymptotic limit.

The computed flow fields for a no-control case and a successful control case, based on a 7-point weighted sum of $\partial w/\partial y|_w$,[32] were examined to investigate the mechanism by which the drag reduction is achieved. The most salient feature of the controlled case was that the strength of the near-wall streamwise vortices was drastically reduced. In Fig. 14.22, contours of streamwise vorticity in a cross plane are shown. The reduction of the strength further substantiates the notion that a successful suppression of the near-wall streamwise vortices leads to a significant reduction in drag. Note that for the controlled case, the wall actuations were applied at both walls.

The probability-density function of the wall-shear stress in the streamwise direction is shown in Fig. 14.23. It is evident that the control case is very effective in suppressing large fluctuations,

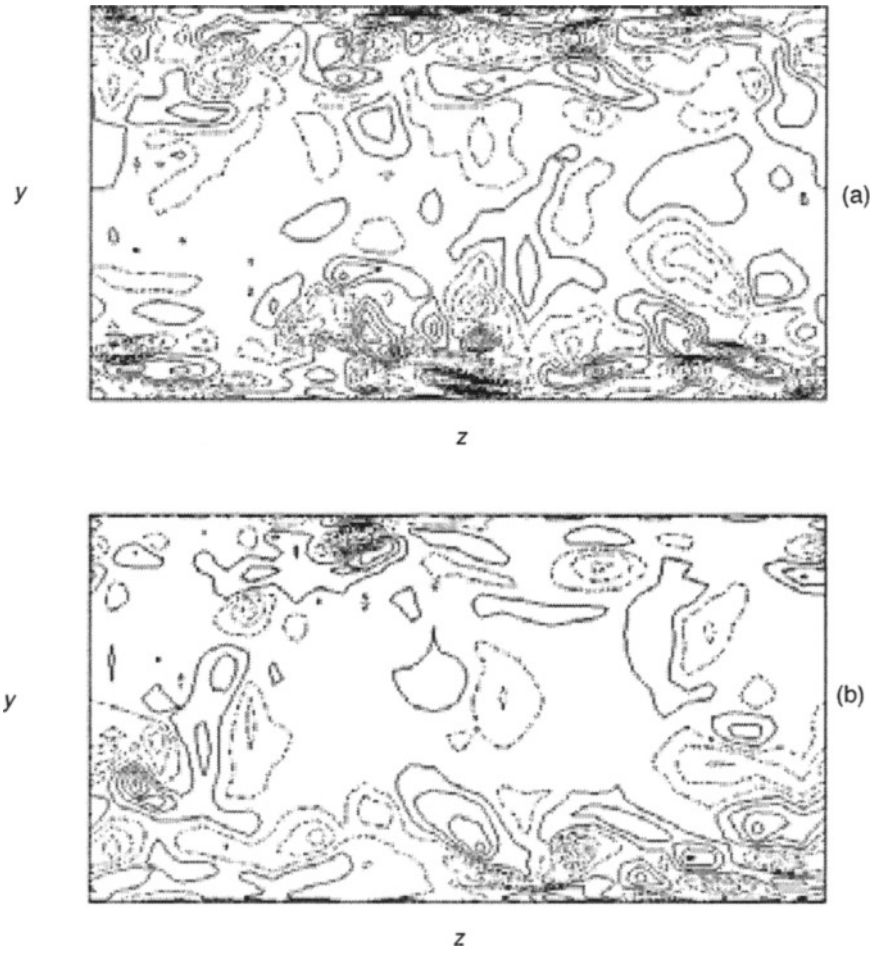


Fig. 14.22. Contours of streamwise vorticity in a cross-flow plane. (a) No control. (b) control using seven fixed weights. The contour level increment is the same for (a) and (b). Negative contours are shown by chains of dots.

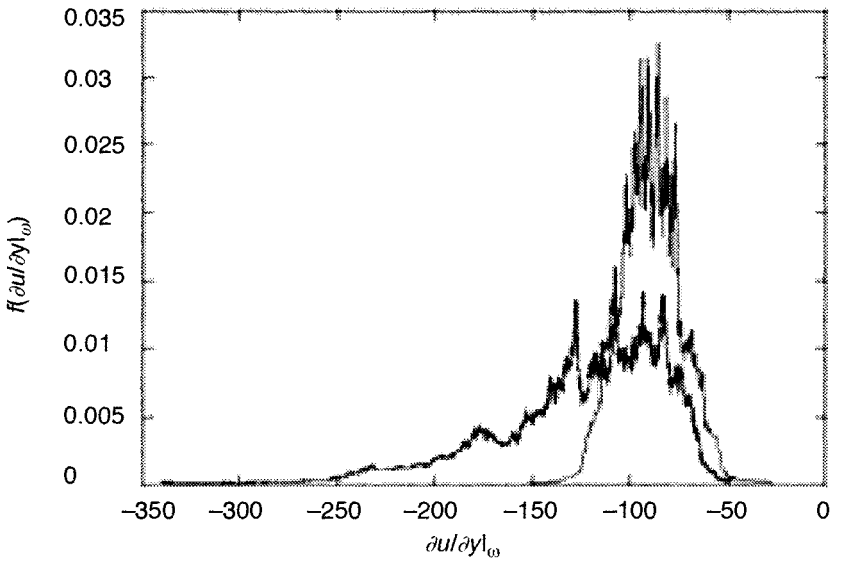


Fig. 14.23. Probability-density function of wall-shear stress: line showing wide distribution is with no control: line showing narrow distribution is with control and seven fixed weights. Area under each curve is normalized to one.

thus reducing the mean skin friction. Furthermore, the rms values of turbulent fluctuations in the wall region are also reduced, as shown in Fig. 14.24(a). The same trend was observed by Choi *et al.*[10] The rms vorticity fluctuations are also significantly reduced, except for ω_x, very close to the wall. The increase for ω_x is caused by additional $\partial v/\partial z$ due to the wall actuation. The rms fluctuations of ω_z at the wall, which are mainly due to $\partial u/\partial y$ at the wall, are also decreased. This decrease indicates that the control scheme led to a reduction in the mean shear and its variance at the wall by suppressing large fluctuations, as shown in Fig. 14.23. The reduction in the rms fluctuations in ω_x and ω_y indicates that the control scheme indeed reduced the strength of the near-wall streamwise vortices and the wall-layer streaks.

Figure 14.25 compares the distribution of wall actuation used in our control with that from Choi *et al.*[10] v-control using the information at $y^+ = 10$ for the same wall shear stress distribution.The corresponding wall shear stress distribution is also shown in Fig. 14.25. The wall actuations indicate a strikingly similar distribution to each other, even though the wall actuation of our control is based only on the wall shear stress $\partial w/\partial y|_w$. Basically, our control scheme detects edges of locally high-shear stress regions by measuring the spanwise variation of $\partial w/\partial y$, and applies appropriate wall actuation, as shown in Figs. 14.25(a) and 14.25(b). Since high-shear stress regions are usually elongated in the streamwise direction, only spanwise variation is necessary for detecting the edges. Since $\partial w/\partial y$ is a direct measurement of streamwise vortices, it provides more appropriate information than $\partial u/\partial y$. This is consistent with our finding that $\partial u/\partial y$ at several points in the spanwise direction is enough for good performance of control.[32]

14.6 Electronics

Circuits process the signals from the sensors to find regions of high shear stress. This detection process uses information about the spatial and temporal nature of the vortex-pair streaks. First, the long and narrow aspect ratio leads to building "column"-oriented templates for streak detection. We organize the sensor outputs into thin feature detectors oriented in the direction of the air

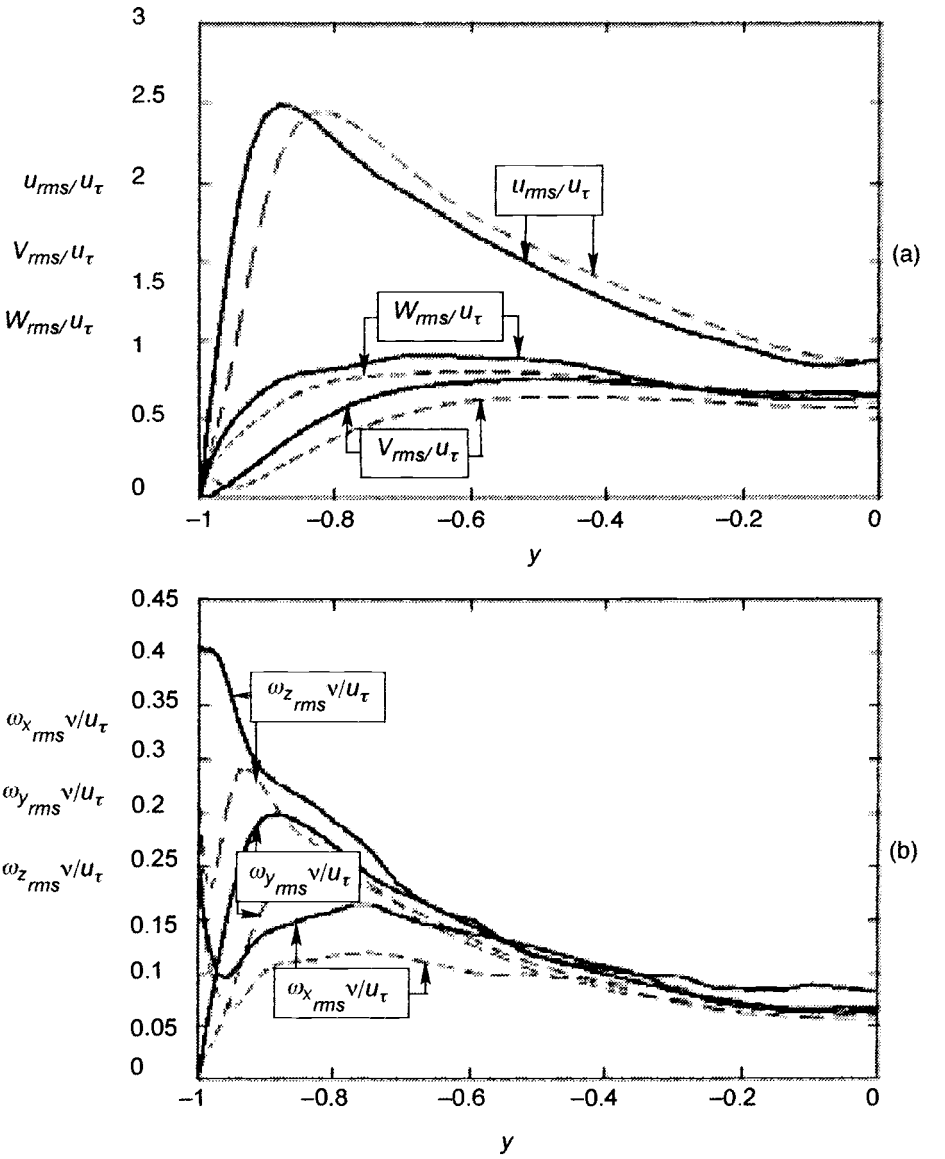


Fig. 14.24. Root-mean-square fluctuations normalized by the wall variables. Solid line, no control; dashed line, control with seven fixed weights. (a) Velocity fluctuations, (b) vorticity fluctuations.

flow. When several sensors in a column register either a larger or smaller output than their neighbors in a spanwise direction, this difference accumulates. If this accumulated difference exceeds a threshold, a vortex pair streak may be present in that column. The appropriate control action raises the associated actuator.

Figure 14.26 shows a plot of the detection and control chip. The constant temperature output sensor signal feeds into a further stage of amplification. A buffer distributes the amplified signal to a nonlinear resistive network composed of horizontal resistor (HRES) circuits.[33] Sensor

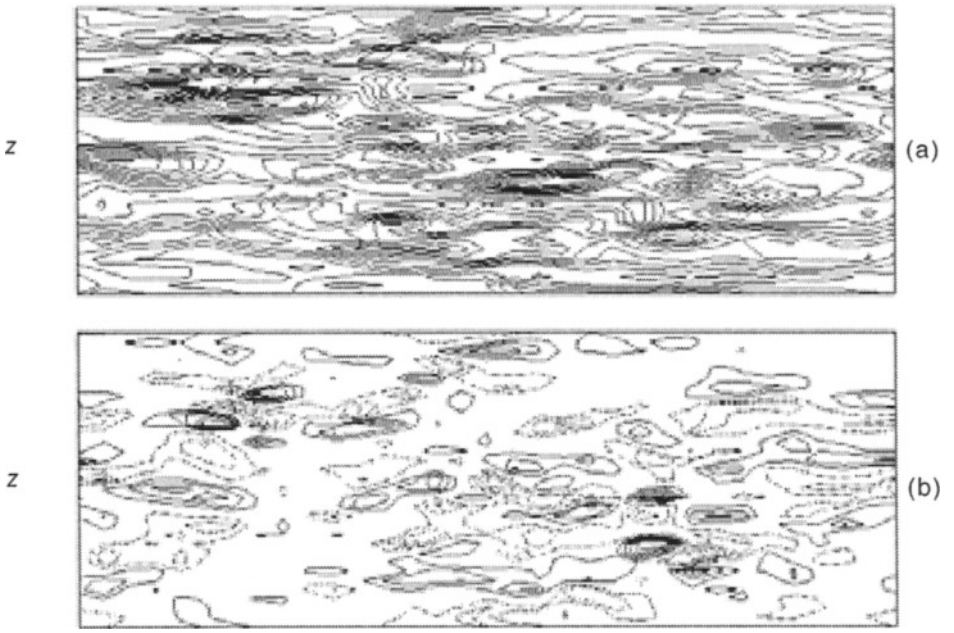


Fig. 14.25. Contours of the wall actuation: (a) control using $\partial w/\partial y|_w$ with 7 fixed weights, (b) control using information at y+ = 10. Negative contours are shown by chains of dots.

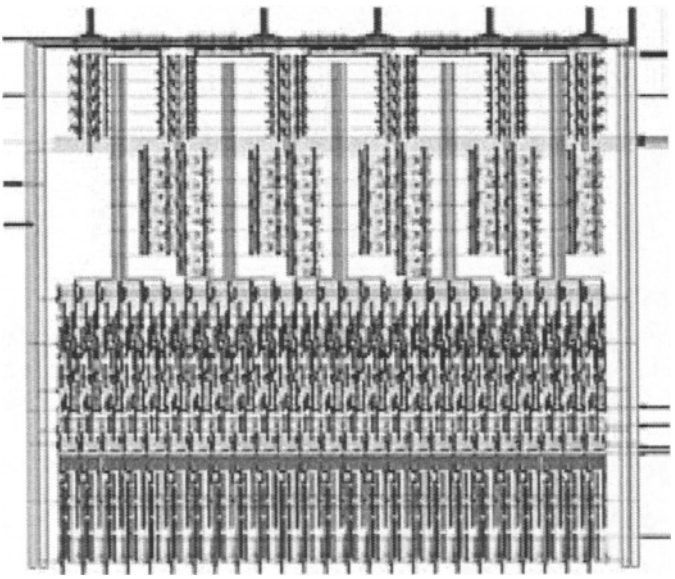


Fig. 14.26. Layout of circuitry on IC.

outputs in the same column and sensor outputs in adjacent columns use different spatial filtering constants. The different constants reinforce activity within a column and discourage activity between adjacent columns. The filtered signals feed to a symmetric antibump circuit.[34] The circuit's

operation mimics that of a soft comparator with an adjustable dead zone. The function of the circuit is to indicate when a particular column has registered a large shear stress value, while the neighboring columns have not registered such a value. The output of the antibump circuit, a current, accumulates for a particular column and is compared to a threshold. If the accumulated value exceeds the threshold, the circuit triggers the actuator by turning on a pull-down transistor.

We designed this system to reduce the fully turbulent drag in our experimental setup. We present the system with a fully turbulent airflow profile. Figure 4.27 graphs the single-column response of the system.

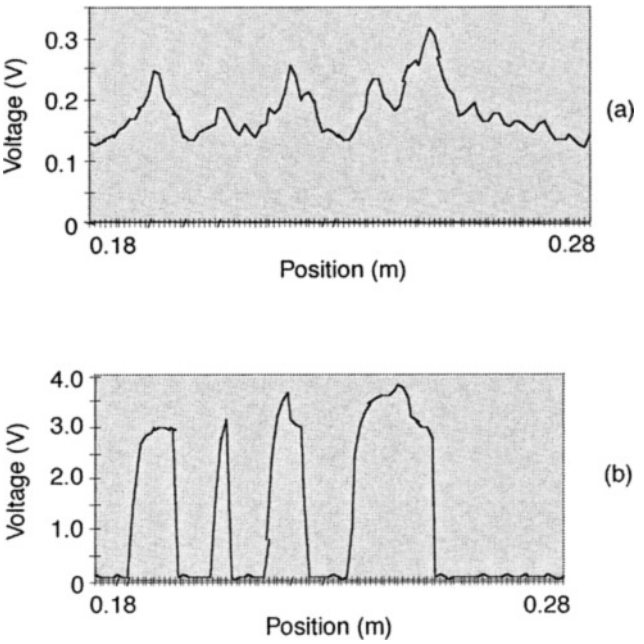


Fig. 14.27. Graph of the output waveforms of one shear stress sensor and the corresponding channel from the detection/control chip. We record the data in a fully developed turbulent flow. (a) Shear stress sensor output, (b) control detection chip actuator control.

14.7 Conclusions

The effort to reduce drag in turbulent boundary layers requires the collaboration of engineers in many different fields. The individual contributions within each field (MEMS, fluids, controls, and electronics) are novel, even as a stand-alone effort. When combined, such efforts can lead to tremendously exciting discoveries. Much work needs to be done in the future to obtain the desired drag reduction, but efforts to date have proved to be extremely promising and worthwhile.

14.8 References

1. M. Walsh, "Riblets as a Viscous Drag Reduction Technique," *AIAA J.* **21** (4), 485–486 (1983).
2. "World Airline Passenger Traffic Growth Rate to Continue Through to 1998," News release, International Civil Aviation Organization (ICAO), 1996.
3. B. Cantwell, "Organized Motion in Turbulent Flow," *Ann. Rev. Fluid Mech.* **13**, 457–515 (1981).
4. D. Bechert and M. Bartenwerfer, "The Viscous Flow on Surfaces with Longitudinal Ribs," *J. Fluid Mech.* **206**, 105–129 (1989).

5. B. Lazos and S. Wilkinson, "Turbulent Viscous Drag Reduction with Thin-Element Riblets," *AIAA J.* **26** (4), 496–498 (1988).
6. S.-R. Park and J. Wallace, "Flow Alteration and Drag Reduction by Riblets in a Turbulent Boundary Layer," *AIAA J.* **32** (1), 31–38 (1994).
7. L. Sirovich and S. Karisson, "Turbulent Drag Reduction by Passive Mechanisms," *Nature* **388**, 753–755 (1997).
8. P. Vukoslavcevic, J. Wallace, and J.-L. Balint, "Viscous Drag Reduction Using Streamwise-Aligned Riblets," *AIAA J.* **30** (4), 1119–1122 (1992).
9. J. McLean, D. George-Falvy, and P. Sullivan, "Flight-Test of Turbulent Skin-Friction Reduction by Riblets," in *Turbulent Drag Reduction by Passive Means*, Proceedings of the Royal Aeronautical Society (1987), Vol. II., pp. 408–424.
10. H. Choi, P. Moin, and J. Kim, "Active Turbulence Control for Drag Reduction in Wall-Bounded Flows," *J. Fluid Mech.* **262**, 75–110 (1994).
11. D. Bushnell, and K. Moore, "Drag Reduction in Nature," *Ann. Rev. Fluid Mech.* **23** 65–79 (1991).
12. J. W. Hoyt, "Hydrodynamic Drag Reduction Due to Fish Slimes," in *Swimming and Flying in Nature*, edited by T. Wu, C. Brokaw, and C. Brenne (Plenum, New York, 1975), Vol. 2, pp. 653–672.
13. D. Bechert, M. Bartenwerfer, and G. Hoppe, "Drag Reduction Mechanisms Derived from Shark Skin," Paper 86-1.8.3, *15th Congress of the International Council of the Aeronautical Sciences*, 1986.
14. J. Gray, "Studies in Animal Locomotion. VI. The Propulsive Powers of the Dolphin," *J. Experimental Biology* **13**, 192–199 (1936).
15. D. Au, and D. Weihs, "At High Speeds Dolphins Save Energy by Leaping," *Nature* **284**, 548–550 (1980).
16. T. Lang and K. Pryor, "Hydrodynamic Performance of Porpoises (Stenalla atenuata)," *Science* **152**, 531–533.
17. M. Kramer, "The Dolphins' Secret," *J. Am. Soc. Naval Engin.* **73**, 103–107 (1961).
18. P. Ferredoxin, A. Johansson, J. Haritonidis, and H. Eckelman, "The Fluctuating Wall-Shear Stress and the Velocity Field in the Viscous Sublayer," *Phys. of Fluids* **31**, 1026–1033 (1988).
19. S. Obi, K. Inoue, T. Furukawa, and S. Masuda, "Experimental Study on the Statistics of Wall Shear Stress in Turbulent Channel Flows," *10th Symposium on Turbulent Shear Flows* (Pennsylvania State University, 1995), Vol. 1, pp. 5-19 to 5-24.
20. J. Kim, P. Moin, and R. Moser, "Turbulence Statistics in Fully Developed Channel Flow at Low Reynolds Number," *J. Fluid Mech.* **177**, 133–166 (1987).
21. R. Goodstein, *Fluid Mechanics Measurements*, Chap. 11 (Hemisphere Publish Corp., 1983), pp. 559–615.
22. C. Liu, Y.-C. Tai, J.-B. Huang, and C. M. Ho, "Surface Micromachined Thermal Shear Stress Sensor,"" in *ASME Application of Microfabrication to Fluid Mechanics* (Chicago, 1994), pp. 9–15.
23. Y.-C. Tai, and R. Muller, "Lightly-Doped Polysilicon Bridges as Flow Meter,"" *Sensors and Actuators* **15** (1), 63–75 (1988).
24. F. Jiang, Y.-C. Tai, J.-B. Huang, and C. M. Ho, "Polysilicon Structures for Shear Stress Sensors," *Digest IEEE TENCON'95* (Hong Kong, November 1995), pp. 12–15.
25. A. Hussain and W. Reynolds, *The Mechanics of a Perturbation Wave in Turbulent Shear Flow*, Scientific Report 70-1655TR, AFOSR (1970).
26. T. Tsao, C. Liu, Y.-C. Tai, and C.-M. Ho, "Micromachined Magnetic Actuator for Active Fluid Control," in *ASME Application of Microfabrication to Fluid Mechanics* (Chicago, 1994), pp. 31–38.
27. R. Miller, Y.-C. Tai, G. Burr, D. Psaltis, C.-M. Ho, and R. Katti, "Electromagnetic MEMS Scanning Mirrors for Holographic Data Storage," *Solid State Sensor and Actuator Workshop 1996* (Hilton Head, SC, 1996), pp. 183–186.
28. F. Jiang, Y.-C. Tai, R. Karan, and M. Garstenauer, "Theoretical and Experimental Studies of the Micromachined Hot-Wire Anemometers," in *Tech. Digest 1994 IEDM* (San Francisco, 1994), pp. 139–142.
29. H. Tennekes and J. L. Lumley, *A First Course in Turbulence* (MIT Press, Cambridge, Mass., 1972)

30. T. Tsao, F. Jiang, R. Miller, Y.-C. Tai, B. Gupta, R. Goodman, S. Tung, and C.-M. Ho, "An Integrated MEMS System for Turbulent Boundary Layer Control," *International Conference on Solid-State Sensors and Actuators (Transducers '97)*, (Chicago, 1997). pp. 315–318.
31. M. Moller, "Efficient Training of Feed-Forward Neural Networks." Ph.D. thesis, Aarhus University, Denmark, 1993.
32. C. Lee, J. Kim, D. Babcock, and R. Goodman, "Application of Neural Networks to Turbulence Control for Drag Reduction," *Physics of Fluids* **9** (6), 1740–1747 (1997).
33. C. Mead, *Analog VLSI and Neural Systems* (Addison-Wesley, Reading, Massachusetts, 1989).
34. T. Delbrück, "Bump Circuits for Computing Similarity and Dissimilarity of Analog Voltages," *Computation and Neural Systems Dept. Memo 10*, California Institute of Technology, Pasadena (1991).

15

Analysis Tools and Architecture Issues for Distributed Satellite Systems

G. B. Shaw,* G. Yashko,† R. Schwarz‡, D. Wickert,** and D. Hastings††

15.1 Introduction

The recent development of several new technologies has made the concept of a distributed satellite system feasible. The term "distributed satellite system" refers to the coordinated operation of many satellites to perform some specific function. This definition encompasses a wide range of possible applications in commercial, civilian, and military sectors. The advantages offered by such systems can mean improvements in performance, cost, and survivability compared with the traditional single-satellite deployments. These improvements make the implementation of these systems attractive and inevitable. The emphasis of this chapter is to highlight the important concepts and issues associated with distributed satellite systems and the implications for small satellite and microsatellite designs.

15.1.1 Vision for the Future

The development of low-cost, single-function satellites offers new horizons for space applications when several satellites operate cooperatively. The vision of what can be achieved from space is no longer bound by what an individual satellite can accomplish. Rather, the functionality is spread over a number of cooperating satellites. This functionality greatly expands the utility of small satellites and microsatellites, allowing them to be used for a much wider range of missions. Further, these distributed satellite systems allow the possibility of selective upgrading as new capabilities become available in satellite technology. In the next 10–20 years, the commercial world will see the development of four types of space-based systems that will be available to both friendly and unfriendly nations, corporations, and individuals on a worldwide basis.

- **Global positioning and navigation services**. While the DOD already has the Global Positioning System (GPS), other countries are developing equivalent systems or augmenting the existing ones. Similar capabilities will be available through the development of personal communication systems. They will enable navigation with an accuracy of less than 1 m.
- **Global communication services.** Several systems are already in production, such as Iridium, Globalstar, and ICO. These systems will provide universal communications services between mobile individuals to almost anyplace on the surface of the Earth. These systems will work transparently with local cellular systems and will enable rapid telecommunications development in underdeveloped parts of the world.
- **Information transfer services.** These services will enable data transfer between any two points on the surface of the Earth at rates ranging from a few bits per second for paging, to mega and gigabits per second for multimedia applications. Proposed systems include Orbcomm, Spaceway, Cyberstar, Astrolink, and Teledesic. Individual users will be able to access

*Space Systems Laboratory, Massachusetts Institute of Technology, Cambridge, Massachusetts (MIT)
†Space Systems Laboratory, MIT
‡Space Systems, Loral Space and Communications
**U.S. Air Force
††Department of Aeronautics and Astronautics, MIT

large amounts of data on demand. DirectTV from broadcast satellites is a harbinger of what will be possible.

- **Global reconnaissance services.** These services will provide commercial users with multispectral data on almost any point on the surface of the Earth with meter-scale resolution. Data will span the range from the radio frequencies (RF) to the infrared (IR) through the visible into the ultraviolet (UV). The data will be available within hours of a viewing opportunity and approximately a day from the time of a request. Proposed systems include improvements to the French SPOT, as well as Orbimage, World View, Earthwatch, and various types of radar satellites.

Each of these four systems will be part of the global infosphere. Persons wishing to use this infosphere will be able to locate themselves on any point on the Earth, communicate both by voice and computer to other points on the Earth, and have a good picture of the local environment. Both the services and the technologies that enable them will be commercially available all over the world. Given the enormous magnitude of the commercial market, military and NASA communications will have to be fully integrated with, and technologically dependent on, the exploding market-driven communications technologies. There are opportunities for cooperation between the DOD, civil, and commercial ventures. The distributed sensors required for many defense applications are sufficiently small to be added to most or all new or replacement satellites of commercial constellations. This possibility opens avenues for multiplying defense capability and bandwidth at low incremental cost. Continued U.S. leadership in space systems will require understanding what can be achieved with distributed satellite systems and using this paradigm shift. The paradigm shift that is currently starting in the space arena is analogous to what has happened in the scientific computing market. This market started with large mainframe computers and then moved to powerful workstations because more functionality was provided for the money. The market is now moving to distributed sets of workstations to handle larger problems than cannot be attacked by single workstations. The same reasoning is resulting in the move to distributed satellite systems.

15.1.2 Level of Distribution

A distributed satellite system can have two different definitions:

- **A system of many satellites that are distributed in space to satisfy a global (nonlocal) demand.** Broad coverage requirements necessitate separation of resources. At any time, the system supports only singlefold coverage of a target region. The local demand of each region is served by the single satellite in view. Here, the term "distribution" refers to the fact that the system is made up of many satellites that work together to satisfy a global demand.
- **A system of satellites that gives multifold coverage of target regions.** The system therefore has more satellites than the minimum necessary to satisfy coverage requirements. A subset of satellites that are instantaneously in view of a common target can be grouped as a cluster. The satellites in the cluster operate together to satisfy the local demand. Note that the cluster may be formed by a group of formation-flying satellites or alternatively from a subset of satellites that are just passing close to each other. The cluster size and orientation may change in time, as a result of orbital dynamics. In any case, the number of satellites in the cluster is bounded by the level of multifold coverage. In this context, "distribution" refers to the fact that several satellites work together to satisfy a local demand. The entire system satisfies the global demand.

The most important characteristic of all distributed systems, common to both of the above concepts, is that more than one satellite is used to satisfy the global demand. This is the basic

distinction between a distributed system and a single-satellite deployment. Within the classification of a distributed system, the main difference between the two concepts described above lies in the way that the local demand is served. Specifically, the distinction is the number of satellites used to satisfy this local demand. The set of satellites that are used to serve the local demand is defined as a cluster. The cluster size, defined as the number of satellites making up the cluster, is therefore a valid measure of the level of distribution. The lowest level of distribution, with a cluster size of one, corresponds to the first meaning of distribution described above. Larger cluster sizes correspond to higher levels of distribution.

15.1.3 Applications and Generalization to Information Transfer Systems

There are numerous possible applications for distributed satellite systems within the military, civil, and commercial sectors. Some applications are distributed implementations of traditional single-satellite deployments, and some represent new capabilities that would be otherwise impossible to achieve. Although it is unlikely that distributed systems are suitable for all applications, there are many missions for which small sensors on many satellites scale very well and give cost-effective solutions.

All current and envisioned satellite applications involve providing some kind of service in communications, sensing, or navigation. The common thread linking these applications is that the satellite system must essentially perform the task of collection and dissemination of information. Data that contain pertinent information are gathered by the satellite, either from other components of the system (on the ground, in the air, or in space) or from the environment (local or remote). Some interpretation of the data may be performed, and the satellite then disseminates the information to other system components. The generalization made is that all satellite systems are basically information transfer systems, and that ensuring information flow through the system is the overall mission objective. This objective is easily understood for communication and remote-sensing systems. Perhaps more surprising is that navigation systems such as GPS are also information disseminators. The GPS control segment uploads the satellites with information. The satellites use this information to construct a signal that is retransmitted to the ground. Users of GPS can use the information in the received signal–including not only the navigation message contained therein but also the phase of the signal itself–to determine a navigation solution. As with communications and remote sensing, the performance of the system relies on the flow of information through the satellite network.

The format and routing of the information being transferred may be different for different applications, but it is always subject to the same rules of information theory. This common thread linking all systems (navigation, surveillance, communications, and imaging) establishes a context for a generalized analysis, and is particularly useful in the study of distributed systems. In distributing the functionality of a system among separate satellites, the system is essentially being transformed into a *modular* information-processing network. The satellites make up individual modules of the system, each with well-defined interfaces (inputs and outputs) and a finite set of actions. Such systems are analogous to the distributed inter- and intranet computing networks, and as such, are subject to similar mathematics. Distributed computing is a rapidly developing field, and a great deal of work has been done to formalize the analyses.[1,2] Much of this groundwork can be adopted for distributed satellite systems, leading to significant insight about the likely performance and problems associated with these systems.

The remainder of this chapter addresses the fundamental issues related to the use of distributed architectures for space applications. The reasons supporting their implementation are first introduced and explained. This explanation is followed by a discussion of the factors that must be

considered in the design of distributed architectures. Finally, an analysis methodology is presented for quantifying the cost, capability, and adaptability of distributed satellite systems compared to single-satellite deployments. This last section represents work in progress, and is part of a large study of distributed satellite systems currently being undertaken at MIT.[3] Throughout the chapter, simple examples are used to demonstrate important concepts. At the end of Sec. 15.4, an example is constructed to show how to apply the new analysis methodology, unifying many of the issues discussed (see Example 10).

15.2 To Distribute or not to Distribute?

There are many reasons why a distributed architecture is well suited to some space applications. Unfortunately, the arguments for or against distribution are fraught with subjectivity and firmly entrenched opinions. It currently seems that most of the satellite design houses in the country are internally split between the proponents and opponents of distribution. Each camp supports one side of the debate vehemently and can find a seemingly endless stream of supporting arguments to back their claims. The "radicals" claim that the development of large constellations of small satellites leads to economies of scale in manufacture and launch, reducing the initial operating costs. They also maintain that the system becomes inherently more survivable due to the built-in redundancy. Conversely, the "traditionalists" debunk these arguments, reminding everyone that you can't escape the need for power and aperture on orbit, and that building even 100 satellites does not imply significant bulk-manufacturing savings. They assert that the lifetime operating costs for large constellations will far outweigh the savings incurred during construction and launch.

In fact, most of the statements made by both sides are true, but only when taken in context. Clearly, a distributed architecture is not the panacea for all space applications. It is tempting to get carried away with the wave of support that the proponents of distributed systems currently enjoy. Care must be taken to temper this enthusiasm. Also best avoided is the naive, but commonplace, application of largely irrelevant metaphors supporting the adoption of distributed systems; the unerring truth that ants achieve remarkable success as a collective is really not an issue in satellite system engineering!

This section is intended to summarize the real reasons supporting the use of distributed satellite systems, and should hint toward the type of applications for which they are best suited. The short list given here is probably not complete; there are likely many other reasons that support or oppose the use of a distributed architecture for some particular application. Rather, this section is meant to highlight the most important and fundamental factors that are both relevant to this debate and play a common role in system architecture studies.

In order for a distributed architecture to make sense, it must offer either reduced cost or improved performance compared with traditional single deployments. This rationale is of course conditional upon the perception of how performance is measured. Before proceeding to the reasons supporting distribution, some discussion of measurable performance is therefore necessary.

As described in the opening section of this chapter, all envisioned applications for small satellite systems involve some kind of information collection and dissemination. This process always requires the detection of information-bearing signals in the presence of noise and interference. In most cases, this information is in the form of a digital data stream.* The instantaneous performance of digital data transfer systems can be characterized by four important parameters

* Even those systems featuring analog detection, such as optical imaging, almost always feature analog-digital conversion before transmission to the end user.

relating to the detection process: signal isolation, information rate, information integrity, and information availability.

- **Signal isolation** measures the system's ability in isolating (in signal space) different information signals that originate at different sources. Multiple access schemes for communication systems are methods of signal isolation. Analogously, the resolution of an imaging system allows isolation of the signals from adjacent pixels. The system must be able to sufficiently isolate and identify each of the signals within the demand space.
- **Information rate** is a measure of the rate at which information symbols are transferred through the system. This parameter is most familiarly associated with the data rate for communication systems. The pixel sampling rate is the corresponding parameter for imaging systems. The system must sample information symbols at a rate that matches the characteristic bandwidth of the source. For instance, a high-speed cruise missile must be tracked with a high sampling rate.
- **Information integrity** measures the error performance of the system. The integrity is most commonly represented by the probability of making an error in the interpretation of a signal based on noisy observations. For communications, the integrity is measured by the bit error rate; for a search radar system, the integrity is measured by the false alarm rate.
- **Information availability** measures the instantaneous probability that information is being transferred between the correct origin-destination pairs at the correct rate and with the desired integrity. Note that this is a functional definition; the availability is the probability that the system can perform specific functions. In this way, the availability is more than a statement about component reliabilities. A failure-free system will have a low availability if it cannot satisfy the requirements on isolation, rate, and integrity. A loss of availability can arise from component failures, from signal attenuation due to rain or clouds, or simply from random fluctuations in performance. Availability is also related to coverage. A system that cannot support continuous coverage of a region will have a low availability for real-time applications.

The four parameters of signal isolation, information rate, information integrity, and information availability characterize the performance of a satellite system. A system architecture that offers improvements in any of these parameters should be given serious consideration during the system design.

The system cost is the total resource expenditure required to build, launch, and operate the satellite system over the expected lifetime of the system. This cost includes the baseline cost of developing, constructing, launching, and operating the components of the system, and also the expected costs of failure. These expected costs arise from the finite probability of failures occurring that could compromise the mission. Should such failures occur, economic resources must be expended to compensate for the failures.

All of the reasons supporting the use of distribution therefore relate in some way to improving the performance characteristics or to reducing the baseline or failure compensation costs. The following subsections detail these reasons. A later section takes this process one step further by introducing quantitative metrics based on measurable performance and cost, allowing comparative analysis between many different system architectures. Only in this way, by quantitative analysis, can the question "to distribute or not to distribute" be fairly answered.

15.2.1 Signal Isolation Improvements

A system's ability to isolate and identify signals from different sources within the field of view (FOV) is a critical mission driver for many applications. Obviously, a system cannot satisfactorily transfer information between specific origin-destination pairs unless the individual sources and

sinks can be identified. The various methods used to isolate the different signals are termed "multiple access schemes." For communication systems, common multiple access schemes separate the signals in frequency (frequency division multiple access [FDMA]), time (time division multiple access [TDMA]), or signal space (coded division multiple access [CDMA]). Also, individual spot beams can be used to access multiple sources that are spatially separated.

The same techniques can be applied to radar systems. Doppler frequency shifts are used for identification of the target velocity and clutter rejection, and time gating is used for target ranging. Scanning a small radar beam over a large area allows separate targets to be isolated in space to within a beamwidth.

For imaging and remote-sensing systems, the same principles apply. Different sources can be identified if they are detected in different frequency bands, or if they are sampled at times that match the source characteristics. Spatially separated sources can be isolated using a high-resolution detector. An aperture can distinguish between sources that are separated by a distance no less than the resolution of the aperture. Note the one-to-one correspondence between:

- The resolution of an optic and the beamwidth of an antenna or a radar
- The frequency of radiation from a remote-sensing pixel, the carrier frequency of a communication signal, and the Doppler shifts of a radar signal

Distribution can be beneficial for signal isolation for the following reason. By separating resources spatially over a large area, the geometry of the signal collection is different for each detector. This geometry can assist in the separation of the different signals due to FOV changes, different times of flight, or different frequency or phase of the received signals. Larger spatial separation of the apertures means that the phase difference between signals arriving at different detectors is increased, further separating the signals in signal-space. This reasoning is demonstrated in Ex. 1.

Example 1. Isolation Improvements: Spacecraft Arrays

The advent of economical, fast integrated-circuit technology has recently surpassed the previously restrictive data-processing requirements of forming large sparse, synthetic apertures in space. Many people have now started to claim that use of this technology offers potential benefits by reducing the mass and cost of remote-sensing systems for high-resolution imaging.

The angular resolution of any aperture scales with the overall aperture dimension, expressed in wavelengths:

$$\theta_r = \frac{\lambda}{D}$$

where D is the size of the aperture.

An array is an aperture excited only at discrete points or localized areas. The array consists of small radiators or collectors called elements. An array with regularly spaced elements is called a periodic array. To avoid grating lobes in the far-field radiation pattern, the elemental spacing of a periodic array should be less than one-half of the wavelength. An interferometer is the most basic form of a thinned array. In an interferometer, only two elements are used, with measurements taken at differing separations to fill out the sparse aperture sampling.

A random array is a thinned array with random positions of the array elements. The spacing of the elements is usually much larger than one-half of the wavelength, leading

to fewer elements for a given overall aperture dimension. Grating lobes are avoided because there are no periodicities in the elemental locations.

The concept of the spacecraft array involves forming a large, thinned aperture from a set of satellites, with each satellite acting as a single radiator element. Since the spacing between satellites is very much greater than characteristic wavelengths, grating lobes can be avoided only by positioning the satellites randomly.

The signal-to-noise ratio (SNR) behavior of sparse arrays is identical to a filled aperture of the same physical area. That is, a sparse array of N elements, each of area A, will achieve the same SNR as a filled aperture of area NA. This of course makes sense, since the same amount of energy is collected over the same physical aperture in both cases. However, the resolution of sparse arrays can be very much larger than that of an equivalent filled aperture. This resolution arises from the enlarged overall aperture dimension that results from splitting and separating the aperture into elements.

Consider an imaging system capable of 1-m resolution at a wavelength of 0.5 μm (green visible). A geostationary satellite would require diffraction-limited optics 18 m across. Similar resolution for lower frequencies (e.g., X-band) requires even greater aperture sizes. This requirement is clearly impractical for filled aperture systems. A filled aperture must be supported over its entire extent, leading to heavy structures. Even if mass can be kept low through the use of advanced materials, impressive deployment techniques would be required to stow such an antenna within the launch shroud. The question arises as to how big a filled aperture can be built and launched.

Clearly, a sparse aperture can be made very large indeed. The only requirement is that the current paths connecting elements be of the same optical length. Widely separated elements connected through light "tethers" or "booms" could easily extend over length scales of 10 to 100 m. For even larger aperture sizes, a sparse array of separated spacecraft allows resolutions in the submilliarcsecond range.

15.2.2 Information Rate Improvements

For many applications, the requirements for a high information rate drives the designer toward very large apertures and high-power payloads. The probability of correctly detecting information symbols in the presence of noise is a function of the energy in each information symbol. Collecting or transmitting symbols at a high rate therefore requires high-power signals. This requirement, in turn, leads to high-power transmitters or large apertures (to collect more power or to concentrate the power radiated). As an example, consider the HS-601 GeoMobile satellites built by Hughes Space and Communications to provide mobile telecommunications services. To communicate with the handheld subscriber units, these satellites need an antenna 12.25 m in diameter, and payload power of 7 kW.[4] Although such enormous payloads can be accommodated on large satellites like the HS-601, they are clearly infeasible for small-satellite designs. Thus, the range of applications for which small satellites can be used is limited in singular deployments.

There are some applications for which even the largest, highest-power satellite buses available today are too small. For instance, the U.S. Air Force recently laid out performance requirements for a space-based radar system to replace the aging Airborne Warning and Control System (AWACS) aircraft. The Space and Missile Center (SMC) undertook a "quick-look" study in response to these requirements, and was forced toward designs featuring truly massive satellites, beyond the capabilities of today's technology (30 kW of RF power and 12,300 m^2 of aperture).[5]

The solution to these requirements for high information rates and massive satellites lies in multiplying the capabilities of several smaller satellites, such that their combined operation satisfies

the overall mission requirements. This solution can be achieved by division of the top-level task into smaller, more manageable tasks that can be allocated among the elemental components of the distributed architecture. Example 2 illustrates this solution.

Example 2. Rate Improvements: A Distributed Space-Based Radar

The driving mission requirement for a space-based search radar is to detect targets quickly enough to allow a defensive response to be taken. A mean time to detection of 10 s is considered reasonable.[5] Using conventional monolithic designs, this requirement leads to very large power-aperture products. The resulting singular-deployment satellite system would be prohibitively expensive (and probably infeasible) to build and launch.

A distributed version of the space-based radar satisfies the mean-time-to-detection requirement using a group of small satellites working in collaboration. Several satellites search the same area independently, each with a mean detection time that is actually longer than 10 s. The detection rate from a single satellite is therefore insufficient to satisfy the overall detection rate requirement. However, the cumulative detection rate of the whole group is the sum of the rates from each satellite:

$$\text{System Detection Rate, } R_s = \sum_{sats} R_{sat}$$

where R_{sat} is the detection rate for each satellite. The overall mean time to detection is then the reciprocal of the system detection rate R_s. The combined performance from several lower-performance components can therefore satisfy the desired mean-time-to-detection requirement for the system. Through subdivision of the task, distribution reduces the requirements of each individual component.

15.2.3 Information Integrity Improvements

The error performance of data collection and transfer systems is a critical issue in their design and operation.[6] Generally, the probability of erroneously interpreting an information signal depends on the decision rule used to distinguish between data symbols. The detector uses an observation Y_k of the signal plus noise to make a decision about each information bit X_k. Since each X_k has a finite number of possible values, the detector must make a decision from a finite number of alternatives. An error can occur if noise or interference degrades the information signal in such a way that an incorrect decision is made about the observation. These errors can be as benign as a single bit error in a communication message, or as consequential as a false alarm for an early warning radar system.

The probability of error for a single measurement is the likelihood that the interfering noise power exceeds some threshold, equal to the difference between information data values. Consider, for example, the simplest case of an amplitude-modulated binary communication channel (binary power amplitude modulation [PAM]). The two data values {0,1} are represented by two different power levels of the passband carrier wave. The separation between these power levels is d watts. If the noise component of the signal has a power level greater than d watts, a data symbol {0} can appear in the observation as a {1}, or vice versa. The probability of an error of a single bit is then the probability that the noise power is greater than the separation between data symbols. For additive Gaussian noise, this probability reduces to an exponential function of the SNR. The interfering noise can arise from several sources:

- Thermal noise from resistive heating of electrical components in the receiver
- Noisy radiation sources in the FOV of the instrument

- Jamming from unfriendly systems
- Interaction with the transmission medium (rain, bulk scatterers)
- Background clutter

The random statistics of these noise sources is the basic reason supporting the use of distribution to improve the error performance. If several instruments are used to measure the same signal, the total signal power collected increases linearly with the number of detectors. However, the noise sources, being characteristically represented as independent random distributions, are incoherent between detectors. The compound effect of the noise collected or introduced at each detector is therefore partially canceled. Essentially, this partial cancellation is a direct result of the assumption of independence between the different samples of the noise signals. This partial cancellation is especially true of errors arising from instrumental sources of interference. The thermal receiver noise is obviously incoherent between different detectors. The interference from noisy radiating bodies in the FOV of one satellite may not be an issue for a second satellite, due to the differing viewing angle of the scene. Jamming interference is also satellite specific; an enemy can easily disrupt a single satellite but would struggle to jam an entire group of satellites that may be spatially separated.

Therefore, errors made in the interpretation of signals are likely isolated to specific detectors or satellites. In this way, the overall integrity of the compounded information from several instruments is improved. This improvement is simply a consequence of eliminating instrumental errors through averaging. Multiple measurements also allow voting algorithms to be implemented, further reducing the probability of erroneous interpretation.

15.2.4 Information Availability Improvements

If carefully designed, a distributed architecture can often lead to improved performance of the satellite system by increasing the availability of system operations. A system is termed unavailable if it cannot collect and disseminate information between known and identified source/sink pairs at the required rate and integrity. Losses of availability can result from component failures, from signal attenuation due to blockage or weather, or from random performance fluctuations. Since most systems have to transfer information between locations distributed throughout the entire coverage area, a loss of availability can also be attributed to poor coverage statistics, and can be limited to isolated locations. For example, defense reconnaissance satellites may have to image scenes over two or more continents, relaying the data to multiple downlink stations across the world. There will be times during the orbits of these satellites when they are not passing over important targets. The system is unavailable at these times since images of the targets cannot be recorded. The revisit time of the satellites effectively specifies the availability built into the system. Of course, very high availabilities can only be achieved by constellations giving continuous coverage over the target regions. The availability of a system is related to the variance of the supportable rate and integrity, and as such, is sensitive to worst-case scenarios. Since a loss of availability represents an inoperational state, any measures that can be taken to improve the availability of a system are desirable. There are several methods by which distributed architectures can lead to increased availability through reductions in the variance of performance. These reductions can lead to the following improvements:

- Improve coverage of the demand
- Reduce impact of component failures

The methods by which distribution can lead to these improvements are described in detail in the following sections.

15.2.4.1 Matching a Distributed Demand

Some applications require the reception of signals at many different locations. Such applications are characterized as having a distributed demand. A worldwide communications consumer base, or sampling locations for a global mapping of the geomagnetic field, are examples of a distributed demand. The architectural options for these applications are to place sensors everywhere there is a demand, to have a single sensor that maneuvers to the demand locations, or to adopt some strategy somewhere in between these extremes. The trade-off here is between the cost of additional hardware resources and the cost of additional expendables, such as fuel and time. A system with a few satellites that can maneuver to different sampling locations (either by thrusting or by utilizing orbital mechanics) requires less dry mass on orbit, possibly leading to lower costs. However, the additional cost of fuel, or the opportunity cost associated with the loss of availability due to sequential sampling, may sway the balance in the other direction. A question presents itself: how should spacecraft resources be distributed to best match a distributed demand? The answer to this question is, unfortunately, neither simple nor general. The best option for one application may be unsuitable for another. There are, however, some general trends.

Clearly, a distributed architecture is the only option for applications requiring simultaneous sampling at all demand locations. This is equivalent to a continuous coverage requirement. Consider, for example, a global mobile communication system. A single satellite cannot serve the entire globe, forcing the designer toward a constellation of satellites that can guarantee continuous coverage.

Some applications involve a coupling between measurements at different sample locations, especially during processing of the information. An example of this coupling is the combining of signals collected by the apertures of an interferometer—an essential operation in the construction of an image. This sharing of information necessitates interfaces between the satellites of the system. These interfaces are expensive and add complexity. Furthermore, the transmission of information between satellites requires energy expenditure. Electrical energy, like propellant, is a valuable expendable resource. In some cases, the energy cost of data transmission between satellites can be more expensive than the equivalent fuel expense of maneuvering. This is especially true for satellite systems relying on nonrenewable energy sources (e.g., batteries, fuel cells). For these applications, if sequential sampling can be tolerated, the savings in hardware from having fewer satellites can offset the opportunity costs associated with losses of availability.

Some tasks involve no coupling between the different sampling locations. In these tasks, the processing of signals at the different locations can be performed independently. Separate, independent sensors can satisfy the demand without the sensors having to interface among themselves. Without any of the energy costs or complexity of intersatellite links, a very distributed architecture may be favorable for these applications—the improvements in availability outweighing costs of extra on-orbit hardware.

Example 3. Matching a Distributed Demand: The Separated Spacecraft Interferometer

Optical interferometers collect light at widely separated apertures and direct this light to a central combining location, where the two light beams are interfered. Fringes produced by the interference provide magnitude and phase information from which a synthesized image can be generated. Space-based optical interferometers can be implemented as single spacecraft, featuring collecting apertures separated by tens of meters, or as separated spacecraft, where baselines of hundreds or thousands of meters enable measurement with submilliarcsecond angular resolution.

The collector spacecraft sample the distant starlight at several different baselines (separation and orientation) in order to construct the image. The locations of the sampling points define a distributed demand. Clearly then, a possible modification to the basic configuration that could offer improved availability is a system with an increased number of collectors. By distributing the collectors at the desired sampling locations, many different baselines can be made from the numerous combinations of collector pairs. In this way, many baselines can be measured simultaneously (or at least without additional maneuvers), and the image can be filled out more quickly.

There is a distribution of sampling locations that has been shown to fill out the image optimally for a given number of baselines. This distribution is known as the Cornwell distribution, and specifies the locations of points to place collectors such that measurements at all baselines result in optimal sampling of the image. Obviously, sampling from a distribution with a larger number of Cornwell points results in a higher quality image. Unfortunately, sampling at more locations requires either more collectors or more maneuvers. A system with more collectors requires fewer maneuvers to sample at all pairs of Cornwell points. In order to sample all pairs of points from an m point Cornwell distribution, a system consisting of n collectors must make $({}^{m}C_{n} - 1)$ separate maneuvers. In choosing the system size, a trade-off is therefore made between the cost of additional collectors and the cost of propellant for maneuvers. This trade-off can be demonstrated with a simple calculation.

Consider a system designed to meet similar objectives to that of the New Millennium Interferometer, proposed by NASA.[7] The requirement is to image 500 objects, with five revisits, over a 15-year lifetime. This requirement translates into approximately one image every 2 days. Provided each image is completed in this 2-day period, the availability of a candidate system is considered satisfactory. The total system cost corresponding to configurations with different numbers of collector spacecraft can be estimated for various image quality requirements. To simplify the calculation, the system cost can be represented as total system mass; this is a reasonable approximation to first order, and allows the important trends to be seen. The total mass is the sum of the mass of all the satellites M_s and the total propellant mass M_f:

$$\text{Total mass} = M_s + M_f$$

The cumulative satellite dry mass M_s is simply the number of collectors n_s multiplied by the dry mass of each collector, assumed to be 200 kg. The total fuel mass depends on the number of maneuvers that must be made. To form an m-point image, the total number of maneuvers N_m for a configuration of n_s collectors is given by:

$$N = {}^{m}C_{n_s} - 1 = \frac{m!}{n_s!(m-n_s)!} - 1$$

The total fuel mass is then simply:

$$M_f = N_m M_f NI$$

where m_f is the propellant mass for a single maneuver and NI is number of images. This value can be simply calculated. Assume a hydrazine propulsion system on each collector, with an $I_{sp} = 210$ s and a thrust $F = 25$ N. Model each maneuver as a constant acceleration of fixed duration t_1, followed by a drift period $t_{2,i}$ (specific to each maneuver), and then a constant deceleration of the same fixed duration t_1. For any particular maneuver, the total distance to be moved s_i is shown in Eq. (15.1):

$$s_i = \frac{1}{2}at_1^2 + at_1t_{2,i} + \frac{1}{2}at_1^2 \tag{15.1}$$

where $a = F/m_s$ is the acceleration.* Recall that the total time T to perform all maneuvers required per image is bounded to 2 days:

$$T = N2t_i + \sum_i^N t_{2,i}$$

Rearranging and substituting for $t_{2,i}$ from Eq. (15.1) then allows us to solve for t_1 in terms of the known quantities. The propellant mass used for a maneuver is then:

$$m_f = 2t_1\dot{m}_f ,$$

where the propellant mass flow rate $\dot{m} = F/(gI_{sp})$.

These calculations were evaluated for several different system sizes and Cornwell distributions. The results are shown in Fig. 15.1. Notice that the optimum number of collector satellites varies depending on the size of the Cornwell distribution. For low-quality images with a low number of Cornwell points, it is more efficient to have small numbers of collectors that maneuver frequently. Conversely, for large sampling densities, it becomes more efficient to increase the number of satellites. Notice, however, that there exists an optimum number of satellites, beyond which the increases in dry mass outweigh the benefits from reductions in maneuvering. This is a trend seen in a wide range of applications involving a distributed demand.

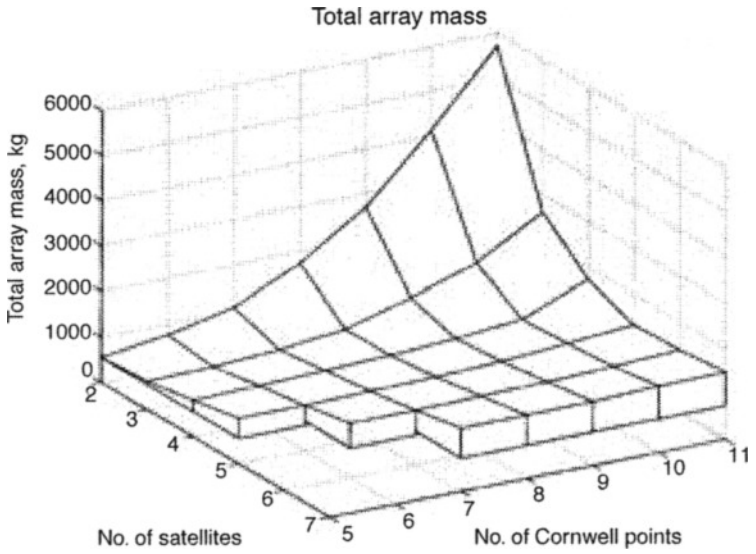


Fig. 15.1. Separated spacecraft interferometer mass vs configuration size and image quality. (S/C is satellites.)

*s_i is known from the Cornwell distribution.

15.2.4.2 Redundancy and Path Diversity

A loss of availability due to component failures, blockage, or rain/cloud cover can be avoided if there are redundant information paths. The redundancy can reduce the impact of component failures. This redundancy can be provided by distributed architectures featuring multifold coverage. For example, a mobile communication user can select, from all of those satellites in view, the operational satellite with the clearest line of sight. This redundancy can reduce service outages and improve availability. This concept extends across almost all applications.

Note, however, that distribution can improve availability only if there is redundancy in the design. A distributed architecture with total resources that can only just satisfy the demand is a serial system and is subject to serial failure modes. A failure in any component will lead to a failure of the system. The system availability is the product of the availabilities of the components. Since the availability of a component is always less than unity, the overall availability of the system decreases geometrically with the number of serial components. Only by adding redundancy can a distributed architecture take advantage of parallel reliability. System failure of a redundant architecture requires all parallel paths to fail.

In general, most architectures will require some redundancy to satisfy the availability requirements throughout the expected lifetime. Frequently, the cost associated with this redundancy is less for a distributed architecture than it is for traditional systems. This availability (or reliability) cost accounts for the production, storage, and launch of on-orbit or on-the-ground spares necessary to maintain availability. For a distributed system, these spares often represent only small fractions of the initial deployment.

For example, consider two alternate designs for some arbitrary mission. The initial deployment of system A features a single satellite that can meet the isolation, rate, and integrity requirements with an availability of 80%. System B is a distributed architecture. Without any redundancy, system B can satisfy the isolation, rate, and integrity requirements with three satellites, each a third as capable as the satellite of system A and with the same satellite availability of 80%. The initial availability of system B is therefore poor at $0.8^3 = 51\%$. If we then raise the requirement on availability to be 90%, both systems require redundancy. System A requires a single spare. The system can tolerate a failure of a single satellite, so the combined availability is simply 1 minus the probability of both satellites failing: $1 - 0.8^2 = 96\%$. System B requires two additional spares to give a configuration of five satellites. The system can tolerate failures in any two satellites, to give an availability of 94% (the probability of two or fewer satellites failing). Both systems satisfy requirements, with similar eventual availabilities. However, the additional availability expenditures are different. Whereas system A requires additional expenditure equal to the initial deployment cost, system B needs availability expenditure of only two-thirds of the initial costs.

15.2.4.3 Visibility and Coverage Geometry

In some instances, distribution and multifold coverage improve the availability by reducing the variability of system performance. By making the behavior of the system more predictable, the probability of operating within acceptable performance bounds is increased. The performance of the system is particularly sensitive to coverage variations; it is here that distribution can lead to improvements. The multifold coverage characteristic of distributed architectures supports consistent levels of performance in two ways:

- **Reducing the variance of the visibility, defined as the number of satellites in view from a ground station.** Generally, the visibility is a function of both space and time. The number of satellites in view from a location changes in time and is usually different at other locations.

The performance of a satellite system is frequently dependent on the visibility. Large variations in visibility can therefore cause large fluctuations in performance. The designer faces the choice of sizing the system for the worst-case coverage, or accepting losses of availability at times when the visibility is below average. Increasing the number of satellites in the constellation not only increases the visibility, but also reduces the variance. According to the Central Limit Theorem, as the numbers of satellites is increased, the minimum visibility converges toward the average value. This theorem assists the designer, improving the availability of systems based on average coverage characteristics.

- **Reducing the impact of performance variability (of individual satellites) by taking advantage of favorable coverage geometry.** The geometry of the coverage over target regions can have a large impact on the sensitivity of the system. Frequently, the rate or resolution that can be supported by a single satellite can be spatially and time varying, depending on the viewing angle, the transmission path, and the detector characteristics. Favorable coverage geometries minimize the impact of these variations, ensuring that the combined operation of the distributed configuration achieves consistent levels of performance.

These two concepts are easily understood with the help of Example 4.

Example 4. Visibility and Geometry: Distributed Space-Based Radar

We return to the distributed space-based radar system that was introduced in Example 2. Recall that this system achieved the requirement on detection rate by summing the capabilities of several small radars that independently search the same target area. The cumulative detection rate is therefore directly proportional to the number of satellites in view of the target area. Variations in the visibility translate directly into variations in the achievable detection rate. This can result in a loss of availability if the visibility drops below that necessary to support the required detection rate. The availability can be improved if the system is designed to use an even greater number of smaller satellites to satisfy the detection requirement. As the number of satellites increases, the spatial and temporal variations in the visibility are reduced. The minimum visibility approaches the average value, and the achievable detection rate changes over a much smaller range.

Furthermore, larger configurations of satellites result in more favorable coverage geometries. The multifold coverage leads to a wide distribution of viewing angles surrounding the target. This is particularly important for slow-moving targets. The radar return from slow-moving targets is difficult to distinguish from the ground clutter. Normally, the different velocities of the target and the ground relative to the radar results in different Doppler shifts that separate the target and clutter in frequency, allowing detection. The return from slow-moving targets is often buried in the clutter because of the low relative velocities. A viewing angle parallel to the target's velocity maximizes the Doppler shift between the target and the ground in the frequency spectrum, increasing the signal isolation and improving the probability of detection. Since the target's velocity vector is unknown *a priori*, receivers must be placed at all possible viewing angles to ensure detection. With receivers located at all angles around the target, the distributed space-based radar concept increases the probability of detecting slow-moving targets. This makes the system less sensitive to the target velocity, effectively increasing the availability by reducing the probability of failing the detection rate requirement.

15.2.4.4 Availability Improvements—A Case Study

The availability considerations were so instrumental in the design of the current GPS configuration that this design is worthy of special attention.

Example 5. The Navstar GPS

The most important performance objectives that impacted the design of the GPS system were as follows:[8]

- High-accuracy, real-time position, velocity, and time for military users on a variety of platforms, some of which have high dynamics, e.g., high-performance aircraft. "High accuracy" implies 10-m three-dimensional (3D) root-mean-square (rms) position accuracy or better. The velocity accuracy requires errors to be less than 0.1 m/s.
- Good accuracy to civilian users. The objective for civil user position accuracy is 100 m or better in three dimensions.
- Worldwide, all-weather operation, 24 h a day.
- Resistance to intentional (jamming) or unintentional interference for all users, with enhanced jamming resistance for military users.
- Affordable, reliable user equipment. This eliminates the possibility of requiring high-accuracy clocks or directional antennas on user equipment.

These objectives effectively shaped the architecture of the GPS system. The space segment consists of 24 satellites in 12-h orbits. Each of the satellites transmits a ranging signal consisting of a low-rate navigation message, spread over a large bandwidth by a high-rate pseudorandom noise (PN) code. The resulting signal is used to modulate a carrier at two frequencies in the L-band. The PN codes are chosen such that the signals from different satellites are orthogonal, providing a multiple-access technique. Two different codes are generated: the coarse acquisition (C/A) code has a short period of 1 ms, while the precision (P) code has a long period of 37 weeks. A ground receiver simultaneously tracks several satellites. The pseudorange from the user location to a satellite is measured by cross-correlating the received signal with a time-shifted replicated version of the code, therefore estimating the transmission delay. The actual observable is a pseudorange because it includes the user clock bias, the ionospheric, and tropospheric delays, plus relativistic effects and other measurement errors. The ionospheric group delay is corrected using dual frequency observations. By making at least four pseudorange measurements to different satellites, the unknown user position and clock bias can be estimated.

The triangulation algorithm used for positioning in GPS has many useful features. For example, errors in any pseudorange measurement to a satellite have a reasonably small impact on the navigation accuracy. This robustness arises from the inherent geometry: all the pseudoranges are measured in different directions. Obviously, some coverage geometries are more favorable than others in reducing the impact of errors. The quality of the coverage geometry is measured by the geometric dilution of precision (GDOP), which directly relates the rms position error to the rms ranging error. Small values of GDOP correspond to geometries supporting a high rms position accuracy.

The triangulation method of positioning, either by code ranging or phase measurements, was not the only option available to the GPS engineers. The Navy's Transit system, a predecessor to GPS, differed greatly in its system architecture. Transit relied on

users measuring the Doppler shift of a continuous 400-MHz tone broadcast from satellites orbiting in 600-nm polar orbits. The maximum rate of change of Doppler shift in the received signal corresponded to the point of closest approach of the Transit satellite. The range to the satellite at this point was deduced from the difference between the "up" Doppler and the "down" Doppler exhibited by the signal. In this way, a user who knew the altitude and the broadcast ephemeris of the Transit satellites could calculate a navigation solution to within a few hundred meters.

The most notable difference between Doppler positioning used for Transit and that used for GPS is that in the former system, a navigation solution was obtained using signals from only one satellite. Recall that the GPS navigation solution requires the reception of signals from at least four satellites. It would be tempting, therefore, to think that Doppler positioning offers some availability benefits over the triangulation adopted by GPS, since the requirement of having at least one satellite in view is more easily satisfied. This assumption is, however, entirely incorrect. Furthermore, distinct disadvantages are inherent in Doppler positioning that made it unsuitable for GPS. In fact, Doppler positioning has characteristically poor availability. For Transit, mutual interference problems limited the number of satellites to five, resulting in very intermittent coverage. Although multiple-access methods could allow a larger constellation, alleviating this problem, other factors contribute to the poor availability. The navigation algorithm requires the user to have knowledge of altitude and velocity (in order to cancel self-induced Doppler shifts). This requirement alone makes Doppler positioning unsuitable for aircraft and high-dynamic platforms. The availability of satisfactory navigation is, therefore, extremely sensitive to the user platform characteristics. In addition, the navigation accuracy is very sensitive to geometry for Doppler positioning. The GDOP is characteristically poor, and any errors in the range measurement propagate significantly into the navigation solution. This problem is particularly severe at higher elevations. Clearly, Doppler positioning could not satisfy objectives of the GPS system. The triangulation algorithm was therefore adopted for positioning in GPS.

A geostationary Earth orbit (GEO) was not chosen for the GPS constellation because of the requirement for coverage of high latitudes. Furthermore, the GDOP for a geostationary constellation is poor because a user on the Earth can only see satellites to either the north or the south, depending on the hemisphere in which they are located. A GEO orbit would therefore be compromised by poor availability.

The designers of GPS were, therefore, forced toward a lower altitude constellation of satellites. In order to guarantee low values of GDOP, the visibility of the GPS constellation had to be high. This is generally true of the chosen 24-satellite constellation. At some latitudes, it is possible for as many as 11 satellites to be in view simultaneously. Nevertheless, spatial and temporal variations in the GDOP remain one of the largest problems with the current GPS system. At midlatitudes, the visibility of the GPS-24 constellation is poor, with only four satellites in view for approximately 0.4% of the time. Alternate constellations could have offered improved visibility statistics. However, GPS-24 minimized the impact on availability of a failure in a single satellite.

The satellites of a lower altitude constellation have a relative motion over the Earth, causing Doppler shifts in the signals received. These Doppler shifts have the advantage of rotating the phase of the received signal, removing any direct current (dc) biases from tracking channels. Unfortunately, these same Doppler shifts cause interference problems from cross-correlation sidelobes. This interference is only a problem for the

decorrelation of the short-period coarse/acquisition (C/A) code, but is a limiting factor against significantly increasing the number of satellites in the constellation. Mutual interference from other GPS satellites could seriously degrade the ability of a receiver to decorrelate the C/A code. This interference could lead to large errors in the extraction of timing and navigation information from the C/A code. Since the C/A code is used to acquire the P-code (which is largely unaffected by Doppler due to its continuous spectrum), a receiver that cannot track the C/A code would not be able to produce a navigation solution at all.

15.2.5 Reducing the Baseline Cost

When a given satellite constellation is initially deployed, there are costs associated with development, production, and launch of the system's original complement of satellites. Additional expenditures beyond the initial deployment costs, termed "availability costs," are necessary to maintain the constellation over a given time period. Availability costs include the production, storage, and launch costs associated with the on-orbit or on-ground spares needed to ensure the availability of the system. The sum of the initial deployment costs and the availability costs make up the baseline system cost. This cost is discussed in the introduction to Sec. 15.2. Baseline costs are typically very high. For distributed satellite systems to be considered viable, they must be at least competitive in cost, as compared with traditional systems.

Conventionally, system cost estimates can be made using basic parametric models such as the USAF Unmanned Spacecraft Cost Model (USCM)[9,10] or the Small Satellite Cost Model.[11] These models consist of a set of cost-estimating relationships (CERs) for each subsystem. The total cost of the system is the sum of the subsystem costs. The CERs allow cost to be estimated as a function of the important characteristics, such as power and aperture. Frequently, the CERs are expressed as a power law, regressed from historical data. For example, the USCM estimate for the theoretical first unit (TFU) cost of an IR-imaging payload is based on aperture and is shown in Fig. 15.2.

Care must be taken in applying the SSCM to distributed systems. Although each satellite in a distributed system may be small, the SSCM was derived assuming single-string designs and modest program budgets. These assumptions clearly don't apply to a distributed system of perhaps 1000 satellites, with a total system cost of several billion dollars. Unfortunately, the use of USCM generally leads to high costs for distributed systems, for two reasons:

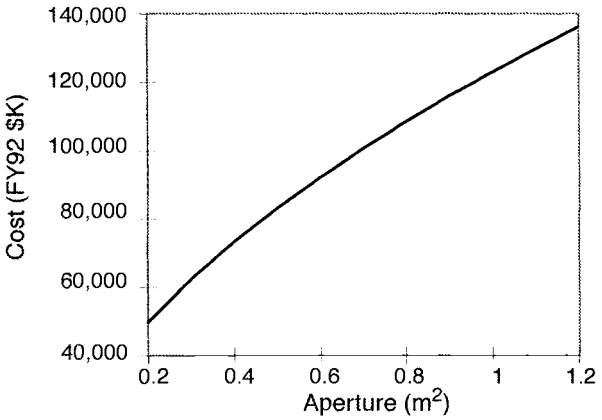


Fig. 15.2. The USCM cost-estimating relationship for IR payloads.

- In partitioning the mission and allocating tasks among separate components, total hardware resources required on orbit are often increased because of having to add redundancy to overcome serial reliability problems, as discussed in Sec. 15.2.3. Consider a single satellite satisfying a demand with reliability of 0.9. To achieve the same overall reliability with satellites of half the size, an extra redundant satellite is necessary. In this example, total resources on orbit for the distributed system are therefore 50% more than for the single deployment. Increases in total resources can also result from the nonlinear relationship between detection probability and SNR. A set of smaller apertures is often less efficient at detection than a single aperture of the same total size. Since the CERs base cost on characteristic resource, the net result of this increase of total hardware is an increase in cost.
- Typically, the USCM power laws in the CERs are nonlinear, with an exponent less than unity. The marginal price per kilogram of mass, or per meters squared of aperture, is higher than the price for smaller systems. Figure 15.2 demonstrates this trend. As a result, it is more expensive to divide a large system (especially aperture or power) into smaller components.

It would therefore appear that distributed satellite systems are characteristically more expensive than singular deployments. However, there are additional factors that can sway the balance in favor of distribution.

First, there is a question about the validity of using the USCM for estimating the cost of modern distributed satellite systems. The basic problem is that the model is based on regression from historical data of past military satellite programs. As such, the CERs of the USCM may not reflect modern trends or practices. The programs from which the model was derived were not subject to the same budget constraints as modern systems. Stated simply, past military satellite programs were expensive because they were allowed to be. Second, conventional cost models, being based on historical data, reflect an industry that was crippled by a conservatism and a reliance on risk avoidance. The high baseline cost of space systems was perhaps the largest reason for the conservatism. The enormous initial expenditure, added to the characteristically high risk, led to a reliance on tried and tested practices and established technologies. Unfortunately, this doctrine was self-supporting, being usually more costly than modern alternatives, and thus serving only to refuel the conservatism.

There is, however, some indication that changes are occurring. The advent of small satellite technology has hailed a new era of satellite engineering that minimizes costs by risk management rather than risk avoidance.[12,13] A willingness to accept some risk can lower the cost of satellite programs, enabling more missions to be flown and allowing new technology and innovative techniques to be implemented.[11,14,15] The use of commercial-off-the-shelf (COTS) technology can lead to substantial cost savings in development and operations (legacy systems often require specially trained operators). By accepting high risk and implementing strategies to manage failures, small satellites have been successfully designed, built, and operated at a fraction of the cost of traditional systems.[16] Should distributed satellite systems really proliferate in the market, they will achieve low costs by lowering the requirements on individual satellite reliability, taking advantage of the redundancy built into the architecture.

The changes in the space industry have not been restricted to the small satellite arena. The commercial satellite industry is just now beginning to realize the benefits of modernized design practices. Hughes Space and Communications and Lockheed Martin are moving away from the concept of the "handcrafted" satellite. They are therefore enjoying enormous savings from adopting the "production-line" approach to satellite design and construction. Standardized bus designs with modular interfaces to many different payloads reduce the development time and simplify assembly and test. Recent developments in commercial distributed satellite systems (e.g.,

Iridium, Teledesic, Orbcomm) reflect this production-line approach to satellite manufacture, and result in cost reductions that were previously unheard of in the satellite industry. Whereas the CERs of the USCM assume a single-string design, favorable economies of scale can result from bulk manufacture. The production of a larger number of small units allows quicker movement down the learning curve. Lockheed Martin is apparently observing a 15% discount rate in the production of the 66-satellite Iridium system.[17] This discount is made possible by economies of scale in manufacture and by modifying the way that satellites are built and assembled. For example, Lockheed requires that the subcontractor (Raytheon, Inc.) for the main mission antennas of the Iridium satellites perform full subsystem testing prior to delivery of each unit. No further testing is done until full integration. Components that fail the integration test are returned to the manufacturer, and a new antenna is taken from the storeroom. This practice has greatly reduced costs and assembly time.[18] Such practices are poorly represented by existing cost models.

The cost of launching a satellite system can make up a significant portion of the baseline costs. This is especially true of distributed satellite systems featuring many small satellites. Typically, launch costs do not scale linearly with mass. The price per kilogram is higher for lower mass payloads. Unless bulk-rate contractual agreements can be made with launch providers, learning curve discounts do not apply to launch costs. This would suggest that the launch costs of distributed systems are very high compared to those of traditional singular deployments. However, although each satellite in a distributed system may be small, when the entire system is considered as a whole, it can be huge. Economies of scale support the larger launch vehicles. Therefore, subject to volume and orbit constraints, it is cheaper to deploy the initial constellation using large launch vehicles. An entire orbital plane of satellites could be deployed on a single launch, giving the added benefit of distinct performance plateaus. The initial launch costs of distributed systems therefore scale more like those of large satellites, and should be priced based on the total constellation mass rather than on the individual satellite mass. Note that replacement satellites (for system augmentation or for compensation of failures) can be launched on dedicated small vehicles, such as Pegasus or Scout, or as secondary payloads, utilizing the spare launch capacity on larger boosters. The cost associated with this replacement scales more like that of small satellite launches, and can offer significant savings compared to the replacement launch costs of traditional satellite systems.

Distributed systems also offer the possibility of being able to ramp up the investment gradually, in order to match the development of the market. Only those satellites needed to satisfy the early market are initially deployed. If and when additional demand develops, the constellation can be augmented. The cost of constructing and launching these additional satellites is incurred later in the system lifetime. Because of the time value of money, the delayed expenditure can result in significant cost savings.

Each of these factors helps to offset the apparently high costs suggested by conventional parametric cost modeling. Consequently, the baseline cost associated with a distributed satellite system may actually be smaller than for a comparable large-satellite design. However, this is not generally true, since the baseline cost is extremely dependent upon the application. Some missions are more suited to distribution than others. An example of a mission that is well suited to distribution is passive IR imaging of the Earth, as shown in Example 6.

Example 6. Baseline Costs: Distributed IR Imaging System

For midwavelength IR payloads on low-altitude satellites, the payload costs scale with the resolution and the swath width of the instrument. Small swaths require less expensive satellites, but require more of them. The effect of these scalings can be quantified.

The payload cost for a single midwavelength IR satellite is the sum of the costs of the optics, the focal plane array of electrooptic detectors, and the computational resources needed to process the image. Canavan suggests that the cost of the optics for instruments of this type scales with volume rather than area.[19] The volume of an optic scales as D^2f, where D is the aperture diameter and f is the focal length. To achieve a resolution of Δx, the aperture size for diffraction-limited optics is:

$$D = \lambda r_{max}/\Delta x,$$

where r_{max} is the maximum range to the target. For a satellite at a low orbital altitude h, covering a swath of half-width W, the slant range is given by:

$$r_{max} = (W^2 + h^2)^{0.5}$$

and may be dominated by the cross-range component. For a constellation of satellites, the swath width of the instrument is dependent on the number of satellites in the constellation and the revisit time required of the system. Small revisit times require more satellites and larger swaths. The revisit time T for a constellation of N satellites is given by:

$$T \approx \frac{4\pi R_e^2}{2zVWN}$$

where R_e is the Earth's radius, V is the along-track velocity of the satellite, and z is a constant (~3) that depends on the extent and uniformity of coverage in latitude.[19] Inverting this relationship gives the swath half-width in terms of revisit time and constellation size.

The focal length of the optics is related to the resolution requirements and to the size Δd of the IR detectors that are available:

$$f = h(\Delta d/\Delta x)$$

Smaller detectors lead to smaller focal lengths, and a great deal of effort has been expended in trying to shrink IR detectors. Currently, several commercial detectors are available in the midwavelength band, with sizes ranging from 17 to 100 μm. This gives the cost of the primary optics as:

$$\text{Optics cost} = a\frac{h^3\lambda^2\Delta d(1 + W^2/h^2)}{\Delta x^3}$$

where a is the cost density ($/$m^3$). Canavan suggests $\$10M/m^3$ is a reasonable cost density for modern optics.[20] To be conservative, let us assume that the optics cost an order of magnitude more than Canavan's estimate: $a = \$100M/m^3$.

The cost of the focal plane array (FPA) scales directly with the number of detectors in the focal plane. This number is dependent on the swath width and on the dwell time requirements of the detectors. Long dwell times mean that the detectors cannot be scanned as quickly, and more detector elements are needed. The dwell time t_d can be calculated from the required sensitivity of the IR device. This sensitivity is measured by the noise equivalent temperature difference, $NE\Delta T$, which quantifies the minimum detectable change in apparent scene temperature from one pixel to the next during a scan:[21]

$$NE\Delta T \approx \frac{C(\lambda)}{\lambda^2 \sqrt{t_d}}$$

where $C(\lambda)$ is a function of wavelength for each detector. For HgCdTe detectors with $\Delta d = 40$ μm, $C(3\ \mu m) = 1.3 \times 10^{-12}$. A $NE\Delta T$ of ~0.5 K is considered a good IR system. By inverting this relationship, the dwell time can be calculated. This calculation then allows an estimation of the number of detector pixels in the instantaneous FOV:

$$N_{FPA} = \frac{N_x N_y t_d}{P},$$

where N_y is the number of pixels scanned in the along-track direction over an orbit, $N_y = 2\pi R_e / \Delta x$, and N_x is the number of pixels scanned across track, $N_x = 2W/\Delta x$. The cost of the FPA is then calculated assuming a cost of $1 per detector, based on current levels of technology.[20]

The computation costs scale with the number of instructions that must be carried out each second. This number is equal to the product of the number of pixels across the swath width of the instrument (N_x), and the rate at which they are crossed ($V/\Delta x$). A computer capable of 100 millions of instructions per second (MIPS) can now be flown for about $100K, so the computation cost density is ~$0.001 per instruction per second.[20]

The total payload cost is the sum of the costs of the optics, the FPA, and the computation costs. The bus costs can be estimated by assuming a 20% payload mass fraction and a constant $77K per kilogram of mass. This payload mass fraction represents a compromise between that for typical large satellites (30%) and that for a small satellite (10%).[22] The payload mass is needed for this calculation and is estimated by assuming an average mass density of the optics of 1 g/cm^3, with an additional multiplicative factor of 2 to account for some extra margin.[21]

Total constellation cost can then be estimated by summing the costs for optics, FPAs, computers, and dry mass for each satellite, and multiplying by the number of satellites in the constellation. A discount factor to account for an expected learning curve must be applied, depending on the number of satellites produced. The discount factor is assumed to be 5% for less than 10 satellites, 10% for 10 to 50 satellites, and 15% for more than 50 satellites.[23]

Launch costs do not have to be calculated because, as discussed earlier, they should scale only with total mass on orbit. Since we already account for total dry mass, adding launch costs only alters the total system cost by a constant amount, without altering the trends.

Incorporating these equations in a spreadsheet lets us examine the effect of constellation size on cost for different orbital altitudes. Figure 15.3 shows this relationship for a system with the following performance parameters:

- Revisit time, $T = 25$ min
- Ground resolution $\Delta x = 30$ m
- $NE\Delta T = 2$ K
- HgCdTe detectors, tuned to 4 μm, $\Delta d = 40$ μm

The curves for hardware costs exhibit a minimum at a given amount of distribution. Increased constellation sizes reflect a separation of the overall task among more

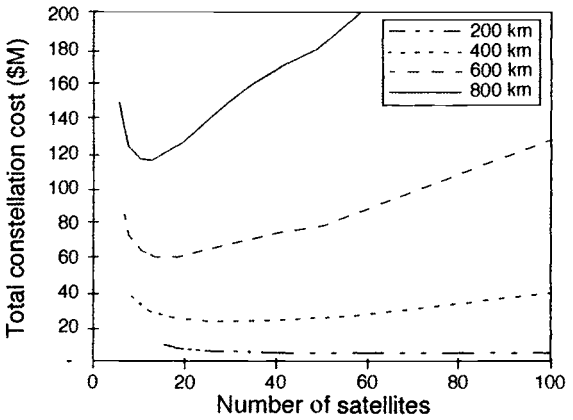


Fig. 15.3. Constellation cost for 25-min revisit time.

components, reducing the swath that each satellite is responsible for imaging. There appears to be an optimum swath corresponding to the level of distribution at the minimum point in the curves for each altitude. The existence of the optimum swath is a direct result of the quadratic nature of the optics cost with swath, and the hyperbolic relationship between swath and the number of satellites in the constellation. Neglecting learning curve effects, the total optics costs over the system therefore scale as (N + 1/N). Constellations with fewer satellites than optimum feature larger swaths, and consequently larger optics, FPA, and computation costs. Systems with more satellites than the optimum have increased costs, because the swath does not decrease fast enough to offset the increasing costs of more satellites. This is a good example of when distribution can lower the baseline costs. However, if the revisit time is increased to 60 min, the benefits of distribution begin to diminish. This is shown in Fig. 15.4. For revisits of longer than an hour, distribution incurs a cost penalty. This is because the swath for long revisit times does not need to be very big for a constellation of any size, and a large distributed system has too much wasted resource.

A candidate architecture can be chosen from these curves. The system parameters for a viable, low-cost architecture are shown in Table 15.1.

If the proposed microsatellite systems become a reality, the existing costing paradigm will change completely. Cost models that scale with unit cost, modified only by a learning curve, are not really applicable to microfabrication or batch-processing techniques. The microfabrication of solid-state components involves huge production runs, and the cost is reasonably insensitive to the actual number of components produced. An interesting caveat to be considered is the increased component reliability resulting from mass manufacture. As a result of the manufacturing process, mass-manufactured products have a very low variability in production standards and therefore have a characteristically high reliability. Consider, for instance, the reliability of home video recorders, an electromechanical component with many moving parts and extremely high tolerance requirements. Based on volume production, a typical VCR costing only a few hundred dollars has a reliability exceeding 0.999, the infamous triple-9 standard that drove the cost of the Apollo program.

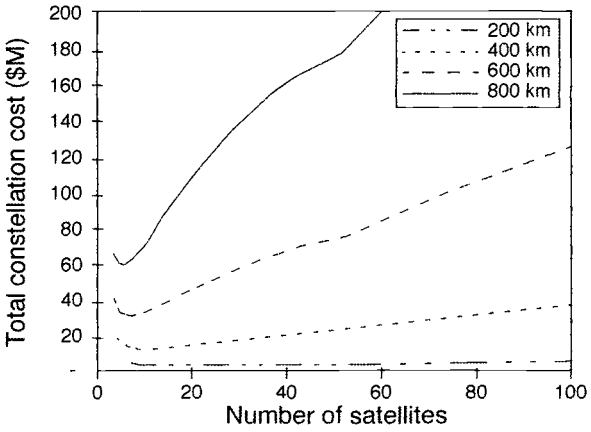


Fig. 15.4. Constellation cost for 1 h revisit time.

Table 15.1. Distributed IR Imaging System Parameters

	Value	Notes
No. of Satellites	50	Chosen for low cost
Orbital altitude	400 km	200 km too low due to drag
Revisit time	25 min	Requirement
Resolution	30 m	Requirement
Aperture diameter	6 cm	$D = \lambda r_{max}/\Delta x$
Focal length (m)	55 cm	$f = h(\Delta d/\Delta x)$
Payload mass (kg)	4 kg	1 g/cm^3 with 100% margin
Dry mass (kg)	20 kg	20% payload mass fraction

15.2.6 Reducing the Failure Compensation Cost

In addition to the baseline costs, failure compensation expenditure is necessary to replace components that fail during the lifetime of a satellite. The expected value of these costs can be estimated based on the failure probabilities of the system components. Clearly, the availability costs and failure compensation costs are closely related. The former is expenditure to reduce the chance of failures, while the latter is the expected cost of those failures. Generally, a larger expenditure in improving availability leads to fewer replacements and smaller compensation costs. Together, these costs can make up a significant amount of lifetime costs.

For certain satellite missions, a distributed architecture may lower lifetime costs by reducing the availability costs and the replacement costs. The impact of distribution on the availability costs was briefly introduced in Sec. 15.2.4. Distribution also affects the replacement costs, because they are closely related. A recent design study at MIT[24] showed that distributed systems appear to yield the greatest cost savings under two conditions:

- When the components being distributed make up a large fraction of the system cost. It is prudent to distribute the highest cost components among many satellites: *don't put all your eggs in one basket!*

- When the component being distributed drives the replacement schedule of the spacecraft within the system.

Some of these savings manifest themselves as follows:

- Replacements on average occur later, resulting in larger savings from discounting to constant year dollars.
- There are fewer replacements of overall components.
- The cost of replacing a single module in a distributed system is much less than that associated with the replacement of the entire satellite in a traditional design.

Example 7. Replacement Costs: Polar-Orbiting Weather Satellites

Instruments aboard polar-orbiting weather satellites,[24] such as the proposed National Polar Orbiting Environmental Satellite System (NPOESS), are classified as either primary or secondary. Because the primary instruments provide critical environmental data records, failure of a primary instrument necessitates replacement. A secondary instrument is one whose failure may be tolerated without replacement. If an orbital plane's complement of sensors are all located on a single satellite, failure of any primary sensor will require redeployment of all the plane's sensors. By distributing the primary instruments intelligently across a cluster of several smaller spacecraft, it may be possible to reduce the cost of the system over its lifetime because the plane's entire complement of sensors are not redeployed after every failure.

Consider the following three configurations, illustrated in Fig. 15.5. The blocks labeled as *A*, *B*, and *C* represent three primary instruments required in a given orbital plane. The total costs over a 10-year mission life were calculated for each of the three cluster configurations. As shown in Fig. 15.6, the costs over the 10-year period are broken up into three categories: initial deployment, required spares, and expected replacements. Initial deployment includes the development, production, and launch costs for each orbital plane's original complement of spacecraft. The number of required bus, payload, and launch-vehicle spares were derived from a Monte Carlo simulation of the mission, assuming given component reliabilities.[24]

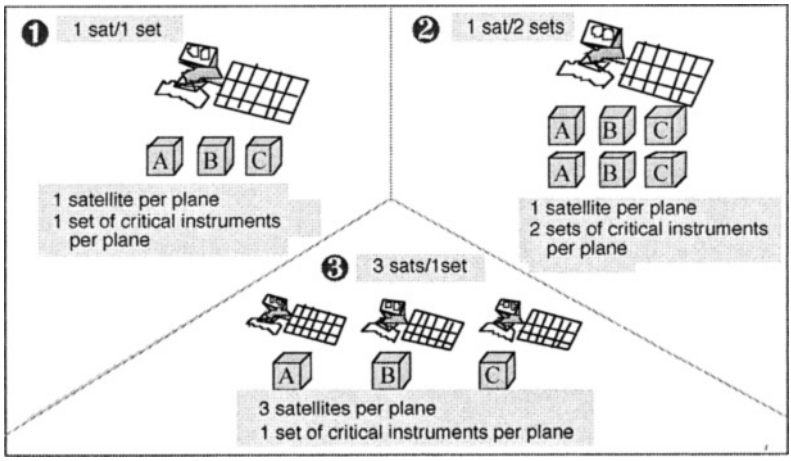


Fig. 15.5. Satellite and sensor configurations.

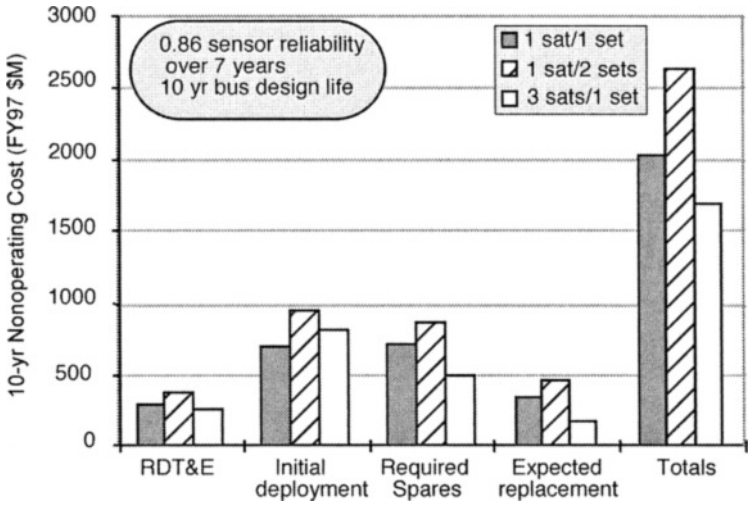


Fig. 15.6. Total costs over 10-year mission life.

Figure 15.6 shows that the initial deployment cost is least expensive for the 1 sat/1 set configuration. Adding a redundant sensor to the single satellite configuration greatly increases initial deployment cost in terms of larger bus size, additional instruments, and more expensive launch vehicles. The 3 sat/1 set configuration, although being launched on a less expensive vehicle, is slightly more expensive than the 1 sat/1 set configuration because of the duplication of bus subsystems and some sensors on each of the three smaller satellites. The figure also shows that adding a redundant sensor increases the cost, compared with configurations with a single primary instrument. The slight decrease in failure densities as a result of redundancy does not make up for the expense of additional sensors.

Distributing the primary instruments among three satellites significantly increases the reliability of each individual satellite. Higher satellite reliability and lower launch costs to costs to replace the satellites that fail result in the 3 sats/1 set configuration having the lowest expected replacement cost. Once again, the slight increase in reliability gained from adding redundant primary instruments for the 3 sats/2sets configuration is outweighed by the higher bus, payload, and launch costs.

To summarize, distribution within a satellite mission may reduce the replacement costs over the lifetime of a mission. A modular system benefits not only because a smaller replacement component has to be constructed, but also because there are huge savings in the deployment of the replacement. These savings are the greatest when the component(s) being distributed make up a large fraction of the system cost and drive the replacement schedule.

15.3 Issues and Problems

There are some issues that are critical to the design of a distributed architecture that were irrelevant to the design of traditional systems. Depending on the application, these issues might be minor hurdles, or could be so prohibitive that the adoption of a distributed architecture would be unsuitable or impossible. Some of the important considerations, characteristic of all distributed architectures, and particular to small satellite and microsatellite designs, are presented in this section.

15.3.1 Modularity vs Complexity

The potential benefits from distribution of satellite resources were stressed in Sec. 15.2. It was shown that improvements in cost and performance can result if architectures are carefully designed to utilize a segregation of resources into smaller, more modular components. By allocating individual system functions to separate satellites, and by adding redundancy, it was suggested that significant cost savings result from an increased availability and a reduced failure compensation. Furthermore, the enhanced capabilities offered by distributed architectures greatly expand the useful applicability of small satellites and microsatellites. For these reasons, an important issue to be addressed is the level to which a system should be distributed. How much can the system be divided into smaller components and still offer the benefits discussed previously? The central issue here is the trade-off between the advantages of modularization and the cost of complexity.

15.3.1.1 Modularization

As discussed in Sec. 15.1.3, the distribution of system functionality among separate satellites means that the system is essentially transformed into a *modular* information-processing network. The satellites make up individual modules of the system, each with well-defined interfaces (inputs and outputs) and a finite set of actions. Such systems are analogous to the distributed inter- and intranet computing networks, and are subject to similar mathematics. Distributed computing is a rapidly developing field, and a great deal of work has been done to formalize the analyses.[1] A lot of insight can be gleaned by adopting much of this groundwork.

One beneficial aspect of modularization comes from an improved fault tolerance. System reliability is by nature hierarchical, in that the correct functioning of the total system depends on the availability of each of the subsystems and modules that constitute the system. Early reliability studies[2] showed that the overall system reliability was increased by applying protective redundancy at as low a level in the system hierarchy as was economically and technically feasible. In addition, the reliability was increased by the functional separation of subsystems into modules with well-defined interfaces, at which malfunctions could be readily detected and contained. Clearly, subdividing the system into low-level redundant modules leads to a multiplication of hardware resources and associated costs. However, the impact of improved reliability over the lifetime of the system can outweigh these extra initial costs.

There are additional factors supporting modularization that are specific to satellite systems. As discussed in Sec. 15.2.4, baseline costs associated with a system of small satellites may be lower than the costs for a larger satellite design. Of even greater impact is the lower replacement costs required to compensate for failure. A modular system benefits not only because a smaller replacement component has to be constructed, but also because of the huge savings in its deployment.

All of these factors suggest that a system should be separated into modules that are as small as possible. However, there are some distinct disadvantages of low-level modularization that must be considered. The most important of these disadvantages are the costs and low reliability associated with complexity.

15.3.1.2 Complexity

The complexity of a system is well understood to drive the development costs and can have a significant impact on system reliability. In many cases, complexity leads to poor reliability, as a direct result of the increased difficulty of system analyses: failure modes were missed or unappreciated during the design process. For a system with a high degree of modularity, these problems can offset all benefits discussed previously.

Although each satellite in a distributed system might be less complex because it is smaller and has lower functionality, the overall complexity of the system is greatly increased. The actual level of complexity exhibited by a system is difficult to quantify. Generally, however, the complexity of a system is directly related to the number of interfaces between the components of the system. Although the actual number of interfaces is architecture specific, a distributed system of many satellites has more interfaces than a single satellite design. Network connectivity constraints mean that the number of interfaces can increase geometrically with the number of satellites in a distributed architecture. This is an upper bound; systems featuring satellites operating in parallel with no intersatellite communication (later defined as collaborative systems) exhibit linear increases in interfaces with satellites. Complexity of a distributed system is therefore very sensitive to the number and connectivity of the separate modules.

The impact of this additional complexity is difficult to evaluate, especially without a formal definition of how complexity is measured. Recent studies at MIT[25,26] would, however, suggest that complexity can cause significant increases in development time, increases in cost, and losses of system availability. For these reasons, the level of modularization must be carefully chosen. Only with thorough system analysis and efficient program management can the impacts of complexity be minimized.

15.3.1.3 A Lower Bound on the Size of Component Satellites

From very basic analyses, a lower bound on the size of the satellites of a distributed system can be estimated. To be a contributing element to the system, each satellite must receive information from some external source. Once this information has been received, the satellite may perform some processing and reduction before relaying the information to the next node in the network. This node may be a ground station or another satellite in the system.

Regardless of the eventual destination, information must flow through the satellite. Although some data reduction may be done, the actual flow of "primary information" must be conserved. Primary information is specified by the top-level requirements; for an early warning radar system, primary information represents target detections. Because of the continuity constraint on primary information, the satellite must be able to communicate with another node of the network at a rate conducive with system requirements. The delivery of the primary information to the destination node need not be immediate. For some applications, a store-and-forward method of delivery is preferable. In this method, the information is stored on board the satellite until it can be transmitted to the destination node. The continuity constraint therefore ensures that, at all times, information flowing into a satellite must be either stored or transmitted.

This required flow of information through the satellite leads to two simple statements, with associated bounds. The first statement says that in order to maintain extended operations during a single-duty period, the net information transmitted by a satellite must be equal to that received. The energy conversion system (solar arrays) of the satellite must be able to support this net transmission of information over the same duty period. If the satellite cannot provide enough energy to allow transmission of this quantity of information, requirements cannot be satisfied. The amount of energy required depends on the integrity requirements (driving the energy per bit), the distance to the destination node (free space loss), and the transmitter/receiver characteristics. The satellite must also provide the energy needed to receive the information in the first place. This is a small factor for passive signal detection, but can dominate in active systems. These are systems, such as radar and ladder, that illuminate a target and detect the return. The satellites must transmit a signal with enough energy to make the round-trip journey to the target and back while satisfying detection requirements. The target adds the information to the signal, but returns only a fraction

of the incident energy, depending on its cross section. Note that under this definition, and contrary to intuition, communication satellites are also passive; while they do retransmit information sent from another source, the definition of an active system relates to the detection of the information. It is assumed that satellites must eventually relay the received information to a destination node.

The second statement relates to data storage requirements placed on satellites: the amount of new information stored on the satellite at any instant is the difference between rate of information collected and rate of transmission. The net storage at some time t is the integration of this difference, from an initial time at the start of the duty cycle to the current time. For store-and-forward systems, the value of this integral initially increases with time as more data are stored, and then decreases to zero at the end of the duty cycle, when all the data have been downloaded. The maximum value of this integral defines the data storage capacity requirement. Note that this requirement can be very costly to satisfy, especially for remote-sensing systems. Consider, for example, a single panchromatic image of a 25-km^2 scene at 1-m resolution with 256 shades of gray. This single modest image requires 625 Mbytes of storage capacity.[27] Compression techniques can help, reducing this figure by as much as an order of magnitude. However, in order to store any reasonable number of images, a great deal of storage capacity is required on the satellite. Data storage devices are typically heavy and power hungry, and can consume a substantial portion of the satellite's resources. This is true for large satellites such as SPOT and Landsat, which weigh more than 1000 kg, so it would seem unlikely that small satellites with modest resources could satisfy similar storage requirements. Solid-state storage devices are now available that relieve this problem somewhat. The current state of the art in solid-state storage can buffer 1Gbit of data in static RAM, consuming only 5 W of power.[16] Nevertheless, data storage requirements place a severe constraint on the smallest size for a remote-sensing satellite. Of course, in distributing the task of collecting data among the many elements of a distributed system, the storage requirements for each satellite are reduced. This reduction may actually enable large constellations for use in remote-sensing applications.

Example 8. Data Storage: Distributed IR Imaging System

Returning to the distributed imaging system described in Example 7, we can estimate the storage requirements on each satellite. Assume that each pixel has a 4-bit value, corresponding to 16 shades of gray. The data rate of the IR detector on each satellite is then given by multiplying this 4 bits by the product of the number of pixels across the swath width of the instrument (N_x), and the rate at which they are crossed ($V/\Delta x$). The data must be stored until an opportunity for downlink arises. Maximum downlink interval is set by the requirement on responsiveness of the system. For near real-time applications, downlink opportunities must come frequently. This frequency is helpful for distributed systems, since storage capacities are limiting, and the interval between downloads must be as short as possible. Assuming 25-min revisit time for imaging, 5-min interval between downlinks, and minimum elevation from the ground station to the satellites of 20 deg, Fig. 15.7 shows data storage and downlink communication requirements of the satellites.

The figure shows that a 1-Gbit storage device is sufficient for systems with more than approximately 10 satellites. Communication data rates are high, but manageable for constellations with greater than 50 satellites at altitudes above 400 km.

Table 15.2 shows the important parameters for an architecture featuring 50 satellites at 400 km altitude. The number of downlink stations is calculated to ensure a 5-min downlink interval.

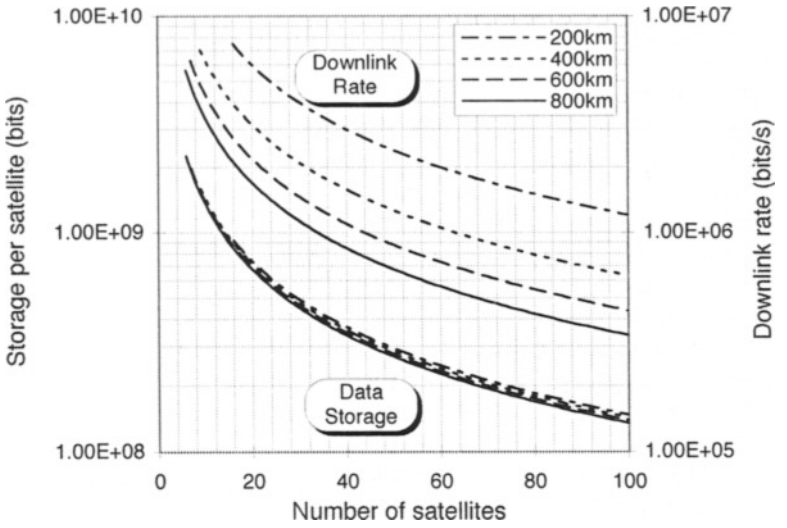


Fig. 15.7. Data storage and communication data rates for 25-min revisit time, with 5-min intervals between downloads.

Table 15.2. Distributed IR Imaging System Communication and Storage Parameters

	Value
No. of Satellites	50
Orbital altitude	400 km
Revisit time	25 min
Downlink interval	5 min
Maximum data storage	0.29 Gbits
Downlink data rate	1.25 Mbits/s
Number of ground stations	43

15.3.2 Automation in Control and Operations

Current levels of automation in satellite systems reflect an incremental evolution that is based on a high level of human involvement. Historically, human involvement has resulted from the desire to reduce risk and from limited technological capabilities. Because of the dependence on humans to perform tasks, operation costs can make up a significant portion of the life cycle costs for a satellite system.[28] In addition, human error continues to be a major cause of spacecraft anomalies and failures.[29] With the introduction of large constellations or clusters of satellites, some automation of operations will be required to reduce costs while maintaining availability (the probability of meeting system requirements at a given time).[30–32]

Despite the recognition of the need for automation, there is reluctance to implement automation. The large investments and high risks involved in space ventures have led to a conservative industry. In addition, the desire to reduce cycle times for new programs favors significant re-use of proven technologies, and thus low levels of automation. A methodology that can be used to predict the effects of automation on a satellite system has been developed at MIT,[33] and may

enable more cost effective systems to be considered. The definition of performance metrics can enable operations concepts to be quantitatively compared. In particular, operations concepts have a profound impact on the system availability (probability of meeting system requirements) and life cycle costs (sum of development, operations, and opportunity costs).

Figure 15.8 qualitatively represents the cost and availability characteristics of a hypothetical satellite system with respect to an increasing level of automation. As low levels of automation are introduced into the system, the operating costs decrease, principally due to a decrease in the number of human operators (Fig. 15.8a). At some point, however, the increases in design and development costs due to software development outweigh the decrease in operation costs.

As shown in Fig. 15.8b, availability may decrease, increase, or be unaffected by the level of automation. For tasks that are simple, well understood, or periodic (such as routine stationkeeping on a geostationary satellite), availability may increase with increasing automation (Task A). This increase is true when human errors are more likely than software errors, or when the impact of unanticipated situations is negligible. For complex, rare, or unexpected functions, availability may decrease as humans are removed from the loop (Task C). This decrease can occur when the automation is unable to resolve problems that could have been resolved by a human, or when the automation fails to accurately inform the human of the situation. There may be some functions for which the availability is nearly independent of the level of automation (Task B).

An increase in system availability translates into an increase in revenues for a commercial system or an increased ability to perform an objective for science or military systems. Thus, the potential for failure can be represented by an opportunity cost that represents revenues forgone as a result of increased system downtime. For commercial systems, the opportunity cost can be added to the development and operations costs to form the life cycle cost. Determination of opportunity costs requires additional data, such as the relationship between a particular function and revenue. For science or military applications, the definition of an opportunity cost may be difficult. In such cases, the development and operations costs would be compared against the availability, without attempting to define the life cycle cost. The combination of the two curves from Figs. 15.8a and 15.8b are represented in Fig. 15.8c, showing the overall life cycle cost, which is defined as the sum of development, operations, and opportunity costs. In the example of Fig. 15.8, there exists an optimum level of automation at which life cycle cost is minimized. Because of the complexity of the satellite system, a methodology is needed that can model the effect that automation has on costs and system availability. Such a methodology would enable system engineers to identify those functions that should be automated.

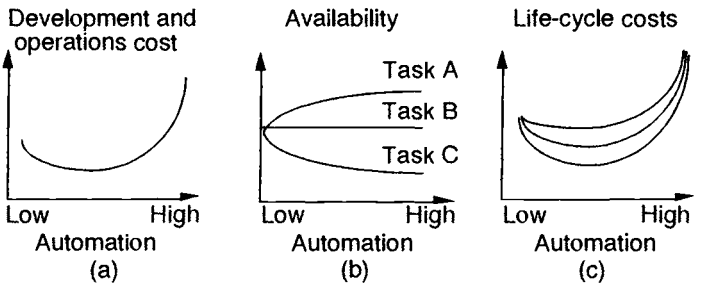


Fig. 15.8. Effect of automation on cost and availability.

15.3.2.1 Levels of Automation

In recent years, a great deal of attention has been paid to the interfaces between machines and their human operators.[33] Of specific interest in the space community is the problem of remote operation. Despite the popularity of the topic, however, most of the efforts have been involved in isolating failure modes of the combined system and providing insight into control issues. The idea that automation can be introduced gradually is recognized, but definitions of discrete levels tend to be fuzzy.

In general, increasing the level of automation shifts responsibility from the human operator to the ground processor or spacecraft processor. As can be seen in Fig. 15.9, there are two processor interfaces to which automation may be applied: (1) between the human and the ground computer processor, and (2) between the control center and the spacecraft processor. Therefore, the overall level of automation for a given task or function must be defined by each of the levels of automation at the two interfaces.

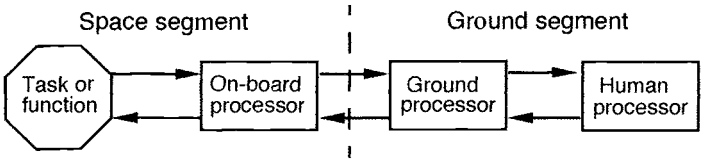


Fig. 15.9. Processor interactions to perform a task.

Regardless of the level of automation, one of the processors is generally responsible for performing the task, and is termed the primary processor. Sometimes, the primary processor may be unable to complete the task. This may be due to a hardware or software failure, to human error, or to a state of the system for which the software was not designed. Also, there may be times when the software has been designed to not attempt to perform the function under certain conditions. The inability of the primary processor to perform the task is termed a processor deficiency. When such a deficiency has occurred, a secondary processor may be defined to take over the function. A tertiary processor may also be defined to recover from a secondary processor deficiency.

Satellite systems automation can be accomplished by placing responsibility for tasks on space- or ground-based computers rather than on a human operator. However, even in an "automated" system, the human is often not completely out of the loop, but may monitor the computer processor. In order to preserve generality in the model, it is important to capture this subtlety of adding automation. To this end, discrete levels of automation have been defined that can describe the application of automation incrementally. Thus, a gradual transition is allowed from fully unautomated (human control) to fully automated (no human intervention). Each level of automation defines the degree to which each processor (human or computer) is responsible for the completion of a task.

The following sections describe the levels of automation associated with the satellite-control center interface (remote interface), shown crossing the dashed line in Fig. 15.9. The definitions for the levels corresponding to the control center-human interface are analogous to those for the remote interface.

There are six levels of automation defined,[25] ranging from full automation to no automation. Each level represents a variation in the responsibilities and information passed between the space processor and control center. Thus, the assignment of primary and secondary processors varies with the level of automation.

- **Fully automated.** Full automation implies that the space processor is the primary processor, and performs the function with no intervention from the ground control center. Since information regarding the task is not shared with the ground, if the space processor fails to correctly perform the task, the ground would not be notified. An example of a fully automated system is a "fire and forget" air-to-air missile fully responsible for finding and destroying the target, but the pilot is unable to enter the loop to ensure that the mission is accomplished.
- **Paging.** With paging, the task is performed by the space processor, but the control center is notified at the occurrence of any failure or processor deficiency. Thus, the on-board processor (OBP) is the primary processor, and the control center is the secondary processor. The major advantage of paging is that the control center is informed of problems as they occur, but dedicated ground computers or personnel do not continuously monitor the satellite. An example of a system with paging would be the computer diagnostics system of a modern automobile. The computer checks and adjusts variables such as fuel-air mixture to compensate for poor performance, as indicated through the car's oxygen sensor. If the on-board system is unable to bring the oxygen reading within desired bounds by itself, the "service engine soon" light on the dashboard goes on. The mechanic must be paged, or, in this case, the car itself is "paged" and brought into the shop. Note that until the car has trouble, the mechanic in the shop is fully unaware of the car's status.
- **Supervision.** The task is nominally executed by the space processor (primary processor). However, the space processor's actions are monitored continuously by the control center (secondary processor). The ground may also be informed through an alarm if a primary processor deficiency should occur. The ground station may assume responsibility for the task at any time. When the human is the secondary processor, supervision effectively increases the probability of a timely response (compared to paging), since the human would already be monitoring the control center. Also, because the ground is continuously monitoring the task, the human would be able to watch the evolution of events unfold, and therefore might be more likely to correctly interpret the reason the primary processor could not perform the task. Examples of a system operating under supervision might be an aircraft flight management system or a nuclear power plant. Although a computer is responsible for maintaining the safe operation of the system, human operators are required to closely watch for deviations. Safety in a nuclear environment demands this response time to be small, and therefore supervision is used rather than a level of automation such as paging. With supervision, the secondary processor implicitly agrees with each of the primary processor's actions.
- **Cueing.** At the cueing level, the system may require ground intervention at various points during the task's execution. The satellite still attempts the task by itself and therefore remains the primary processor, but must first obtain authorization from the ground. This is analogous to a PC prompting the user to verify any delete actions. Although this required intervention decreases the probability of incorrectly performing the function, it may also increase the time required to perform the task. Thus, increased availability may come at the cost of a more time-consuming process. With cueing, the secondary processor must explicitly agree with the primary processor's actions.
- **Data filtering.** At this level of automation, the remote system is not responsible for performing the task, but aids the ground by filtering the downlinked information. Therefore, the control center is the primary processor. During a failure of the function (e.g., flight hardware failure), the satellite could send a more detailed set of data to the ground. This could tend to lessen the data load on the ground and should ease the identification and analysis of failures. A good example of data filtering can be found in many modern satellite operations systems

(e.g., Integral System's Epoch 2000, or FIXIT).[34] Rather than scrolling raw telemetry to the screen, as was done in the past, some newer systems present the data in a graphical format, which allows the data to be interpreted more easily and accurately by a human operator.

- **No Automation.** For this level of automation, there is no presorting of data before transmission to the ground. The data must be interpreted by the ground in raw form, and the task is fully performed by that processor. This may result in some development cost savings relative to a more automated system, due to less software, but may increase operations costs, due to the required operator workloads.

15.3.2.2 Automation in Constellations of Satellites

A model has been developed at MIT[33] to determine the effects of satellite system automation on system costs and availability. Several preliminary studies have been performed to demonstrate the model's capabilities and to obtain some general results. The model's cost and availability outputs are sensitive to the model inputs, so only general trends are discussed. More specific answers to the automation trade-off require a more detailed analysis, as described in Sheridan.[33]

Constellations of satellites introduce a twist in the automation trade-off, since automation is usually associated with high nonrecurring costs. As the level of automation increases, operations costs are expected to decrease, and development costs are expected to increase. As the number of satellites increases for a given level of automation, the operations costs increase linearly. Development costs, however, lag a linear relationship, due in part to a learning curve and in part to the much lower recurring costs of automation relative to the nonrecurring cost. Thus, functions not suitable for automation for a single satellite may be desirable for a constellation. In fact, crossover points can be identified that define constellation sizes over which certain levels of automation are optimal. In the generic communication satellite system used in the MIT study, for constellations below 60 satellites, the optimal level for the payload was no automation. Above 60 satellites, however, paging resulted in the lowest life cycle cost per satellite. The actual locus of each crossover point and the optimal levels of automation are sensitive to the model inputs as well as to the function being automated.

Although the results stated above may seem obvious, they are important. When considering automation for a large constellation of spacecraft, the engineer must be careful in using previous automation trade-off results. The optimal level of automation for a function on a single satellite is not generally the optimum for a constellation.

A general result that was consistent throughout the studies performed was that automating to the cueing level resulted in the highest life cycle costs. This is because with cueing, there are high development costs associated with automating to the point that a computer can suggest a solution, but there is also a high operations cost since the human is always required to authorize any actions. In effect, the automation is developed, but not trusted. Although specific implementations of cueing may not have the same results, system designers should be wary. Often, high aspirations lead to capable systems that are not trusted, and the results obtained here may be observed.

Although high levels of automation may be desirable for some functions and some missions, the fully automated level is rarely optimal. With full automation, the human is never made aware of failures when they occur. If the automation is unable to recover from a failure, the failure becomes permanent. Thus, there is some chance that a repairable failure will not be repaired, effectively increasing the probability of a permanent failure. The paging level of automation requires very little human involvement, and operations costs are very small. However, even limited human involvement is useful when the automation software makes a mistake.

15.3.2.3 The Future: Microsatellite Constellations

Constellations of microsatellites will likely contain a very large number of satellites in order to perform a useful mission. As discussed previously, for a given level of automation, as the constellation size increases, operations costs will scale linearly, while development costs will take advantage of economies of scale. Therefore, operations costs for these large constellations will dominate the life cycle costs, and high levels of automation will be required for many functions.

Despite the potential benefits, automation alone may not be enough to make constellations of microsatellites more cost effective than smaller numbers of larger satellites. Automation may need to be taken a step further by the widespread use of passive functionality. An example of passive functionality is the use of a gravity gradient for attitude control. The need for a computer or human to perform attitude control may be greatly reduced, or eliminated. In turn, the satellite's operations and development costs for the attitude control function may be greatly reduced or eliminated. Of course, there may be some degradation of performance, but as shown in Sec. 15.2, individual spacecraft performance degradation may in some cases be tolerated if satellites work together to perform a task. Passive functionality may also be used for other functions. Power generation for spinning satellites is to some extent passively controlled since no array pointing is required. Passive communications in the form of omnidirectional antennas may also permit the degradation of pointing accuracy. For navigation and propulsion, passive functionality may mean eliminating all forms of orbit control. Since there are so many satellites, it may be cheaper to allow them to drift naturally, and reinsert replacements into sparsely populated orbits as needed. In essence, passive functionality helps to reduce the complexity of the system, thereby reducing both development and operations costs.

In addition to reduced functions, cost savings may result through a forced reduction in human-intensive failure recovery operations. Teledesic has adopted this strategy. Teledesic's operations strategy only allows humans to be involved in a recovery for 4 h. If the problem is not fixed after the allotted time, the spacecraft is deorbited and replaced. For such large constellations, the development cost per satellite may be low enough to allow such a strategy. For huge constellations of microsatellites, truly disposable satellites may be cost effective. In other words, it may be less expensive to never involve humans in recovery actions, and to just let satellites fail. The law of large numbers will ensure that individual failures will be uniformly distributed over the constellation. The system may be oversized to anticipate the fractional degradation of the constellation due to individual failures.

Another possible problem area with operations of huge constellations may be on-orbit reconfiguration and reprogramming. One solution to this problem takes advantage of proposed system architectures that feature several different kinds of satellites, each tailored to perform specific tasks. In such a configuration, a few larger "mother ships" may be deployed that contain instructions for the microsatellites. This allows the centralization of certain mission-specific information. By designing each microsatellite to use the mother ships as configuration databases, updates can be made without contacting each of very many individual microsatellites. This concept may be taken a step further by incorporating multicasting, thereby allowing several species of microsatellites to take commands from the same mother ship.

15.3.3 Clusters and Constellation Management

In Sec. 15.2, it was argued that by combining the capabilities of many individual elements, systems of small satellites or microsatellites can be used for high-rate or high-resolution applications. Several architectures designed to utilize this concept are gaining popularity within the space community. It has been suggested that satellite clusters can offer significant advantages for remote

sensing and communications. The creation of a cluster requires that several satellites can simultaneously view the same target. The number of satellites in the cluster is the same as the level of multifold coverage supported by the system.[35]

For many applications, the relative positions and dynamics of the satellites in the cluster are a critical factor in the design. There are two options:

- **Local clusters**, in which a group of satellites fly in formation. The relative positions of the satellites are controlled within specified tolerances.
- **Virtual clusters**, in which a subset of satellites from a large constellation make up the cluster. At any time, those satellites that are in close proximity can make up the virtual cluster. The actual constituent satellites and their positions constantly change, subject to the orbital dynamics of the constellation.

The most suitable choice depends on the application. Consider, for example, using a cluster to form a sparse aperture for high-resolution imaging of terrestrial or astronomical targets. By coherently adding the signals received by several satellites, the cluster would create a sparse aperture many times the size of a real aperture. The phasing and the optical paths of the electromagnetic waves would have to be carefully controlled so that the signals could combine coherently. The tolerance is typically $\lambda/20$. Stationkeeping is therefore a large problem for a (static) local cluster. Provided that the satellite positions are controlled within the $\lambda/20$ tolerance, the processing requirements on the satellites are reduced. Conversely, the element positions of a virtual cluster continuously vary; therefore, to correctly phase the signals, these locations must be known with a high degree of accuracy at all times. As a result, the virtual cluster has slack stationkeeping requirements, but needs a great deal of intersatellite communication and processing to ensure coherence. This coherence issue is discussed in Sec. 15.3.4. The remainder of this section details the propulsion requirements necessary to maintain the relative positions of satellites orbiting in a local cluster.

15.3.3.1 Local Clusters

Satellites in a local cluster orbit in formations such as those shown in Figs. 10 (a) and 15.10 (b). Although one or more reference satellites are in standard Keplerian (i.e., inertial) orbits, maintaining the formation requires the other satellites to orbit in planes parallel to the reference orbits. These noninertial orbits are characterized by either a focus that is not located at the Earth's center of mass (Fig. 15.10 [a]) or orbital velocities that do not provide the proper centripetal acceleration to offset gravity at that altitude (Fig. 15.10 [b]). As expected, the Earth's gravitation acts to move these satellites into Keplerian orbits.

The satellites are subjected to "tidal" accelerations that are a function of the cluster baseline and orbit altitude. To remain within the allowable relative position tolerances, these accelerations must be counteracted by thrusting. Options are to use either a series of impulsive thruster firings at regular intervals or to apply a lower, continuous thrust.

Figure 15.11 compares the ΔV expended for continuous thrusting to that for impulse thrusting. The figure shows that as the position tolerances on the cluster tighten, the ΔV expended for impulsive thrust approaches that expended for continuous thrust. This is because the thrusting intervals must be very short and frequent. At these short thrusting intervals, only minimal propellant savings result from impulsive thrusting. For a given mission life, a satellite stationed far from the reference orbit consumes more fuel than a satellite stationed closer. It may be desirable, therefore, to rotate the positions of the satellites during their lives to distribute the fuel consumption among all the satellites in the cluster. Figure 15.12 shows that the ΔV budget can be reduced to 54% that of the worst-case satellite in the cluster.

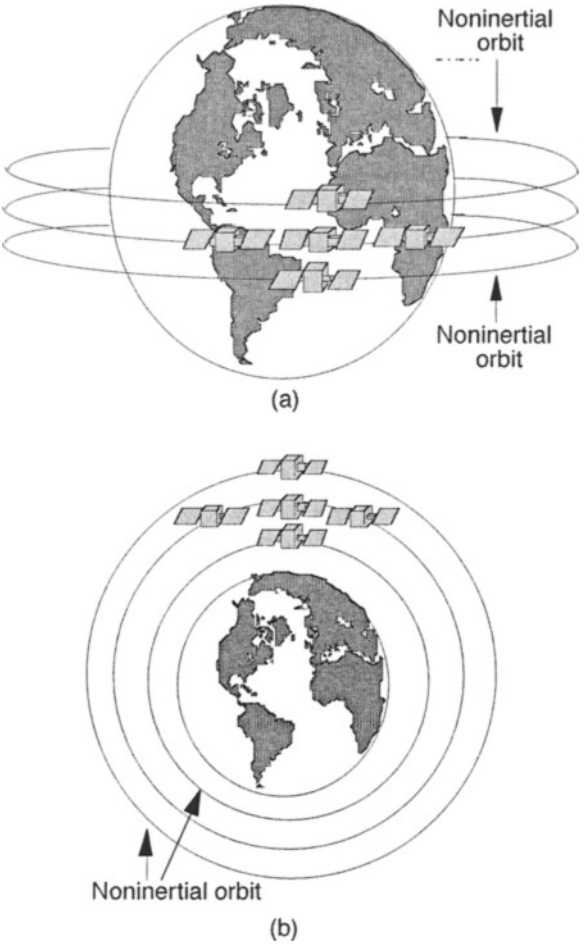


Fig. 15.10. (a) Cluster formation normal to reference orbit plane, (b) Cluster formation in plane of reference orbit.

At a rate of one rotation during the mission, the ΔV expended is about 65% of the worst case. As rotation rate is increased beyond that shown in the figure, the ΔV required just to maintain the motion around the cluster begins to dominate. At a rate of a few hundred rotations during the mission, the ΔV ratio increases above unity, indicating it is no longer beneficial to rotate cluster satellites. For a virtual cluster, all satellites travel in inertial orbits, and the satellite array is formed using whatever satellites are available as they fly over the region of interest. The ΔV for maintaining the virtual cluster is therefore zero.

15.3.3.2 Propulsion System Requirements

Based on acceleration levels presented in the previous section, the propulsion system requirements for maintaining a local satellite cluster can be calculated. The feasible range of specific impulse (Isp) and efficiency (η) are the primary characteristics of interest. Specific impulse of a thruster is a measure of the thrust-per-unit mass flow rate of propellant, while efficiency is defined as the ratio between the realized propulsive power and the input power. To determine feasible

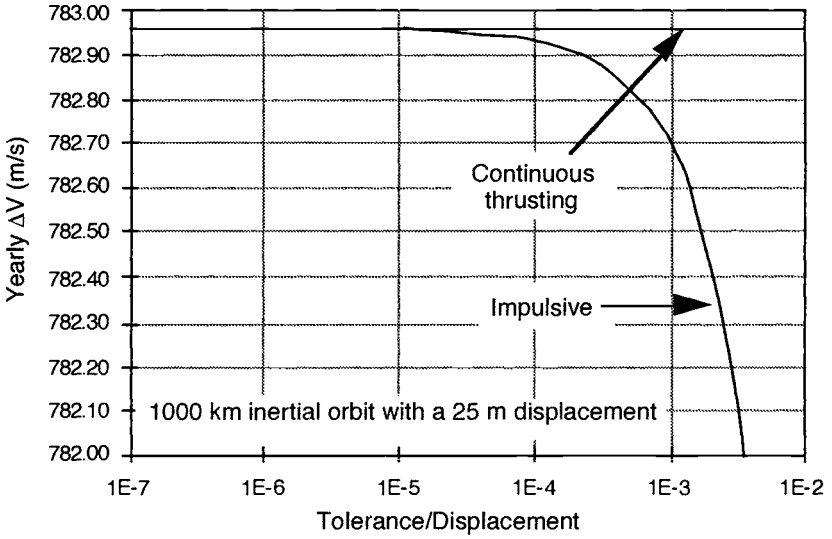


Fig. 15.11. Propellant savings from impulsive thrusting.

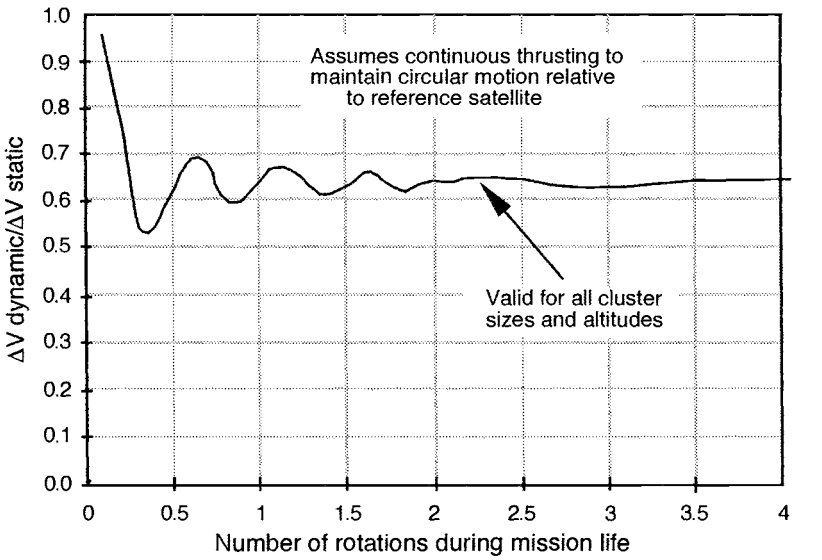


Fig. 15.12. Reduction in cluster maintenance ΔV.

ranges of thruster Isp and η to maintain a local satellite cluster, some constraints must be placed on the power, size, and mass allocated to the propulsion system. These constraints allow margin for the satellite to accomplish tasks other than simply operating the thrusters. There must be sufficient volume, mass, and power available on board the satellite to perform payload operations. For the examples presented here, the propulsion system was constrained to be less than 30% of the satellite mass and was allocated 20%, at most, of the power resources. The size of the tanks was limited to a third of the diameter of the spacecraft.

Figures 15.13 and 15.14 illustrate the feasible regions of thruster-specific impulse vs efficiency for clusters orbiting at various altitudes, assuming a satellite mass of 100 kg. The feasible regions are indicated by the triangular areas in the figures; the capabilities of current thrusters, by the shaded regions. At specific impulses below the feasible regions, the propellant required to maintain the cluster during the 5-year mission life exceeds the assumed 10% maximum initial propellant mass fraction. Combinations of specific impulse and efficiency above and to the left of the feasible regions result in thruster power requirements greater than the allowable 20% of

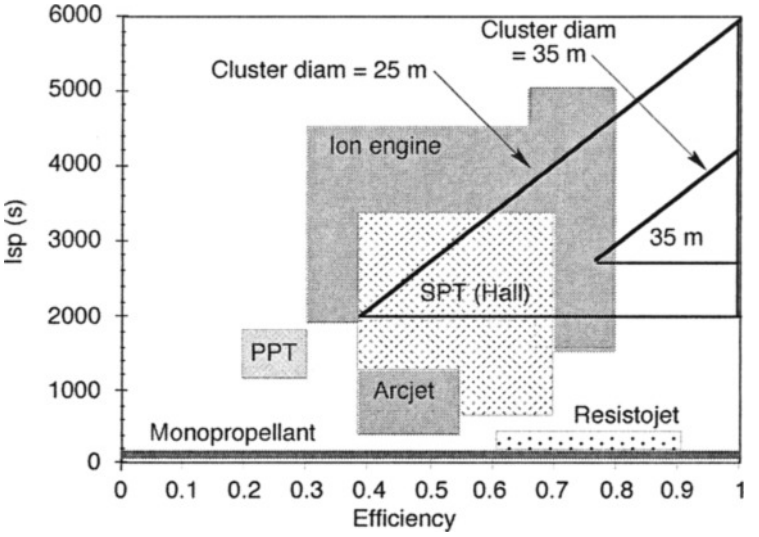


Fig. 15.13. Feasible Isp vs η at 1000 km.

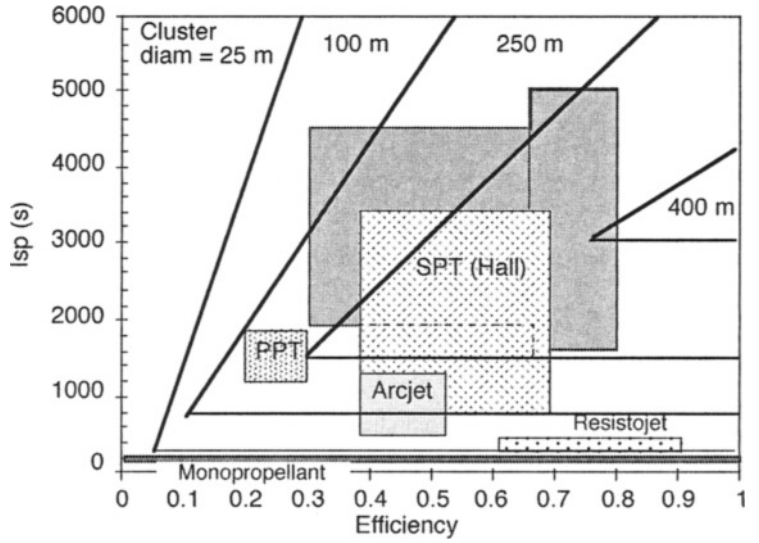


Fig. 15.14. Feasible Isp vs η at 10,000 km.

spacecraft power. An increase in thruster efficiency allows for higher specific impulses without exceeding the power limitations.

As shown in Fig. 15.13, a satellite in a cluster orbiting at 1000 km altitude with a 25 m baseline requires thrusters operating at a minimum specific impulse and efficiency of 2000 s and 40%, respectively. Stationary plasma thrusters (SPTs) and ion engine technology can currently achieve these ranges. If the baseline is increased to 35 m, thruster requirements increase to 2700 s specific impulse and 80% efficiency. Similar effects can be seen in Fig. 15.14. At 10,000 km, cluster baselines in the hundreds of meters can be achieved using SPT or ion engine technologies. Note that at no altitude can chemical propulsion maintain cluster formations for the 5-year lifetime.

15.3.4 Spacecraft Arrays and Coherence

Some have proposed the use of large constellations of satellites, each with a small antenna, for forming extremely large sparse apertures in space. This spacecraft array concept has received a great deal of attention from many sectors of the space community, mostly because of the potential it offers for high-resolution sensing and communications. There are two different ways to form these apertures. A static array can be defined using a local cluster, in which the relative satellite positions are controlled.

Alternatively, a virtual cluster can be defined by a group of satellites from a large constellation. At any time, some subset of the constellation will be in view of some desired region of the Earth that is to be sensed. If the group of satellites is in low Earth orbit (LEO), its relative motion means that as time progresses, the subset of satellites used to form the array changes. The array is therefore changing continuously, with the separations and the numbers of elements following a reasonably random pattern.

The spacecraft array idea was discussed in Example 1. To recap, a large thinned aperture is formed from a set of satellites, with each satellite acting as a single radiator element. The angular resolution of any aperture scales with the overall aperture dimension, expressed in wavelengths. The SNR achievable by the array is directly proportional to the number of constituent elements. The spacing of the elements is much larger than one-half of the wavelength, so grating lobes are avoided only if there are no periodicities in the elemental locations. The positions of the satellites must therefore be randomly distributed. For local clusters, this constraint effectively slackens the stationkeeping requirements, but introduces accurate position knowledge. Provided the satellite positions are always known to a high accuracy, there seems little reason to expend fuel to maintain a random location. A virtual cluster achieves the random distribution of satellites by default.

Unfortunately, there are many technical difficulties involved with the design and construction of such a system, mainly due to the requirement for signal coherence between large numbers of widely separated apertures. This is especially true for systems intended for Earth observation. Interferometric techniques are not well suited to Earth observation from orbit, since the Earth forms an extended source, unlike the astronomical sources that lie embedded in a cold cosmic background. This extended source forces a need for very high SNRs and high sampling densities,[36] leading to designs featuring a large number of satellites. For high-resolution imaging applications requiring either long integration times or high SNRs, the situation is made worse by forward motion of LEO satellites limiting the time over the target. This limited time forces more simultaneous measurements to be made in order to reach the required SNR and therefore requires even greater numbers of elements. Furthermore, although there are no grating lobes to consider, the thinned and random array exhibits large average sidelobe levels. In general, the ratio of the power in the main lobe to the average sidelobe power is equal to the number of elements in the array.[37] For most detection applications, the maximum sidelobe power should be much lower

(more than 15 dB lower) than the main beam power. Using this measure results in large bounds (on the order of 50) on the minimum number of satellites that must be used to form the sparse array.

The formation of sparse apertures using a large number of satellites is complicated by data processing, presenting a barrier to the adoption of sparse array technology. The most generic problems, common to both static and virtual clusters, involve using spacecraft arrays as receivers. The signals from each element of a receiver must be combined coherently. The data-processing requirement scales quadratically with the number of elements, and the equipment becomes very costly as the aperture size grows. The actual exchange of signals between receivers and combiners also poses a difficult challenge. For an interferometer, the exchange is done simply by routing the analog signals from the pair of collectors to a common combiner, constraining the optical paths to be equal in each case. It is difficult to adopt the same strategy for arrays with many elements. Since the satellites are remote from each other, there is no easy way of simultaneously combining the signals from all of the different elements in an analog form. For these arrays, the combining is easier done in discrete time, after digitizing the signals while preserving the phase.

Thus, the applicability of spacecraft arrays for passive sensing is effectively limited. A passive receiving spacecraft array must record information over a reasonably long period of time to integrate the SNR. This long recording period then necessitates enormous storage capacity on board each satellite, since all the phase information must be preserved. Sampling the carrier wave at the Nyquist limit with 8-bit quantization would result in storage rates of 96 Gbps for an X-band detector. Even with high-speed buffers, the required storage capacity after only a few seconds of integration time is prohibitive. Of course, the receiver can filter and mix the input signal down to a lower intermediate frequency (IF) before the analog-to-digital (A/D) conversion, greatly reducing the load on the data processing. This process results in no loss of information, provided the *information bandwidth* is known to be small compared to the carrier frequency. The information-bearing component of the signal is the complex amplitude of the electromagnetic wave. The phase and magnitude of this complex amplitude must be preserved during any signal processing prior to the combining operation. In general, the bandwidth of this complex amplitude may be as high as the receiver bandwidth, disallowing any mixing down to lower frequencies. Sometimes, however, the information signal of a target is known to be band limited over a reasonable range (kilohertz to megahertz). In this case, digitization and storage can still be problematic, but are at least manageable.

Spacecraft arrays are also unsuitable as active transmitters for tracking applications. These spacecraft radars track targets with a narrow beam, optimizing the SNR from the target while nulling the clutter and noise. Correct phasing of the array at the desired angle and range to illuminate the target must be performed in real time. Returns from the target are used by a feedback controller to vary the phase at each element in order to steer the beam. To do this, each array element must have accurate information about the relative position of all other array elements. Continuous communication between satellites is needed. Furthermore, the time constant for the detection (including signal reception, combining, processing, and phasing of the transmissions) must match the dynamics of the target. For small local clusters, the slow dynamics of the array may allow this procedure to be carried out if the processing capability exists. For virtual clusters, this task would be very tricky, given the dynamic nature of the system.

The problem can be avoided by adopting a multistatic architecture in which a single monolithic transmitter illuminates the scene with a wide-area beam. A receiving array of spacecraft then tracks the target from the reflected illumination. The time constant of the detection process is much greater, allowing signals to be combined and processed independently of the transmitter.

The system would then require either several large, localized GEO transmitters, or many moderate-sized transmitters distributed in LEO or medium Earth orbit (MEO). This architecture has no real "deal breakers," and with enough careful design, the technology exists to construct an effective working system.

15.4 Generalized Analysis

The previous sections have introduced some of the reasons that support the use of distributed architectures for space applications and have also identified the areas where distribution can lead to difficulties. Most of the arguments used in these discussions were qualitative. This qualitative approach was necessary to assist understanding and to highlight the important points. Qualitative reasoning is, however, the realm of lawyers and politicians. The engineer must instead rely on thorough quantitative analysis to support any design decisions. This section therefore introduces a method for quantitative analyses of space system architecture design. There are many different analysis methods that can be used in aerospace system engineering. Whereas most methods apply to traditional deployments and specific applications, the methods described here are applicable to almost all unmanned space missions, and can be generalized to both distributed and traditional architectures.

15.4.1 Classifications

For generalized analyses, it is helpful to categorize the different classes of systems considered. Categorizing the classes allows the issues and problems characteristic of each class to be assessed. The analyses can then be conducted in a logical and orderly fashion.

In Sec. 15.1.2, it was pointed out that the level of distribution exhibited by a system is defined by the cluster size. Although the cluster size is the primary form of system categorization, additional classifications are necessary. Specifying a cluster size of 10, for example, indicates nothing about the way that the satellites coordinate to satisfy the local demand, nor does it provide any information about the constellation design. Classifications are therefore necessary that identify the system and define the important system characteristics.

15.4.1.1 Distribution

The first level of system classification is based on the coordination exhibited by the system elements.

- **Collaborative.** Each separate satellite operates independently and is able to isolate signals to the required resolution and transfer information at the correct rate and integrity. The cluster size can be as low as unity, but may be more if multiple satellites are needed to support the required availability (e.g., improving coverage statistics, adding redundancy, reducing performance fluctuations). Examples of collaborative systems are commercial distributed imaging systems such as SPOT, GEROS, and Resource 21.[38] These systems feature constellations of several satellites, each capable of recording images with around 10 m resolution. The size of the constellation determines the coverage and revisit time of the system. Distribution improves the functionality of the system by supporting more frequent observations over larger areas.
- **Symbiotic.** The separate satellites cannot operate alone. Rather, they have a symbiotic relationship with the other satellites in the system. No single satellite can isolate the signals or transfer information at the correct rate or integrity. Only by the coordinated operation of several elements can the system perform the design function. The cluster size of symbiotic systems must be greater than unity. An example of a symbiotic system is the spacecraft array for communications and sensing. Another example is the proposed separated spacecraft interferometer.

- **General.** The system cannot be categorized as either purely collaborative or symbiotic, but instead as a combination of the two classes. Some level of coordination is required between system elements, although each element has high-level functionality. Individual satellites can isolate the signals and can transfer information at the correct rate or at the correct integrity, but not both. The cluster size for general systems must be greater than unity. Examples of a general system include the proposed broadband communication systems like Teledesic; each satellite can support local point-to-point communications, but relies on the constellation for connectivity across the network. Another example is passive imaging for military reconnaissance. Each satellite can record images with high resolution; however, only by using data fusion between satellites or collateral information from other sources can the images be interpreted with a high degree of confidence.

15.4.1.2 Architectural

The second level of system classification specifies the level of homogeneity exhibited by the system architecture.

- **Constellations.** These are systems that feature a large number of similar satellites in inertial orbits, each orbit with its own unique set of parameters. Walker Delta patterns or Molniya orbits are possible types of constellations. Systems such as Teledesic, GPS, and Iridium are characterized as being constellations. Virtual clusters, with sizes greater than unity, can be formed if the constellation supports multiple coverage of target regions.
- **Clustellations.** Some proposed systems involve local clusters of spacecraft that orbit in close proximity. The clusters can be made up of formation-flying satellites or can even involve the physical tethering of satellites. The system may involve more than one cluster. An architecture that utilizes local clusters is classified as a clustellation, since it features a constellation of clusters.
- **Augmented.** An augmented system has a hybrid architecture featuring primary and adjunct dissimilar components that perform different subsets of the mission. The system is designed so that the combined performance of all components satisfies the overall mission objective. An example of an augmented system would be the combined use of different platforms or sensors to perform active and passive surveillance. Within this analysis framework, the Space Based Infrared Systems (SBIRS), SBIRS Low and SBIRS High, are collectively classified as augmented. Another example of an augmented system is the proposed concept of using both uninhabited Ariel vehicles (UAVs) and space assets for tactical reconnaissance of a battlefield.

15.4.1.3 Operational

The third level of classification groups systems according to their operational characteristics. This is the most abstract type of classification. The following list is by no means exhaustive and covers only several examples of the operational classifications.

- **Active or passive sensing.** Remote sensing may be active or passive, with marked differences in capability and cost. These differences are primarily due to the additional power requirements and the r^4 scaling associated with active systems.
- **Track, search, or imaging.** Tracking targets using staring sensors scales differently from searching for targets with scanning sensors. The detailed imaging of a static scene differs from either tracking or searching.

15.4.2 Development of Metrics

Satellite systems can be designed to perform essentially the same task in many different ways. To compare alternate designs, a metric is required that fairly judges the performance of the different systems in carrying out the required task. In today's economic climate, there is also a requirement to consider the monetary cost associated with different levels of performance. Because of the extremely large capital investment required for any space venture, satellite designers must provide the customer with the best value. For a distributed system to make sense compared with another kind of system, it must offer reduced cost for similar levels of performance. The importance of cost indicates the benefit of a definable cost per performance metric. Performance and cost metrics can be used as design tools by addressing the sensitivity in performance and cost to changes in the system components, or by identifying the key technology drivers. This leads to the definition of the adaptability metric, which quantifiably measures the sensitivity to changes in the design or role of the system. The last metric used for the generalized analysis is the capability metric, which assesses the potential capabilities of the system.

- **Cost per performance.** The cost per performance metric is a measure of the cost to provide a common level of performance and includes expected development, launch, and operations costs, with a special consideration for contingencies within the system lifetime. (One example of such a contingency cost was that associated with the shortened lifetime of a U.S. Defense Satellite Communications System [DSCS] satellite that had to be rephased during Operation Desert Storm.)
- **Adaptability.** The adaptability metric judges how flexible a system is to changes in its required role, component technologies, or operational procedure. The adaptability offered by a distributed network of satellites allows the design function to be changed by augmenting or replacing elements in the constellation or by simply reprogramming the system software. The first issue is that the satellites of a distributed system will be smaller and less complex than large single deployments, allowing replacements to be made at low incremental cost. This means that elements can be replaced to keep in step with the current state of technology, possibly improving performance or reducing costs associated with maintaining obsolete technology. The second issue, concerning adaptability through reprogramming, requires change in the way people think about space systems; on-orbit hardware may be considered a commodity, with the software representing the value added. A good analogy hinting to the benefits offered by this flexible system is the personal microcomputer, which can be updated by changing the processor, by adding expansion cards, or by simply changing the operating system. The question to be answered is whether such a system can be cost effective while maintaining the required level of performance.
- **Capability.** The capability metric is used to assess the potential capabilities of a particular distributed system compared with the traditional single-satellite deployment. The metric also characterizes the value of those systems offering new capabilities that would be impossible to achieve without distribution.

Any metric used for comparative analysis should be quantifiable and unambiguous. A measurable metric therefore requires a formal definition that leads to a calculable expression. Unfortunately, satellite engineering analysis has traditionally been treated on a case-by-case basis. Each new satellite system is designed and judged by its own set of rules for a specific, narrowly defined task. Thus, any formal definition of a metric (for performance or cost) has been specific and relevant only to satellites of the same architecture, be it GEO communications or weather imaging. It is therefore necessary to develop a generalized and formal framework for defining quantifiable metrics for performance and cost, capability, and adaptability.

15.4.3 The Concept of Measurable Performance

Ability to measure the performance of a satellite system is implicit in constructing the metrics described in the previous section. For example, a cost-per-performance metric cannot be calculated unless there is some understanding of what *performance* actually represents. Identifying exactly how performance is measured is therefore the first step in the analysis. This is nontrivial, since previous definitions have been subjective and entirely dependent on the opinions of the engineer. A formal definition is required, independent of function or application, and unambiguous in implication.

Performance is perceived in terms of satisfying a demand of a cooperative or uncooperative market. An example of a cooperative market is a consumer base for communications, and an example of an uncooperative market is the detection of a set of military targets. The demand can be represented by a set of *functional requirements*. Performance should always be defined relative to these requirements. To be unambiguous and quantifiable, performance should represent the likelihood that the system can satisfy the functional requirements, both instantaneously and continuously over its lifetime. In short, *the performance of a system is the probability that the system instantaneously and continuously satisfies the top-level functional requirements that represent the mission objective.*

The mathematical representation of performance follows immediately from this definition. This representation supports the quantitative analyses necessary for construction of the metrics discussed earlier.

It is important to note that performance is distinct from capability, although the two are related. Capability characterizes the level at which a task can be performed, while performance simply measures how well the system performs a task relative to requirements. The definition of capability will follow directly from the construction of the performance metric and is discussed later.

15.4.3.1 The Formulation of the Generalized Performance

The first observation, and primary enabler for a generalized analysis, is that all current and envisioned applications for satellite systems essentially perform the task of collection and dissemination of information. This common thread linking all systems allows a framework for a generalized analysis to be established.

For information collection/dissemination, the performance of the system is always subject to a signal-to-noise constraint, and a detection problem of deciphering the correct signal from noise, clutter, and other (unfriendly) sources. Since the analysis should reflect realistic operational states, the reliability and redundancy of the system should be an intrinsic part of the performance metric. In general, the performance of a system is not deterministic but is statistical. This is due to the random nature of noise sources and the finite probability of failure of system components.

As alluded to in Sec. 15.2, system performance for information disseminators can be defined by a set of four requirements and an associated confidence of compliance.

- **Isolation requirement.** The system must be capable of acquiring, isolating, and identifying information-bearing signals from sources in the coverage region. The ability of the system to acquire and identify individual signals must exceed the isolation requirement S^*. For an imaging system, this sets the requirements that the resolution must correspond to the spatial separation of the signal sources and that the pixels must be accurately mapped to the object scene.
- **Rate requirement.** The system must be capable of satisfying a resolution or rate requirement R^*. This determines the granularity (symbol interval) of the information being transferred. For a communication system, this places a requirement on the mean information data rate. For a search radar, this requirement is represented as the mean time for detection of a target.

- **Integrity requirement.** The integrity requirement I^* represents the error control of the system. The information transferred through the system must satisfy this integrity requirement. For communication systems, this criteria defines the acceptable bit error rate. For the search radar, the requirement on the false alarm rate is the corresponding integrity constraint.
- **Availability requirement.** Availability is defined as the probability that the system is operational at any particular instant in time. The system is said to be operational only if it satisfies the isolation, rate, and integrity requirements. The availability of the system is therefore a measure of the variance of the isolation, rate, and integrity supported by the system. The isolation, rate, and integrity can vary due to component failures, viewing angle and coverage variations, signal attenuation, or simple statistical fluctuations. The availability requirement A^* defines the minimum acceptable operational probability. For instance, a communication system may be subject to a requirement that it can support multiple users at a certain data rate at a given bit error rate, with a 90% probability.
- **Instantaneous performance $c_p(t)$.** This is the instantaneous probability of compliant operation. The instantaneous performance at some time t is the likelihood that the system simultaneously satisfies the three requirements described previously. In fact, because the other requirements are embedded in the definition of availability, the instantaneous performance is simply the probability that the system availability exceeds A^*.

The availability requirement essentially specifies the minimum amount of redundancy (or reliability) required of the system. The instantaneous performance measures the amount of redundancy exhibited by the system over and above this minimum. Before proceeding, it is valuable to solidify the understanding of these definitions with a real example. For this, we return to the availability considerations of the GPS system, in Example 9.

Example 9. Instantaneous Performance of the GPS-24 Constellation

GPS navigation requires measurements to at least four satellites. The possibility of satellite failure, or the likelihood of poor coverage geometry, requires that more satellites than this minimum be in view at any time. To provide acceptable availability, the constellation was therefore designed to have a minimum of five satellites in view above 5° elevation.[8] This number of satellites corresponds to an availability requirement that the Position Dilution of Precision (PDOP) be no greater than six. In most cases, the GPS-24 system achieves far higher visibility than this requirement, and there are no PDOP outages. There is, however, a finite probability that one of the 24 satellites will fail. Should this occur, there are several regions in the mid and high latitudes where the PDOP constraint is violated, and the availability of service falls below requirements. Outages are short lived, typically on the order of 5 min, with a maximum of 24 min per day. These service outages do, however, constitute a failure within the definitions of the availability requirement. The instantaneous performance of the GPS-24 system is therefore the probability that no such failure occurs.

The calculation of $c_p(t)$ can be carried out to any desired level of detail. The system is first represented as a state vector of its components (e.g., number of satellites, power per satellite, aperture size). Defining failure as noncompliance of the availability requirement, the *failure states* are the set of all possible state vectors corresponding to the A^* requirement boundary. By considering all possible transitional paths from the current state to any of the failure states, the probability of failure can be calculated. This calculation involves a Markov chain analysis. The complexity of this

calculation grows exponentially with the number of elements in the state vector. For this reason, the smallest possible state vector, containing only the critical system components, should be used.

In general, there may be different rate, integrity, and availability requirements placed on a satellite system at different times within its life, or at different demand locations within its region of coverage. For example, the mean-time-to-detection requirement for a search radar system may be a spatially varying function; long detection times can be tolerated in regions where there is little or no threat, while short detection times are required around air bases. The system must satisfy the requirements specific to each of the demand locations. The performance $c_p(t)$ of the system at some time (t) is given by the worst-case probability of compliance, over all demand locations within the coverage region. All users will experience at least this minimum level of performance. Defining $x \in X$ as the set of source/sink pairs that constitute the demand, this calculation can be done for each year over the lifetime of the satellite to give the performance profile:

$$c_p(t) = \min[p(Ax, t \geq A^*(x,t) | S^*(x,t), R^*(x,t), I^*(x,t)], x \in X$$

Typically, the availability of a satellite system will change over its lifetime. This is because all system components have finite failure probabilities that generally increase in time. Once on orbit, a satellite system is difficult to repair. A system must be able to operate within acceptable availability bounds throughout its intended life. System reliability is closely related to availability. Reliability refers to the probability of continuous compliance of the availability requirement.

- **System reliability** (at time t) is the conditional probability that the system has been operational over the interval $\{0, t\}$, given that it was operational at an initial time $t = 0$. Note that this is also a functional definition: a system is deemed operational if it can support the requirements on information isolation, rate, integrity, and availability.

Satellite engineers usually favor reliability as a figure of merit over availability. In many cases, the end-of-life (EOL) performance characteristics of the system are of primary importance. After all, the system lifetime is by definition the duration of time over which the system should satisfy requirements. Using the reliability definition, evaluated at the EOL, allows us to form the generalized performance metric.

- **Generalized performance metric C_p.** This metric is the probability that the system has been operational continuously from an initial time to EOL. The metric reflects the likelihood that the system has been able to continuously satisfy the isolation, rate, integrity, and availability requirements described previously. In this way, the generalized performance is a quantitative measure of how well the system satisfies the functional requirements.

The relationship between instantaneous performance and the generalized performance is analogous to the conventional definitions of availability and reliability. The generalized performance metric C_p, relative to a set of requirements (S^*, R^*, I^*, A^*), is therefore of the same form as the instantaneous performance, but with reliabilities used in the calculations in place of availabilities.

Note that a system that is unable to meet requirements will have $C_p = 0$. This value reflects the fact that such systems are not considered as viable candidates. This stresses the importance of setting requirements carefully, ensuring that they properly reflect the expected demand.

15.4.3.2 The Cost-per-Performance Metric

The impact of improved performance on the cost of a system must be determined in order to calculate the cost-per-performance metric. If the value of performance can be quantified, the system cost can be modified to correspond to a common level of performance. The modified system cost should represent the total lifetime cost of a system, where lifetime cost is defined to be the total expenditure necessary to continuously satisfy the top-level system requirements. In this way,

alternate systems with different levels of performance can be fairly compared on the basis of the total lifetime cost.

The *baseline system cost* C_S accounts for the design, construction, launch, and operation of the system components. This baseline cost does not, however, account for the expected cost of failures of system components. Since the system must satisfy requirements throughout its design life, expenditure will be necessary to compensate for any failures that cause a violation of this condition. These additional *failure compensation costs* V_f must be added to the baseline system cost to give the total lifetime cost C_L.[39]

$$C_L = C_S + V_f$$

To calculate the baseline system cost, a cost operator is applied to the system state vector and associated component reliabilities. As long as the operator is used consistently, any parametric cost model could feasibly be used for this calculation. As discussed earlier, however, care must be taken in applying the SSCM, since a distributed system of small satellites is not necessarily a small system. Note that a premium is paid for more reliable components.

Since some costs are incurred at different times within the lifetime of the system, the cost is actually represented as a cost profile. This profile has to be modified to account for the time value of money. Costs incurred later in the system lifetime have a lesser impact on the overall system cost. A dollar is always worth more today than it is tomorrow; capital expenditure can earn interest if invested elsewhere. The yearly costs are therefore discounted according to an assumed discount rate corresponding to an acceptable internal rate of return (IRR). In order to attract investors to commercial systems, the high risk associated with space ventures necessitates a high IRR of approximately 30%.[40]

The discounted cost profile $c_s(t)$ must then be integrated over the system lifetime to obtain the total baseline system cost C_S:

$$C_S = \sum_{life} c_S(t)$$

The failure compensation costs V_f can be estimated from an expected value calculation:

$$V_f = E[V_f] = FV_S \ ,$$

where F is the combined probability of failures that cause a loss of availability, and V_S is the sum of the economic resources required to compensate for the failures. Note that V_S includes the costs of replacement satellites or components, launch costs, and any opportunity costs representing the revenue lost during the downtime of the system. The calculation of V_S is architecture specific and in most cases depends strongly on the nature of the most likely failure mode. A failure mode and effects analysis (FMEA) may be required to estimate the replacement costs. Estimation of the opportunity costs is difficult, requiring a prediction of the nature of the failure, the time the failure most likely occurs, and its duration. Despite these problems, V_S can be estimated with reasonable confidence using predictive methods for simple systems and Monte Carlo simulations for more complex systems.

By definition, generalized performance represents the probability that the system continuously satisfies the requirements. The complement of the performance is the probability of failure, F:

$$F = 1 - C_p$$

It is therefore through the failure compensation costs that performance has an impact on system lifetime costs. A higher performance system will have a lower probability of failure and consequently a lower expected value of the compensation costs.

This gives the lifetime system cost:

$$C_L = C_s + (1 - C_P)V_s$$

Although the lifetime costs are a valid measure of the cost per performance of a system, it is useful to also include in the metric a comparison to the expected demand. This step is called demand matching and is necessary because a system cannot outperform the demand. Extra capacity beyond the market demand brings no additional revenue or benefit, but may incur increased costs.

The product of the required values of rate, integrity, and availability gives the nominal error-free rate of information transferred through the system for each source/sink pair. Comparing this error-free rate to the size of the local demand, and taking the minimum, gives the achievable capacity of the system. This capacity is defined as the *market capture* of the system. Since the size of the local demand Q is almost always time and spatially varying, the demand-matching calculation involves an integration over the entire coverage region for each year of the satellite lifetime, to give a market capture profile $m_c(t)$:

$$m_c(t) = \sum_{x \in X} \text{Min}[(R^*(x,t)I^*(x,t)A^*(x,t), Q(x,t)]$$

A further complication arises if the system operation results in monetary income, as in commercial communication systems. In this situation, the time value of money means that there is also a bias in the relative "value" of market capture, with a weighting toward the start of the system's lifetime. In general, revenue should be earned as close as possible to the time that the associated costs are incurred. For example, revenue earned from the transmission of bits early in the life of a communication satellite is more important than revenue earned late in the satellite's lifetime. For this reason, for each year of the lifetime of the satellite, the capture profile $m_c(t)$ must also be correctly discounted, according to the same discount rate as was used to discount the costs.

The total system capture M_c is then calculated by summing the discounted capture profile over the entire lifetime of the system:

$$M_c = \sum_{life} m_c(t)$$

Having determined the total system capture, the cost-per-performance metric CPP can now be calculated:

$$CPP = \frac{C_L}{M_c}$$

The units of the cost-per-performance metric are dollars per information symbol.

Part of the utility of this cost-per-performance metric is that it permits comparative analysis between different systems with large architectural differences, scaling their cost according to their performance and market capture. Very large and ambitious systems can be fairly compared to smaller, more conservative systems. The cost-per-performance metric can also be used to assess the potential benefits of incorporating new technology in spacecraft designs. New technology should only be included in the design of a new satellite if it can offer reduced cost or improved performance. This can be evaluated with the metric, provided that both the cost and the expected reliability of the new technology can be estimated. Commonly, the largest problem encountered with incorporating new technology in space programs is schedule slip. This slip can have an adverse effect on the overall success of the program, extending the period of capital expenditure, while delaying operations that bring revenue. These effects can also be captured by the cost-per-performance metric. Some typical amount of program slip can be included in the cost profile $c_s(t)$,

and the corresponding delay can be applied to the performance profile. The combined effects of including the new technology will then be apparent, by comparing the cost-per-performance metric to costs corresponding to designs featuring more established technologies.

Example 10. Distributed IR Remote Sensing for Forest Fire Warning

To demonstrate the application of the cost-per-performance metric for a distributed system, we will consider the design of a system to detect forest fires. This choice is deliberate. The excellent text on space systems engineering, *Space Mission Analysis and Design*, by Larson and Wertz,[22] demonstrates the system engineering process using the conceptual design of the fictitious FireSat satellite as a case study. By adopting the same list of requirements as in the case study, a comparison can be made between the distributed system described in the study and the more traditional FireSat. Drawing from the referenced text, the mission objective is "*to detect, identify, and monitor forest fires throughout the United States, including Alaska and Hawaii, in near real time.*"

This mission statement can be translated into a basic set of functional and operational requirements.

- Thermal imaging with four temperature levels and 30 m resolution
- Coverage of continental United States
- Response time (detection and data delivery) of 1 h
- Ninety-eight percent availability
- Ten-year mission life

The requirements chosen for this example differ slightly from those of the quoted text.[23] The main difference is in the response-time requirement: the text specifies daily coverage of the continental United States, with a 30-min data-delivery time. This requirement seems contradictory to the mission statement; a revisit time of 1 day is simply too long for adequate fire-fighting response. Since the mission statement specifies near real-time detection, it is assumed that a fast response time (detection and reporting) is a driving requirement. For this example, choosing a total response time requirement of 1 h seems reasonable and permits a comparison of small constellations and large distributed constellations. If a faster response time were chosen, only large constellations could satisfy the requirement, and no comparisons could be made.

These requirements represent the minimum levels of performance of a viable system. In terms of the definitions used in this section, the isolation requirement specifies that the resolution of each pixel be 30 m. The information symbol size is 2 bits, to achieve four temperature levels. The rate requirement for each of the 30-m pixels in the United States is that the pixels be sampled and reported every hour. The availability requirement is 98% throughout the U.S. coverage region. Note that there is no integrity requirement specified. This requirement was omitted to simplify the analysis. If present, such a requirement would specify the maximum probability of assigning the wrong temperature to a 30-m resolution cell.

Consider a constellation of LEO satellites to carry out the mission. To detect forest fires, sensing is best performed in the midwavelength (3–5 μm) IR region. We therefore return to the system described in Examples 6 and 8. The performance characteristics specified in those examples were:

- Global coverage
- Revisit time = 25 min
- Data delivery time = 5 min

- Sixteen temperature levels

By using these elevated performance characteristics, candidate systems will have a built-in redundancy for the forest-fire warning mission. From the parameters given in Tables 15.1 and 15.2, we can define a feasible, low-cost system. Calling our system *Tinderbox*,* the system parameters are:

- Fifty satellites, at 400 km altitude
- Spacecraft dry mass ~ 20 kg
- Spacecraft wet mass ~ 28 kg (assuming 300-s Isp thrusters and ΔV~ 100 m/s per year for orbit maintenance)
- Aperture = 6 cm
- Forty-three ground stations
- Maximum data storage requirement = 0.3 Gbits
- Storage capacity = 1Gbit
- Downlink rate = 1.25 Mbit/s

Assume that each satellite has a constant reliability of 0.985 over the 10-year lifetime, and that each groundstation-satellite link has a constant availability of 0.9. We proceed to calculate the performance of Tinderbox, relative to the requirements described above.

The resolution of 30 m matches the requirement, and it can be assumed that this requirement will be satisfied throughout the lifetime of the system. Under an assumption of perfect pointing, the system therefore has a 100% continuous probability of satisfying the isolation requirement. As a result, the availability (and in this case, reliability) is just the probability of satisfying the rate requirement; recall we are neglecting integrity.†

The revisit time of 25 min supported by this system means that at least two satellites pass overhead every 50 min. Furthermore, each satellite that passes overhead has multiple opportunities to download the data. The first satellite (A) that overflies has three opportunities to download before its storage capacity is exceeded. The second satellite (B) has two download opportunities before the maximum response time is violated.

The availability is then:

$A = 1 - P$ (both sats fail)
$- P$ (both sats work but all 5 ground links fail)
$- P$ (sat A fails and sat B misses 2 links)
$- P$ (sat A misses 3 links and sat B fails)

Since it is assumed that satellite failures are independent, the availability is easily calculated:

$$A = 1 - (0.015^2) - (0.985^2 \times 0.1^5) - (0.015 \times 0.098 \times 0.1^2) - (0.985 \times 0.015 \times 0.1^3) = 0.9996s$$

---

* The name "Tinderbox" was chosen not only because of the flammable connotations of the word, but also because of the fact that Tin is the 50th element of the Periodic table, thereby providing a small tribute to the forerunner of commercial distributed space systems.

† Note that Tinderbox supports four times the temperature granularity of the baseline requirement, implying that the achievable integrity would be very high. In order to fail requirements (based on only four temperature levels), a pixel would have to be labeled with a temperature that is four levels higher than the true value.

This performance clearly exceeds the availability requirements. However, the generalized performance is the probability of the system degrading to a state such that the availability requirement is violated. To evaluate this performance, we must consider the effects of all the different combinations of satellite and link failures to determine which of them constitute a *failure state*.

Consider first the availability of the system after the failure of a single satellite:

$$A_{1failed} = 1 - P(\text{sat A fails} \mid \text{sat B failed}) - P(\text{sat B fails} \mid \text{sat A failed}) - P(\text{sat A works, misses 3 links} \mid \text{sat B failed}) - P(\text{sat B works, misses 2 links} \mid \text{sat A failed})$$

Assuming the satellite failures are independent, there is an equal probability that the failed satellite is A or B. Therefore, this calculation becomes:

$$A_{1failed} = 1 - (0.5 \times 0.015) - (0.5 \times 0.015) - (0.5 \times 0.985 \times 0.1^3) - (0.5 \times 0.985 \times 0.1^2) = 0.980$$

Thus, a single satellite failure is not sufficient to cause a violation of the availability requirement. Failures must occur on two adjacent satellites for the availability to drop below 0.98. The probability of failures in two or more satellites in the constellation is:

$$P(\text{two or more failures}) = 1 - P(\text{none fail}) - P(1 \text{ fails}) = 1 - 0.985^{50} - 50 \times 0.985^{49} \times 0.015 = 0.1726$$

There is no guarantee that these two satellites will ever consecutively overfly a location. Nevertheless, this calculation provides an upper bound on the probability of failing the availability requirement. Therefore, the generalized performance metric has a lower bound:

$$C_p > 1 - 0.1726 > 0.827$$

This is the performance of the Tinderbox system in satisfying the functional requirements described earlier. The cost per performance can be calculated from the same cost estimation techniques used in Example 6 (Fig. 15.3) to give:

- Satellite hardware cost ~ $25M
- Launch cost ~ $31M (assuming $10K/lb = $22K/kg)
- Total baseline cost ~ $56M
- Replacement cost for a pair of satellites ~ $2M ($750K for hardware + $1.25M for launch)

The lifetime cost is then:

$$C_L = C_s + (1 - C_P)V_s + \$56\text{M} + 0.1726 \times \$2\text{M} = \$56.3\text{M}$$

The demand-matching and market-capture calculations are trivial, since they simply represent the number of pixels imaged by the system over its lifetime. During each revisit cycle, all of the pixels on the Earth are imaged. For 30 m resolution, this amounts to ~ 5.7×10^{11} pixels. These pixels are imaged and delivered within 1 h. For a 10-year lifetime, the total number of pixels imaged is ~ 5×10^{16}.

The cost per performance is therefore CPP = $1.13/gigapixels!

This measure allows us to judge the relative merits of Tinderbox compared with those of other architectures, such as the FireSat system. The referenced text never actu-

ally formulates a final design, so it has been assumed that FireSat features three satellites at 800 km altitude. The satellite reliability and link availability are assumed to be the same as those of Tinderbox. Assuming a revisit time of 1 h for FireSat results in a low availability of 0.7, since all elements must work in series. This value is below the required availability. In order to improve the availability to satisfy requirements, a great deal of expenditure would be necessary. However, using the same cost model assumptions as before, we see from Fig. 15.4 that the cost of the 800-km, three-satellite configuration corresponding to Firesat is already expensive at ~$65M for the hardware alone. There seems little reason to consider such a costly option, when the distributed Tinderbox offers a cheap and capable alternative.

15.4.3.3 The Capability and Adaptability Metrics

The analysis methodology is a work in progress,[3] and as a result, the exact formulations of the adaptability and capability metrics are incomplete. The details given here represent the current stage of development.

First derivatives of the cost and performance metrics, with respect to state elements, yield the sensitivities to changes in system components. This is exactly the adaptability metric discussed earlier. Second derivatives identify the key technology drivers, since these derivatives determine the component that gives the largest change in the gradient of the performance with small changes in state.

The derivatives of the cost and performance metrics must be constrained by the extent of possible increments in the different state components. There are physical, political, financial, and risk constraints to consider. For example, vast increases in the power available to satellites are possible with the use of nuclear technology, but this technology is unlikely to be considered feasible due to policy and financial constraints.

The capability of the system is defined as the quadruplet of isolation, rate, integrity, and availability (S, R, I, A) that can be achieved with the satellite system. This metric is *not* measured relative to requirements but reflects the possible operating conditions of the system. The capability metric is somewhat ambiguous, since any given system may be able to support several different (S,R,I,A) quads; the metric is nonunique. This lack of uniqueness results because the calculated availability is strongly dependent on the isolation, rate, and integrity quantities. The capability metric is therefore a family of characteristics in (S,R,I,A) space. The capability metric is useful for assessing the potential of distributed systems compared to single-satellite deployments, since it identifies where the largest benefits can be exploited.

15.5 Conclusions

The possible utility of small satellites and microsatellites is somewhat limited by their characteristically modest payload resources (power and aperture). Distributed architectures are enabling for small satellite designs by expanding their useful range of applications to include high rate and resolution sensing and communications. The capabilities of many small satellites are combined to satisfy mission requirements.

A distributed architecture makes sense if it can offer reduced cost or improved performance. Functional requirements specify minimum levels of acceptable performance and include resolution, rate, integrity, and availability requirements. Viable systems must satisfy these requirements throughout their lifetime. Compensation must be made following failures that cause a violation of requirements. "Improved performance" thus relates to the ability of the system to satisfy requirements with a higher probability. "Reduced cost" corresponds to lower lifetime costs that

include the expected failure compensation costs. Because the performance requirements, and the associated probability of satisfying them, are embedded in the lifetime cost calculation, it is a useful metric for architecture analysis. This calculation leads to the definition of a cost-per-performance metric, supporting quantitative analysis of distributed satellite systems. The evaluation of this metric allows alternate architectures to be compared and judged, and can demonstrate the effects of incorporating new technology into satellite programs.

Distribution can offer improvements in isolation (resolution), rate, integrity, and availability. The improvements are not all-encompassing, and in many cases are application specific. Nevertheless, it appears that adopting a distributed architecture can result in substantial gains compared with traditional deployments. Some of the major advantages that distribution may offer are:

- Improved resolution, corresponding to the large baselines that are possible with widely separated antennas on separate spacecraft within a cluster
- Higher net rate of information transfer, achieved by combining the capacities of several satellites in order to satisfy the local and global demand
- Improved availability through redundancy and path diversity. Frequently, the cost of adding a given level of redundancy is less for a distributed architecture
- Improved availability through a reduced variance in the coverage of target regions. This variance reduces the need to "overdesign" and provides more opportunities for a favorable viewing geometry
- Lower failure compensation costs, due to the separation of important system components among many satellites; only components that break need replacement
- Permits new missions that cannot be done otherwise

There are some problems, specific to distributed systems of small satellites, that must be solved before the potential of distributed architectures can be fully exploited. The most notable of these problems are:

- An increase of system complexity, leading to long development time and high costs
- Inadequacy of the data storage capacity that can be supported by the modest small satellite bus resources
- Difficulty in maintaining signal coherence among the apertures of separated spacecraft arrays, especially when the resolution requirements are high or the target is highly dynamic

The resolution of these problems, and the proliferation of microtechnology, could lead toward a drastic change in the satellite industry. It seems clear that distribution offers a viable and attractive alternative for some missions. Large constellations of hundreds or thousands of small satellites and microsatellites could feasibly perform almost all the missions currently being carried out by traditional satellites. For some of those missions, the utility and suitability of distributed systems look very promising. More analysis is warranted in order to completely answer the question of where and when distribution is best applied, but the potential prospects of huge cost savings and improvements in performance are impossible to ignore. It therefore seems inevitable that massively distributed satellite systems will be developed in both the commercial and military sectors. We are living in a time of great changes, and the space industry has not escaped. Over the last few years, "faster, cheaper, better" has been the battle cry of those engineers and administrators trying to instigate changes to improve the industry. "Smaller, modular, distributed" may be their next verse.

15.6 References

1. N. Lynch, *Distributed Algorithms*, prepress ed. (Morgan Kaufmann Publishers, 1995).
2. *System Reliability and Integrity* (Infotech International Limited, 1978).
3. G. B. Shaw, "Generalized Analysis of Distributed Satellite Systems," Ph.D. thesis, Dept. of Aeronautics and Astronautics, MIT, 1997.
4. GeoMobile Program Office, Hughes Space and Communications Company, El Segundo, CA (1997).
5. D. P. Wickert, "Space Based Radar–System Architecture Design and Optimization for a Space based Replacement to AWACS," Masters thesis, Dept. of Aeronautics and Astronautics, MIT, June 1997.
6. E. A. Lee and D. G. Messerschmitt, *Digital Communication*, 2nd ed. (Kluwer Academic Publishers, 1994).
7. "Technology Development for Separated Spacecraft Interferometers," Proposal for new millennium technology for instruments and microelectromechanical systems (MEMS) IPDT, Jet Propulsion Laboratory, Pasadena, CA, June 1995.
8. "Global Positioning System: Theory and Applications," in B.W. Parkinson and J. J. Spilker, Jr., eds., Vol. 1. *Progress in Astronautics and Aeronautics*, Vol. 163 (AIAA, Inc., 1996).
9. *Space Division Unmanned Spacecraft Cost Model.* 5th ed., USAF Space Division, Directorate of Cost Analysis, El Segundo, CA.
10. *Space Division Unmanned Spacecraft Cost Model.* 6th ed., USAF Space Division, Directorate of Cost Analysis, El Segundo, CA.
11. D. A. Beardon, "Cost Modeling," in J.R. Wertz and W.J. Larson, eds., *Reducing Space Mission Cost* (Microcosm Press, 1996).
12. J. Sellers and E. Milton, "Technology for Reduced Cost Missions," in Wertz and Larson, *Reducing Space Mission Cost.*
13. J. R. Wertz, "Implementation Strategies and Problems," in Wertz and Larson, *Reducing Space Mission Cost.*
14. J. R. Wertz, "Radical Cost Reduction Methods," in Wertz and Larson, *Reducing Space Mission Cost.*
15. R. Parkinson, *Introduction and Methodology of Space Cost Engineering*, AIAA Short Course (28–30 April 1993).
16. R. Fleeter, "Design of Low-Cost Spacecraft," in W. J. Larson and J. R. Wertz, eds., *Space Mission Analysis and Design*, 2nd ed. (Microcosm Press, Torrance, CA, 1993).
17. C. Macquiddy (private communication with Prof. D. Hastings, Sunnyvale, CA, 4 Oct. 1996).
18. B. Francois (private communication with G. Shaw, Raytheon, MA, 28 February 1995).
19. G. Canavan and E. Teller, "Low-Level Satellites Expand Distributed Remote Sensing," *Signal* (August 1991).
20. G. Canavan, D. Thompson, and I. Bekey, "Distributed Space Systems," *New World Vistas, Air and Space Power for the 21st Century* (1996).
21. R. F. Brodsky, "Defining and Sizing Payloads," in Larson and Wertz, *Space Mission Analysis and Design.*
22. M. Socha, P. Cappiello, R. Metzinger, D. Nokes, C. Tung, and M. Stanley, "Development of a Small Satellite for Precision Pointing Applications," (internal memorandum, Charles Strark Draper Laboratory, August 1996).
23. Larson and Wertz, *Space Mission Analysis and Design.*
24. "Project Foresight," 16.89 Space Systems Engineering class final report, Dept. of Aeronautics and Astronautics, MIT, Spring 1997.
25. R. Schwarz, "A Probabilistic Model of Satellite System Automation on Life Cycle Costs and System Availability," Masters thesis, Dept. of Aeronautics and Astronautics, MIT, June 1997.
26. E. S. Dutton, "Effects of Knowledge Reuse on the Spacecraft Development Process," Masters thesis, Dept. of Aeronautics and Astronautics, MIT, June 1997.
27. B. Preston, *Plowshares and Power, The Military Use of Civil Space* (National Defense University Press, 1994).

28. R. Congor, Microcosm Autonomous Navigation System (MANS) (Microcosm Press, Torrance, CA, 1995).
29. *Ibid.*
30. J. Tandler, "Automating the Operations of the ORBCOMM Constellation," *10th Annual AIAA/USU Conference on Small Satellites Proceedings* (Utah, September 1996).
31. R. S. Hornstein, "On-Board Autonomous Systems: Cost Remedy for Small Satellites or Sacred Cow?" *46th International Astronautical Congress* (Oslo, Norway, 2–6 October 1995).
32. J. T. Collins, S. Dawson, and J. R. Wertz. "Autonomous Constellation Maintenance System," *10th Annual AIAA/USU Conference on Small Satellites* (Utah, September 1996).
33. T. Sheridan, *Telerobotics, Automation, and Human Supervisory Control* (MIT Press, Cambridge, MA, 1992).
34. A. J. Weiner, D. A. Thurman, and C. M. Mitchel, "Applying Case-based Reasoning to Aid Fault Management in Supervisory Control," *Proceedings 1993 IEEE International Conference for Systems, Man, and Cybernetics* (Vancouver, B.C., 1995).
35. G. J. Yashko and D. E. Hastings, "Analysis of Thruster Requirements and Capabilities for Local Satellite Clusters," *10th Annual AIAA/USU Conference on Small Satellites* (Utah, September 1996).
36. C. Swift and D. Levine, "Terrestrial Sensing with Synthetic Aperture Radiometers," in *IEEE MTT-S International Microwave Symposium Digest* (1991).
37. B. D. Steinberg, *Principles of Aperture and Array System Design* (John Wiley and Sons, 1976).
38. *The Satellite Remote Sensing Industry*, KPMG Peat Marwick LLP (1996).
39. H. Hecht, "Reliability During Space Mission Concept Exporation," in Larson and Wertz, *Space Mission Analysis and Design.*
40. R. Lovell, "The Design Trade Process," Lecture notes from MIT 16.89 Space Systems Engineering class, MIT, Cambridge, MA, February 1995.

16

Propellants for Microspacecraft

S. L. Rodgers,* P. G. Carrick,* and M. R. Berman†

16.1 Introduction

Microsatellite concepts, with total satellite mass between 0.1 and 20 kg, impose a new set of constraints on the selection of propellants and propellant systems to be used on these spacecraft. The wide range of missions considered for these devices also requires a broad look at the many available propellant systems and possible development of novel systems to meet the unique mission requirements. Storage and handling concerns over the life of the mission, electrical requirements, and mission scenarios all influence the choice of propellants. For example, in the area of chemical propellants, the operational simplicity of monopropellants must be weighed against the greater performance of some bipropellant systems.

The physical and chemical properties of many conventional and newly developed chemical propellants are presented in this chapter. The aim of developing mass-produced microsatellites with the fuel as an integrated component adds processability as a major new factor to be considered in the selection of propellants. This new factor may force the consideration of entirely new propulsion schemes as well as the development of novel propellants.

16.2 Propellant Fundamentals[1–3]

Propulsion of spacecraft is generally derived from the propulsive force of an expanding gas. The ejection of gas imparts momentum to the spacecraft, causing a corresponding change in velocity. The momentum change is directly proportional to the momentum exchange of the expelled propellant. The overall momentum of the ejected propellant is also proportional to the temperature of the mass that is expelled, where higher temperatures result in greater momentum. The temperature of the exhaust can be derived from a variety of sources, such as chemical combustion or electric power.

16.3 Calculation of Theoretical Performance

The theoretical performance of rocket propellants is used in the design of new propulsion systems. Theoretical performance is usually expressed in terms of the ideal specific impulse (I_{sp}), which is the thrust (in pounds) for a flow rate of a pound of propellant per second (pounds of thrust/pounds per second of propellant), defined[3,4] as:

$$I_{sp} = T/\Delta w \tag{16.1}$$

where T is the thrust in pounds of force and Δw is the weight flow rate of the propellants in pounds per second, which can be found from the mass flow rate (Δm) by multiplication by the gravitational constant g (32.174 ft/s^2). Specific impulse values, either ideal or measured, are generally reported in units of seconds.

*Propulsion Directorate, Air Force Research Laboratory

†Air Force Office of Scientific Research, Air Force Research Laboratory

The calculation of the specific impulse for chemical propellant combinations can be conducted using a variety of computer codes, including the Air Force Astronautics Laboratory (AFAL, now Air Force Research Laboratory) Theoretical I_{sp} Program, Micro Version.[5] This program requires input data consisting of the molar or weight percentage of each propellant ingredient, density of the propellant, standard heat of formation of the propellant ingredients, rocket chamber pressure, and exhaust pressure or the expansion ratio. The fraction of any propellant ingredient in the system and the temperature of the combustion chamber, throat, and exhaust can be varied to maximize I_{sp} for the entire system. The rocket chamber pressure and exhaust pressure can be fixed to specific values to give a consistent comparison among the various propellant combinations. These values are generally calculated at 1000 psi and 14.696 psi, respectively, to give "standard" sea-level specific impulse values. Typical vacuum specific impulse comparisons are calculated using 1000-psi chamber pressure with vacuum expansion and an expansion coefficient (given by the symbol ε) of 40. The design of small propulsion systems for microsatellites will almost certainly differ from the standard conditions listed here. However, these equilibrium-based theoretical calculations can be used for modeling the wide range of combustion chamber pressures and temperatures that might be found in microsatellites.

The heats of formation of the various propellants can generally be found in the *JANAF Thermochemical Tables*.[6] A few of the JANAF heat-of-formation values vary considerably from the more recent measurements. For example, the JANAF value for MgH is 67 kJ/mol lower than recent literature values. However, most of the values are generally accurate. New synthetic propellants are generally not listed in these tables and therefore must be determined experimentally or theoretically in order to evaluate their potential usefulness as rocket propellants.

Using this I_{sp} program, the calculations were tested by comparing with the previously reported I_{sp} values. The specific impulse values calculated by this program should only be used for comparison between like systems and not considered as "absolute."

The design of rocket propulsion systems generally uses the measured specific impulse of engines and motors that have undergone extensive test firings. Such test data are used to determine the performance losses in the same design of the motor or engine. Final mission designs use the actual measured thrust and the ideal specific impulse corrected for losses.

The ideal specific impulse calculated with the standard programs gives only an ideal performance level for a particular propellant system. Actual performance will depend on the efficiency of the thruster. The extremely small sizes of microsatellite propulsion systems present a variety of problems in calculating the potential delivered performance. The small volumes in these propulsion systems will result in much greater heat loss to the propulsion structure (higher surface to volume ratio), resulting in a performance loss. Small thruster sizes also result in a reduced residence time of the reacting propellants in the combustion chamber, which can lead to mixing inefficiency and additional performance loss.

16.4 Calculation of Thrust

Determination of the thrust needed for a microsatellite obviously depends on the total mass of the craft and the specific mission that must be completed. The necessary thrust can generally be broken up into categories based on the particular intended use: launch, orbit transfer, and stationkeeping/control. Launch vehicles generally require large thrust levels—ranging from 10,000 to 10 million newtons (N) of thrust—dependent on the overall mass of the system. For example, each solid rocket booster on the Space Shuttle gives 11.79 million N of average thrust.[7] Orbital insertion or correction may also require thrust levels from hundreds to 200,000 N, but orbit transfers may not, depending on the time required to reach the desired orbit or trajectory. For example, the upper-

stage Centaur has a nominal average thrust of 185,000 N.[7] Stationkeeping and attitude adjustment generally require low thrust levels, on the order of 0.05 N or less, depending on the overall mass of the spacecraft, the orbit, and other factors such as atmospheric drag and spacecraft-pointing requirements. Thrust levels range from less than a newton to millions of newtons for chemical-propellant rockets, 0.01 to 100 N for cold-gas flow systems, and from 0.0001 to 100 N for electric propulsion.[8]

Microsatellite systems will certainly require extremely small, self-contained propulsion systems to maintain proper attitude and pointing. Electric propulsion systems that require large solar arrays and battery storage are probably too large and massive for the smallest microsatellites. Large bipropellant chemical systems that contain complex pumps, valves, and plumbing are also less useful for relatively small spacecraft. Most of the satellites in use today utilize small monopropellant thrusters or cold-gas thrusters for stationkeeping and attitude/pointing control. Miniature electric propulsion designs, such as pulsed plasma thrusters (PPTs), may provide sufficient thrust for attitude control. However, large velocity change maneuvers that are required in a relatively small period of time will almost certainly depend on chemical rockets.

The thrust T of a particular rocket system can be determined from the mass flow rate of the propellants, Δm, times the exhaust velocity of the combustion products, V_e, added to the thrust due to pressure P against the exhaust nozzle with area A_n:

$$T_h = \Delta m V_e + P A_n = I_{sp} \Delta w \tag{16.2}$$

where P is the difference between the exhaust gas pressure p_e and the ambient static pressure[9] p_s: $P = p_e - p_s$. At approximately zero ambient pressure (vacuum), the thrust is maximum, since $PA_n = A_n p_e$ (since $p_s = 0$). The thrust required is then determined by the change in velocity ΔV needed to accomplish the desired task and the mass of the spacecraft:

$$T = F = Ma = M\left(\frac{dV}{dt}\right) \tag{16.3}$$

for a fixed total mass M and acceleration a. Since the change in velocity is accomplished by loss of propellant mass, the total mass M also changes:

$$F(dt) = M(dV) + V(dM) \tag{16.4}$$

To simplify matters, assume that the amount of propellant used during the change in velocity maneuver is small compared to the total mass of the spacecraft. Then the thrust needed for a fixed ΔV required in a fixed amount of time is about:

$$\int F dt = M\Delta V \text{(for no change in mass)} \tag{16.5}$$

The quantity $\int F dt$ (or I) is called the total impulse and is given in newton seconds. For example, a 10-kg microsatellite that requires a ΔV of 200 m/s for a deorbit burn would need a total impulse of 2000 N-s, assuming essentially no mass loss during the burn. In reality, since propellant is used during the burn and therefore the total mass of the vehicle M is reduced, the total impulse needed is slightly less. This total impulse can then be used to determine the approximate necessary thrust over a set period of time. In this example, a 1-N thruster would require about 2000 s to achieve the required velocity change.

The amount of thrust required, and therefore the type of propellant system to use, is dependent on the specific mission or operation requirements. For microsatellites, very small overall changes

in velocity (ΔV) are needed for normal attitude adjustments and stationkeeping. However, significant orbital changes may require large ΔV values. Table 16.1 lists common velocity change requirements for various missions.

16.5 Propellant Characteristics

There are a number of factors to consider when selecting a propellant for a specific microsatellite application. Most of these are highly mission dependent. The small size of the microthruster presents both opportunities and restrictions. A careful study of the system and energy balance become crucial since the margins are so small. The following factors all play a role in the selection of a propellant.

16.5.1 Density

The density of the propellant will help determine the volume and weight devoted to propellant storage, the velocity or distance the satellite can achieve, and the usable lifetime of the satellite. Solid propellants generally provide a higher density propellant system than liquid propellants. Earth-storable liquid propellants are easier to handle and typically have higher density than space-storable or cryogenic propellants. Low-temperature liquids or cryogenics are stored under pressure, requiring a heavier, thicker-walled storage vessel. New lightweight materials derived from current nanotechnology research may allow for much higher pressures at a lower weight penalty than in conventional systems. This would make the use of cryogenics or gases more attractive in a small system.

16.5.2 Handling and Usability

Liquid propellants that have boiling points above approximately 80°C are the easiest to handle. Low vapor pressures also are desirable to reduce storage requirements, reduce operation hazards, and avoid cavitation in pump-fed systems. Liquid propellants also allow for restart and a throttling capability, which provide greater mission flexibility than solid propellants. Solid pellets may provide a way to achieve both density and restart goals, but a delivery system for solid pellets would add complexity and weight to the system. High-pressure gas systems are the least complicated of all, requiring only a controllable valve, but come with weight penalties from the pressure tank and suffer from low propellant density. A modular, plug-in propellant cartridge concept for use in microsatellites seems particularly appealing. Such a cartridge would lend itself to mass production and would enable the separation of the microsatellite and potentially hazardous propellants until needed.

Table 16.1. Typical Velocity Changes Required for Various Missions

Mission Type	Typical ΔV Required (m/sec)	Mission Details
N-S stationkeeping	50–100 (per year)	In GEO[10]
Deorbit from LEO	150–250	Orbit dependent[10]
10° Inclination Change	1200–1500	Orbit/time dependent[10]
LEO to GEO transfer	4000–6000	Orbit/time dependent[10]
Launch to LEO	9000–14,000	Orbit dependent[3]
Earth escape	~12,000	Launch dependent[3]
Earth to moon	~15,000	Land on moon [3]

16.5.3 Stability

Optimally, propellants for small spacecraft must be stable to decomposition over a wide temperature range (typically about 0°–100°C), for example, nonreactive to storage and component materials and storable for long periods of time. Liquid fuels can be used to help manage the thermal loads of the propulsion system, often being used to cool the combustion chamber, expansion nozzle, or both. Thus, high thermal stability, high specific heat, and low viscosity are desirable.

16.5.4 Hazards

Highly corrosive or toxic propellants increase storage and handling complexity with a concomitant increase in operation costs. The small quantities of propellant that would generally be needed for microthruster applications may mitigate this problem to some extent. It is also important to ensure the exhaust products are clean to avoid contamination issues during space operations.

16.6 Special Propellant Considerations for Microthrusters

16.6.1 Mission, Manufacturing, and Packaging Encapsulation

Most current small satellites use either a cold-gas or monopropellant propulsion system for attitude control. Cold-gas systems are most effective for spacecraft missions with low total impulse, because of low specific impulse and generally low thrust. Monopropellant systems, typically using hydrazine, are widely used today in satellite systems.

The size of microsatellites and very small propulsion systems suggests implementation of missions requiring multiple spacecraft. Several large industry ventures have proposed the use of constellations of orbiting microsatellites. Such small craft can be launched from smaller, cheaper launch vehicles, resulting in overall cost reduction. The extremely small size of these multispacecraft constellations requires very small propulsion "packages" that can still accomplish all needed mission propulsion. In addition, the requirement for large number of craft lends itself to mass production techniques. The propulsion system could be produced separately in large numbers and then integrated into the spacecraft just before launch. In that way, hazardous propellants could be sealed into the propulsion package, eliminating the need for costly shutdown of launch pad operations during propellant loading of the payload.

16.6.2 Density

For microspacecraft, the propellant density will directly affect the total volume of the spacecraft. For very small spacecraft, volume constraints may not be as strict as for larger systems limited by the payload volumes on the launch vehicle. This allows for the possible use of low-density gaseous propellants for short-duration, low-velocity missions. Such systems are low cost, simple, and require little associated hardware.

16.6.3 Simplicity

Propulsion systems that are simple to manufacture, have few parts, and are low cost will require propellants that are also easy to handle. This includes most cold-gas systems, most solid propellants, and a variety of liquids, although the toxicity of the propellant must be taken into account if the spacecraft propellant is to be loaded at the launch site.

16.6.4 Thermal Management

Many bipropellant engines use one or both of the propellants for thermal management of the excess heat generated during engine firing. Propellants that decompose upon heating (such as hydrazine or hydrogen peroxide) cannot be used for these types of bipropellant engines.

Another important consideration involves the basic properties of the propellant, such as the freezing point and boiling point. Propellants that have high freezing points may need to be heated onboard the microspacecraft to avoid propellant solidification. Such heating would use resources that very small spacecraft could not afford to lose. For example, hydrazine has a freezing point just under 0°C, while n-propyl nitrate has a freezing point of about −100°C. Even so, hydrazine is the monopropellant of choice for most spacecraft. Appendixes 16.A–16.D (at the end of the chapter) provide an abbreviated list of thermodynamic and physical properties of common fuels and oxidizers.

16.7 Propellants: Cold-Gas Systems[11]

The simplest type of thruster is the cold-gas system.[3,9] In this thruster there is no chemical combustion involved. Thrust is provided by a pressurized gas, which is injected into the thrust chamber. This system is the simplest type of thruster and causes little contamination, but also suffers from relatively low performance (50–75 s) and thrust (typically 0.05–0.10 N). Cold-gas thrusters were the most common type in the 1960s and are still used today for systems needing less than 1000 lb/s total impulse. Two examples of the use of cold-gas propellants for reaction-control systems are the Viking Orbiter[12] and the LANDSAT 3.[13] The system typically consists of a high-pressure tank, gas valve, pressure regulator, pressure relief valve, and a series of thrusters with valves. The pressure determines the performance of the system, and relatively high pressures have been used, for example 5000 psia.

Helium, nitrogen, argon, krypton and Freon 14 have all flown, but ammonia, nitrous oxide, Freon 12, and hydrogen are also candidate systems. Selection of the optimal gas must include considerations of density, molecular weight, and weight of the storage tank. Of the gases that have been used, helium performs the best, but leakage problems make it more difficult to work with. Theoretical performance values range from 46 s for Freon 12 to 290 s for hydrogen (at vacuum, frozen equilibrium, area ratio = 100, and gas temperature = 25°C).

Performance for a cold-gas system can be calculated as follows.[9] I_{sp} is given by:

$$I_{SP} = \left[\frac{2kRT_c}{g(k-1)}\right]\left[1-\left(\frac{P_e}{P_c}\right)^{\frac{(k-1)}{k}}\right]^{1/2} \tag{16.6}$$

Thrust for the system is obtained from:

$$T_h = P_c A_t C_f \tag{16.7}$$

The volume and gas weight are derived from:

$$W_p = \frac{I}{I_{sp}} \text{ and } V = \frac{(W_p RT)}{P} \tag{16.8}$$

where k is the ratio of specific heats of the gas at constant pressure and volume (C_p/C_v); R is the gas constant; g is the gravitational constant to convert weight into mass; A_t is the nozzle throat area; C_f is the thrust coefficient, which is the improvement in thrust provided by the expansion nozzle;[3] T is the absolute temperature; P is the static pressure; I is the total impulse; subscript c refers to the chamber conditions; and subscript e refers to exhaust conditions.

Performance of cold-gas systems can be improved by heating the gas with either an electric heater or a chemical reaction. Such warm thrusters have achieved performance numbers ranging from 105 to 240 s at the expense of the simplicity of the system. However, in the microsatellite

such heating may play an advantageous role in the total thermal management of the system. For example, if the propellant storage tanks are placed to act as a heat sink for the combustion chamber, the excess heat could be used to pressurize the tanks for propellant delivery.

16.8 Monopropellants

Monopropellants generate propulsive energy through the exothermic decomposition of a single molecule or formulated mixture of fuel and oxidizer. Typically the monopropellant is a liquid that is decomposed into products by a catalyst. The advantage of such a system is its simplicity and reduction in necessary hardware, such as propellant fuel tanks, valves, pumps, and tubing. The major disadvantage is that it generally is more dangerous, with possibility of detonations. Only three monopropellants have been used in actual flight vehicles: hydrazine (and its derivatives), hydrogen peroxide, and propyl nitrate. Monopropellant thrusters are used primarily for attitude control devices.

16.8.1 Hydrazine

The most commonly used monopropellant is hydrazine,[14] which serves as a good baseline for other microthruster candidates. Hydrazine, a fuming colorless liquid with an ammonia-like odor first detectable at 70–80 parts per million (ppm), has physical properties that are very similar to water. Hydrazine is hypergolic with nitric acid, nitrogen tetroxide, and hydrogen peroxide. Spontaneous ignition in air can occur if hydrazine is spilled on a surface or soaked into even inert substances such as vermiculite. The vapor is known to form explosive mixtures with air. Hydrazine is a stable liquid and can be stored in clean containers for many years. In the presence of impurities or at elevated temperatures, decomposition is accelerated. Hydrazine and its alkyl derivatives are strong reducing agents. Hydrazine has a positive enthalpy of formation (50.42 kJ/mol) and performs well as a bipropellant fuel. Both unsymmetrical dimethylhydrazine (UDMH) and monomethylhydrazine (MMH) are more stable than hydrazine, with a wider liquid range (lower boiling point) but slightly lower specific impulse values. Mixing either UDMH or MMH with hydrazine tends to quench the explosive decomposition of pure hydrazine. The bipropellant system consisting of MMH with N_2O_4 as oxidizer is extensively used in small attitude-control satellite engines.[15]

As a monopropellant, hydrazine can be effectively catalyzed and decomposed to form gaseous nitrogen, hydrogen, and ammonia. Iridium metal is effective as a room-temperature catalyst; iron, nickel, and cobalt have all been demonstrated to work at elevated temperatures. The most common catalyst in use today is the Shell 405 catalyst (available from Shell Oil Co.).This catalyst is a porous alumina substrate deposited with iridium.

The major difficulty with hydrazine lies in its toxicity. OSHA (Occupational Health and Safety Administration) lists its threshold level value (TLV) at 0.1 ppm, and it is known to cause irritation to eyes and skin. It can be absorbed through the skin, inhaled, or ingested. It also is a suspected carcinogen. These factors lead to expensive handling procedures and an interest in replacing its use with less toxic propellants.

Hydrazine monopropellant thrusters[9] typically range in size from 0.2 to 2500 N of thrust. In principle, even smaller thrusters could be constructed for microspacecraft. The difficulty would lie in designing an appropriate catalyst bed that would resist poisoning and clogging. The temperature control of the catalyst bed must also be addressed. Typically, the catalyst beds for hydrazine are heated to ensure quick, reproducible engine starts. However, heating of the hydrazine in the propellant tank must be avoided to lessen the danger of decomposition or detonation.

Microspacecraft that use hydrazine or hydrazine derivatives must also consider materials compatibility issues. For example, silicon-based micro-propellant devices are probably compatible with hydrazine use in the absence of water, but specific materials must be tested for chemical compatibility before beginning large-scale manufacturing.

16.8.2 Hydrogen Peroxide

Hydrogen peroxide when pure is an almost colorless (very pale blue) liquid. It is less volatile than water and is more dense and viscous.[14] The compound is miscible at all proportions in water, in which it forms a hydrate. Because of its low reduction potential, aqueous solutions of hydrogen peroxide spontaneously disproportionate. For the pure liquid the reaction proceeds as follows:

$$H_2O_2(l) \rightarrow H_2O(l) + \frac{1}{2}O_2 \quad \Delta H^\circ = -98.2\ kJ\ mol^{-1} \tag{16.9}$$

The pure compound decomposes slowly in the absence of catalysts or impurities. However, in the presence of trace amounts of metals or alkalis the reaction is strongly catalyzed. Even a speck of dust has been known to initiate explosive decomposition. Storage of the anhydrous compound or highly concentrated solutions is generally done in wax-coated or plastic containers with added stabilizers, such as urea. Hydrogen peroxide can act as an oxidizing and as a reducing agent, which leads to a rich chemistry.

Hydrogen peroxide can function both as a monopropellant and as an oxidizer in a bipropellant system. As a monopropellant it can be catalyzed by a number of catalysts, including MnO_2, Pt, and Fe_2O_3 to form hot water and oxygen. As an Earth-storable oxidizer it has several attractive attributes.[16] Its performance and density are close to or better than the commonly used nitric acid based oxidizers; its lower vapor pressure reduces the toxic fumes; and it is less corrosive to metals. The major problem with its widespread use has always been its instability. Since its decomposition pathway is exothermic, the reaction is self-accelerating. This problem is compounded by the fact that the reaction is catalyzed by almost any impurity. Stabilizers can help, but they tend to make ignition and stable combustion difficult. Nevertheless, significant effort has been put into the development and use of hydrogen peroxide. It was used as the oxidizer in high concentrations on the NASA X-1 and X-15 programs, and was the monopropellant used for small attitude-control thrusters before hydrazine was developed; it was also used in some Jet-Assisted Take-Off (JATO) applications. The Navy has tried to develop it as a "nontoxic" propellant for shipboard use, so far without success.[16] It continues to be used for some gas generator applications. Another issue that must be addressed for micropropulsion thrusters is the relatively high melting point of hydrogen peroxide. The pure material melts near 0°C and must be heated to use.

16.8.3 Other Molecular Monopropellants

Other candidate liquid monopropellants include nitromethane, ethylene oxide, n-propyl nitrate, ethyl nitrate, and tetranitromethane. These have all been extensively studied and tested with various stability additives and in various engines. Ethylene oxide is an important industrial chemical and is prepared in large quantities through the direct catalytic oxidation of ethylene. It is highly reactive because of its strained bond angles and can be decomposed by both acid and base solutions.[17] As a monopropellant it has an I_{sp} of 199 s (at 1000 psi, 14.7 psi). It has been used to power decoys and is stable to high-pressure flow conditions, although the vapor is sensitive.[11]

The nitroaliphatics have had the least application as monopropellants. Nitromethane has been utilized the most. Nitromethane, an oily liquid with a moderately strong disagreeable odor, has been used as a fuel in race cars and in model airplanes. OSHA lists a TLV value of 200 ppm,

considerably higher than hydrogen peroxide. It has an I_{sp} value of 255 s (at 1000 psi, 14.7 psi). The nitro alkyls are usually sensitive to thermal instabilities and to shock.

The alkyl nitrates are normally more shock sensitive because of the less stable nature of the carbon-oxygen-nitrogen bonding as compared to the nitro compounds. The liquid monopropellent n-propyl nitrate has been used in jet engine starters and for decoys. The British have used isopropyl nitrate for engine starters and for gas generators.

16.8.4 Composite Monopropellants

New advanced monopropellants can be developed from combinations of energetic liquid oxidizers and energetic fuels. This offers the advantage of selecting the physical and performance properties to fit the particular need. Large gains in performance and density are possible. For example, one new monopropellant formulation that contains nitroformates [containing $C(NO_2)_3^-$] as the oxidizer component would result in specific impulse values over 50 s greater than hydrazine and density increases of over 60%.[18] One major disadvantage is the potential sensitivity to detonation or combustion, since the fuel and oxidizer are premixed. Another potential disadvantage is the possibility of physical phase separation between the various propellant components. Both of these disadvantages may be overcome with the proper formulation mixture of the monopropellant. Formulated composite monopropellants that have been examined include the amine nitrate-nitric acid combinations, difluoramines with NO-based oxidizers, boranes in hydrazine, and slurries of high-energy solids in liquids (for example, beryllium hydrides in hydrazine). None of these composite monopropellants has found widespread use.

16.9 Bipropellants

A bipropellant engine introduces more complexity into the propulsion system. Tanks, valves, and feed lines for both the liquid oxidizer and the liquid fuel must be present. Advantages can be realized, however, in safety, performance, and a wider selection of propellants and propellant combinations. The small size of the microthruster lends itself, perhaps, to an easier integration of the bipropellant engine into a single unit, mitigating some of the disadvantages.

16.9.1 Oxidizers

Oxygen and fluorine provide the best performers as oxidizers, but are not Earth-storable liquids. Nitrogen is a good carrier for these elements, and several oxidizer candidates come from the nitrogen-oxygen or nitrogen-fluorine families. Hydrogen peroxide has already been discussed as a monopropellant, but can function as an oxidizer in a bipropellant system as well. Likely oxidizer candidates for microthruster applications are discussed below.

16.9.2 Nitrogen Oxides

Nitrogen tetroxide is the most widely used storable oxidizer in the propulsion arena. It is a colorless liquid that dissociates reversibly into NO_2 in the gas phase and is thermodynamically unstable with respect to decomposition into N_2 and O_2. As the temperature of nitrogen tetroxide is raised from its freezing point (−11.2°C) to 135°C (at which point it is 99% dissociated), its color becomes a deep brown due to increasing NO_2 concentration.[14] Nitrogen tetroxide reacts with water to form nitric acid; hence the moist gases can be quite corrosive. It is currently used in the Titan programs, the Space Shuttle reaction control system, and in some other satellite systems.[7] It is hypergolic with many fuels and will ignite with other carbonaceous materials. It is compatible with stainless steel, aluminum, and Teflon. It reacts, however, with most elastomers. OSHA lists a ceiling exposure limit of 5 ppm (9 mg/m^3) with possible health hazards primarily to the lungs, and

irritation to the eyes, nose, and throat. It has a high vapor pressure (720 mm Hg at 20°C), which may be useful in designing a microthruster self-pressurizing propellant delivery system.

Various forms of nitric acid have also been widely used as oxidizers. They generally have a wide liquid range and are hypergolic with the hydrazines. The major drawback to their use has been their high corrosive nature. Water increases the corrosiveness, so much effort has been spent developing 100% nitric acid, known as white fuming nitric acid (WFNA). The Germans found that adding 6% NO_2 to WFNA, making red fuming nitric acid (RFNA), improved the stability.[16] Adding less than 1% hydrofluoric acid (HF) to RFNA was found to reduce the corrosion rate at least an order of magnitude in both stainless steels and aluminum. This combination, called inhibited red fuming nitric acid (IRFNA) in varying concentrations, was found to be the least corrosive and most stable and is the most common nitric acid based oxidizer in use.[11] It ignites spontaneously with furfural alcohol, aniline, and other amines and has been used with gasoline, hydrazine, and other alcohols.[3] It does have a high density, which leads to compact tank design. All these systems are highly corrosive and must be used carefully.

Other nitrogen-oxygen oxidizers include tetranitromethane and nitrous oxide. Tetranitromethane,[11] an overoxidized monopropellant, has a high density and oxidizing potential similar to N_2O_4. Impurities seem to sensitize the molecule to shock impact detonation and decomposition. A eutectic mixture of 64% tetranitromethane and 36% (by weight) of N_2O_4 freezes at –30°C and decreases the sensitivity without much loss in performance. Nitrous oxide, or laughing gas, is an underoxidized oxidizer that is a room-temperature gas (bp –88.5°C). Its primary attraction for microthrusters is its nontoxicity and lack of corrosion or compatibility problems (indeed it is commonly used as the propellant in whipped cream cans). It is inert at room temperature, but when heated above 600°C it dissociates into nitrogen and oxygen.[19] It may be attractive in a cold-gas system where some extra chemical energy is desired.

The nitrogen fluorides are relatively high performing but as a class are generally very sensitive and reactive, readily decomposing, often violently. Nitrogen trifluoride has some attractive features, including low toxicity, compatibility with most metals and plastics, and lower reactivity than fluorine.

All fluorine-containing oxidizers suffer from the one major disadvantage for microsatellite applications, in that the exhaust products will contain HF. Hydrofluoric acid is a highly corrosive contaminant that most satellite applications cannot stand.

16.9.3 Fuels

The most commonly used storable liquid fuels have been the hydrocarbons, hydrazines, and alcohols. Kerosene, a blend of unsaturated and saturated hydrocarbons, is attractive because of its low cost and wide availability. Rocket propellant-grade kerosene is called RP-1 and is essentially the jet fuel JP-4 with controlled aromatic content to reduce carbon formation in regeneratively cooled rocket engines. With an empirical hydrogen-carbon ratio of 1.953,[1] RP-1's performance is representative of most hydrocarbon fuels. Strained-ring hydrocarbons can provide both higher enthalpy and higher density, resulting in increased performance. Quadricyclane, a cyclization product of norbornadiene, has been examined as a nontoxic propellant replacement fuel for the second-stage Delta II[20]as well as other rocket propellant applications.[21] A hydrogenated condensation product of norbornadiene, known commercially as Shelldyne-H™, also has been used as a high-density (1.11 g/cm) thermally stable fuel.[11] Additional candidates include the triangulanes,[22] the cubanes,[23] the prismanes,[24] as well as others.[25] While all these molecular fuels would be more expensive than kerosene, they offer advantages for micropropulsion devices in performance, ease of handling, stability, and most importantly, in density. It is also expected that the

strained-ring candidates would be a more reactive hypergol than kerosene. The small amounts required for microsatellite operations certainly permit the use of these materials. The wide variety of candidates makes specific tailoring of the propulsion system possible. The use of one of these fuels with either hydrogen peroxide or a nitric-acid/N_2O_4 oxidizer in a prepackaged, cartridge-type microthruster would seem to be the optimal microthruster design for high-velocity microsatellite applications. Hydrazine has been discussed as a monopropellant, but it and its organic derivatives also make high-performing fuels in bipropellant combustion. Unsymmetrical dimethyl hydrazine (UDMH) is often used in place of hydrazine because of its more desirable physical properties. It has a lower freezing point, a higher boiling point, and is a more stable liquid. These factors outweigh its slightly lower performance, making it the fuel of choice over hydrazine. Mixtures of the two fuels have also been used; for example, the Titan II uses a 30%–50% mixture of UDMH in hydrazine. Monomethyl hydrazine has been used extensively in spacecraft operations with nitrogen tetroxide. All these hydrazines, however, are highly toxic.

16.10 Solid Propellants

The advantages of solid propellants over liquid propellants lie primarily in their simplicity (few, if any, moving parts and monolithic structures), their typically higher densities and thrusts over liquid propellants, and their long-term storability. The disadvantages are typically lower specific impulse than liquid systems (though higher than most monopropellants), no restart or refill capability, and the potential for catastrophic failure (the solid propellant contains both oxidizer and fuel in one composite material and is in essence a solid monopropellant).

16.10.1 Conventional Solid Propellants

Two types of solid propellants have been traditionally identified: double-based propellants and composite propellants. Double-based propellants are those that are based mainly on combinations of nitroglycerin and nitrocellulose. This forms a rigid, homogeneous structure. Composite propellants consist of a polymeric matrix in which small particles of an oxidizer and fuel are distributed. The most widely used current solid propellants (for example, the solid rocket motor boosters of the Space Shuttle) use ammonium perchlorate as the oxidizer and aluminum as the fuel. The polymer binder is hydroxy-terminated polybutadiene, which also acts as a fuel to some extent. This combination leads to ideal sea-level specific impulses in the 260- to 265-s range at optimum mixture ratios.

Other solid-propellant ingredients include alternate oxidizers such as ammonium nitrate or potassium perchlorate. The nitramines 1,3,5-trinitro-triazacyclohexane (RDX) and 1,3,5,7-tetraazacyclooctane (HMX) are also used as oxidizers, particularly in applications where exhaust products without HCl are desired. However, the nitramines are molecular monopropellants and can detonate. While aluminum is the most widely used fuel in solid composite propellants, other metal powders can be used. Boron will produce higher theoretical performance, but is actually difficult to burn with good efficiency. It has been used as a burn rate accelerator, however.[3] Beryllium will also produce enhanced specific impulse values over aluminum, but is highly toxic. A number of different binder systems have been used, with the hydroxy terminated polybutadiene (HTPB) and carboxy terminated polybutadiene (CTPB) being the most commonly used today.[3] Many other ingredients or additives have been used and characterized over the years as burn-rate or combustion modifiers.[26]

Micropropulsion applications that would benefit from a conventionally designed solid-propulsion system are those that would only require a single, short burn time, have a large-thrust or high-velocity requirement, or one that requires an extremely simple one-package propulsion device.

Several small solid rocket motors, down to 0.4 kg, have been built and demonstrated.[27] Even smaller, hobby-size solid motors for model rockets have been in use for decades.

16.10.2 Gas Generator Propellants

The purpose of gas generator propellants is to produce high-pressure, energetic gas for use in applications such as gas turbines, power production, or jet engine starters. These propellants are attractive for hot-gas thrusters in micropropulsion applications. The solid propellants could be safely stored in a cartridge capable of high pressure until needed. The solid propellant is then decomposed through a chemical reaction or other ignition source to produce a high-pressure thrust device.

The automobile gas bag industry is a good source for readily available, relatively cheap gas generators. Their requirements for nontoxic, clean-burning, relatively low temperature, quickly and easily ignited solid propellants are a good fit for the needs of mass-produced micropropulsion devices. Alkali azides (NaN_3, or KN_3) have been widely used as gas generators and are attractive because of their tendency to decompose smoothly and quantitatively when heated to 300°C.[28] Stabilized ammonium nitrate based propellants have also been used as gas generators [3] and give a smokeless, clean exhaust. Other candidates for such micropropulsion applications include the nitramines, organic azides,[29] the metal hydrides,[14] and other monopropellant solids.

16.11 Propellant Delivery and Combustion Concepts

16.11.1 Mass Production

Microsatellite concepts that require hundreds of spacecraft in low Earth orbit will need mass production techniques for low-cost manufacturing of spacecraft components. The propulsion systems for these spacecraft may be produced as a complete unit and integrated into the spacecraft just before use. This provides greater flexibility for mass production and use of potentially hazardous propellants. The small sizes necessary for the propulsion system may allow the use of simple production techniques, such as molded plastics or composites. Single-piece combustion chamber/nozzles may also permit ease of production.

Material selection for those parts of the propulsion system that come in contact with possibly corrosive propellants or hot zones, such as the combustion chamber, needs to be carefully considered. Material compatibility of a wide variety of materials with both liquid[30] and solid[26] propellants is available. Silicon has been suggested as a structural material for the batch production of microthrusters, fabricated in the same way as those used in the semiconductor industry.[31] Propellant compatibility and high-temperature use of silicon can be greatly enhanced by forming a nitride layer on the contact surfaces. Silicon nitride (Si_3N_4) is almost completely chemically inert; it retains its strength, shape, and erosion resistance even above 1000°C.[14]

Any manufacturing technique will find it difficult to compete on a cost basis with those found in the plastics industry. New generations of plastics and plastics composites, termed nanoreinforced plastics, offer major technology improvements. The nanotechnology offers significantly higher operational temperatures over those for today's existing polymeric resins. The two leading technologies in this area are nanostructured clays[32] and silsesquioxane[33] chemical technology.

Current research in microelectromechanical devices and high-performance polymers may enable mass production of component parts or entire systems for micro-sized satellites. Single-piece units that contain both the propellant storage and feed lines are possible. It may also be possible to include the combustion chamber and expansion nozzle into this single unit, depending on the materials used, the size of the system, and the overall compatibility with the propellants.

For any small satellite system, the propulsion system must be limited in mass to less than half the total to provide enough mass for usable payload. Long missions that need stabilization or attitude adjustment would require most of the propulsion mass to be used for propellant. That leaves very little mass for the propellant storage, feed system, and engine/motor. Volume limitations may also force the feed system lines and engine valves/injectors to such a small size that small-scale physical effects, such as surface tension and capillary action, become important in the design. Such factors may ultimately limit the selection of the propellants and propulsion system. For example, propulsion systems with propellant feed lines that are less than about one-tenth of a millimeter in radius will not be able to use highly dense, viscous propellants, such as some of the dense energetic hydrocarbon fuels.

16.11.2 Cartridge/Combustion-Chamber Combinations

A multiuse cold-gas propellant system for microsatellite control might be constructed from an integrated solid-propellant/high-pressure tank arrangement. Conceptually, such a system might have a series of small solid-propellant canisters arranged around a spherical pressure vessel. Each canister would have an electronic igniter and would be fired when the pressure in the tank reached a preset low value. This refills the tank as needed and also stores the propellant in a highly dense form. The propellants for such a multiuse system would most likely be state-of-the-art, high-energy solid-propellant formulations with ingredients such as RDX or ammonium dinitramide.

Another prepackaged solid-propellant concept involves small pellets of propellant that are individually moved into the combustion chamber by a relatively low-pressure gas supply and valve system. It may also be possible to mechanically move the propellant into the chamber by using a linear "blister film," where each solid-propellant slug is encased in a thin plastic bubble on a strip of plastic. In either case, each pellet package could then be ignited by the focused pulse of a small semiconductor laser to obtain a set impulsive burst.

Both concepts use solid propellants that have higher density and greater ease of handling than, for example, liquid-hydrazine-based monopropellants. However, the blister film concept suffers from lack of quick response times and the incorporation of moving parts, which add to the risk of failure and are more complicated to mass produce. The pressurized system suffers from the low performance of a typical cold-gas propulsion system. It is also probably not amenable to mass production and difficult to incorporate into an extremely small microsatellite.

16.11.3 Nonchemical Propulsion Methods

In addition to conventional chemical propulsion schemes, propulsion systems that utilize electrical energy as a major driving source can be used in satellites. Electrothermal thrusters, arc jets, and ion thrusters have all been demonstrated,[9] yet typically require kilowatts of electrical power. Plasma thrusters and ion engines must also carry their propellant gas (xenon for example), which is ionized and accelerated. These methods thus face the propellant density constraints similar to cold gas thrusters. Ion engines can have I_{sp} values of several thousand seconds with thrusts less than about 100 mN.

Developing a mass-produced fuel that can be integrated into a microsatellite concept may require consideration of novel propulsion concepts. The rapid development of compact, efficient, and lightweight diode laser devices merits their consideration as components in propulsion systems as either ignition sources or propulsion drivers.

An example of a simple, precisely controllable, low-thrust propulsion concept might use a laser to ablate material supported on a transparent substrate, and use the force of the ablated gaseous plume for propulsion. In this example, a short laser pulse (~1 ns per pulse) passes through a substrate and is absorbed by an active propellant supported on the far side of the substrate. The

irradiated portion of the propellant layer is ablated, forming a gaseous plume that leaves normal to the substrate surface with high velocity (several kilometers per second). The active coating layer, perhaps 1 μm thick, could be a decomposable polymer such as polymethylmethacrylate, or nitrocellulose. Dye can be incorporated into the layer to increase absorption, and energetic materials might also be added to increase the energy released, thereby increasing the velocity and mass of the ablated material. The I_{sp} for such a system would ultimately depend on the ablation dynamics that determine the mass and velocity distribution of the atoms, molecules, clusters, and ions that are ejected from the substrate. For this assessment, conversion of 25% of the laser pulse energy to kinetic energy in the ablation plume is a reasonable estimate. The substrate, possibly a thin Mylar film, could be advanced to expose a new spot of active propellant layer following each laser pulse.

Q-switched diode-pumped microlasers with pulse energy of 5 μJ, pulse duration of 1 ns and pulse repetition rates of 20 kHz are available. An electrical power input of less than 1 W produces about 100 mW of average output laser power. If the laser beam is focused through the Mylar substrate onto the active layer with an area of 10^{-6} cm^2, assuming a density of the active layer of polymer of about 1 g/cm^3, approximately 10^{-10} g of material would be ablated per laser pulse. If 25% of the laser pulse energy is converted into kinetic energy of the ablated products moving normal to the substrate, the velocity of these products would be 5000 m/s. Focusing the laser beam more tightly, using a more powerful laser, or adding an energetic (explosive, pyrotechnic, or solid propellant—see above) compound to the active layer can vary the velocity of the ablated material. Operating at the 20-kHz laser repetition rate, approximately 60 g of material could be ablated per year, producing a ΔV of 300 m/s per year for a 1-kg microsatellite. Advances in laser technology and energetic materials in the active layer could greatly increase these figures for this easily storable, precisely controllable method. Many other laser-driven concepts also merit consideration for the unique demands of microsatellite propulsion.

16.11.4 Laser Propulsion

Certainly one laser propulsion technique that may be considered for microsatellite operations involves direct conversion of ground-based, high-power laser light into on-orbit thrust. Recent advances in laser control, pointing, and adaptive optics provide new opportunities for development of this concept. Such a system might use a high-peak-power pulsed laser, at a wavelength that transmits through the atmosphere with minimal loss, to direct concentrated light onto a mirrored collector on the spacecraft. The laser pulses would then be focused onto a thruster to heat a simple gas or liquid propellant. In this way, nontoxic, storable, easy-to-handle, and low-cost propellants (such as water) could be used for propulsion. Another advantage of such a system is that the complicated power source is based on the ground, where it is easily repaired and upgraded.

The primary use for such a laser propulsion system would be stationkeeping and orbit change, including a space tug concept for shuttling payloads from low Earth orbit (LEO) up to geosynchronous orbit (GEO). Since the major equipment complications are ground based (the laser), the target craft (the spacecraft) can be mass produced at a greatly reduced cost per spacecraft. A variety of spacecraft designs could also be used, resulting in greater mission flexibility. For example, a 1-megajoule (MJ) per pulse ground-based laser with 1-m-diameter beam-pointing optics gives a delivered power at a 200-km orbit (assuming very low laser divergence of about 10^{-4} rad and no atmospheric loss) of about 70 mJ/cm^2 fluence. For a 10-ns pulse length, this results in peak powers of about 7 MW/cm^2 for one pulse per second. For a satellite collection area of about 1 m^2, the focused laser power (about 700 J per pulse) is more than sufficient for plasma breakdown in hydrogen, helium, nitrogen, or even water.

16.11.5 Pulsed-Detonation Engines

An alternative to the standard continuous rocket engines is the pulsed-detonation engine. Such an engine operates by detonating a fuel/oxidizer mixture in a partially confined chamber (similar to the internal combustion engine). Potentially higher temperatures and pressures can be achieved for short periods of time. One potential advantage to such engines is the ability to pressure feed the propellants into the combustion chamber. This effectively eliminates large, heavy turbo pumps that pressurize the propellant before injection into the combustion chamber. The simplicity of the pulsed-detonation rocket engines lends itself to low-cost mass production techniques. However, there are limitations on the reduction in size of these devices due to physical limits of heat transfer processes into the containment walls. Very small spacecraft will therefore probably not use pulsed detonation for propulsion.

16.12 Acknowledgments

The authors would like to thank Dr. Tom Hawkins, Dr. Joe Lichtenhan, and Paul Jones of the U.S. Air Force Research Laboratory for comments and suggestions regarding propellants and Professor D. Dlott for useful discussions regarding nonchemical propulsion methods.

16.13 Appendixes

Appendix 16.A. Thermodynamic Properties of Liquid Fuel[11,34]

Fuel (formula)	Standard Heat of Formation (kJ/mol)	Critical Temperature (K)	Critical Pressure (MN/m^2)	Heat of Vaporization (kJ/mol)
Aerozine-50 (50% hydrazine/ 50% UDMH)	+51.51	607	11.935	33.67
Ammonia (NH_3)	−71.71	405.4	11.280	23.56
Diethylcyclohexane ($[C_2H_5]_2C_6H_{10}$)	−726.18	638.6	2.534	38.79
Diethylenetriamine (DETA) ($[NH_2C_2H_4)_2NH]$	−77.40	496	3.710	50.50
u-Dimethyl hydrazine (UDMH) $[(CH_3)_2N_2H_2]$	+51.63	523	5.978	35.02
Ethanol (C_2H_5OH)	−277.65	516	6.378	38.79
Ethylene oxide (C_2H_4O)	+25.15	468.9	7.191	25.48
Furfuryl alcohol (C_5H_5OOH)	−276.35	572	3.5	53.62
Hydrazine (N_2H_4)	+50.42	653	14.693	43.43
Hydrogen (H_2) l	−9.01	33.21	1.296	0.915
Kerosene (RP-1) (H/C = 2.0)	−26.03	676.5	2.172	49.79
Methane (CH_4)	−89.50	191.0	4.64	8.91
Methanol (CH_3OH)	−239.03	513	7.87	39.32
Monomethyl hydrazine (MMH) ($CH_3N_2H_3$)	+54.84	585	8.24	40.38
Nitromethane (CH_3NO_2)	−139.03	587.9	6.314	38.28
N-propyl nitrate ($C_3H_7NO_3$)	−214.47	580	4.05	34.68

Appendix 16.B. Physical Properties of Liquid Fuels[11,34]

Fuel (formula)	Molecular Weight (g/mol)	Freezing Point (K)	Normal Boiling Point (K)	Specific Gravity
Aerozine-50 (50% hydrazine/50% UDMH)	41.8	267.6	343.15	0.899 @ 298 K
Ammonia (NH_3)	17.0	195.4	239.8	0.682 @ NBP [a]
Diethylcyclohexane [$(C_2H_5)_2C_6H_{10}$]	140.3	194.2	447.2	0.804 @ 293 K
Diethylenetriamine (DETA) [$(NH_2C_2H_4)_2NH$]	103.2	234	480	0.953 @ 298 K
u-Dimethyl hydrazine (UDMH) [$(CH_3)_2N_2H_2$]	60.1	216.0	335.5	0.791 @ 298 K
Ethanol (C_2H_5OH)	46.1	158.7	351.4	0.789 @ 293 K
Ethylene oxide (C_2H_4O)	44.01	162	283.6	0.887 @ NBP
Furfuryl alcohol (C_5H_5OOH)	98.1	240	444	1.13 @ 293 K
Hydrazine (N_2H_4)	32.04	274.7	386.6	1.008 @ 293 K
Hydrogen (H_2)	2.016	13.8	20.21	0.0709 @ NBP
Kerosene (RP-I) (H/C = 2.0)	172	228	450–547	0.80-0.81 @ 293 K
Methane (CH_4)	16.04	90.6	111.6	0.451 @ NBP
Methanol (CH_3OH)	32.04	175.4	337	0.791 @ 293 K
MHF-3~ (86% MMH/14% N_2H_4)a	43.41	219	362.8	0.889 @ 298 K
Monomethyl hydrazine ($CH_3H_2H_3$)	46.07	220.8	360.8	0.879 @ 293 K
Nitromethane (CH_3NO_2)	61.04	244.1	374.3	1.135 @ 298 K
N-propyl nitrate ($C_3H_7NO_3$)	105.1	172	383.6	1.058 @ 298 K
Otto Fuel II	186.52	245.2	NA	1.232 @ 298 K

[a] Normal boiling point (NBP)

Appendix 16.C. Thermodynamic Properties of Liquid Oxidizers[11,34]

Oxidizer (formula)	Standard Heat of Formation (kJ/mol)	Critical Temperature (K)	Critical Pressure (MN/m^2)	Heat of Vaporization (kJ/mol)
Chlorine pentafluoride (ClF_5)	−253.13	416.2	5.316	22.22
Chlorine trifluoride (ClF_3)	−185.77	453	6.626	27.53
Florox (ClF_3O)	−165.69	455	NA	32.22
Fluorine (F_2)	−12.96	144.5	5.583	6.32
Fluorine/oxygen (70/30) (Flox)	−36.04	144.5	5.583	6.48
Hydrogen peroxide (H_2O_2)	−187.78	730	21.58	47.07
Nitric acid-Type IIIAb	−140.16	544	8.867	34.10
Nitric acid-Type IVb	−181.59	540	9.846	34.52
Nitrogen tetroxide (N_2O_4)	−19.58	431.4	9.938	38.12
Nitrogen trifluoride (NF_3) g	−131.50	206.5	5.026	11.59
Oxygen (O_2)	−12.12	154.8	5.375	6.82
Oxygen difluoride (OF_2)	+24.52	215.4	4.958	11.13
Ozone (O_3)	+118.41	260.7	5.537	14.27
Perchloryl fluoride (ClO_3F)	−21.42	368	5.372	19.96
Tetrafluorohydrazine (N_2F_4)	−21.8	309	7.8	15.23

Appendix 16.D. Physical Properties of Liquid Oxidizers[11,34]

Oxidizer (formula)[a]	Molecular Weight (g/mol)	Freezing Point (K)	Normal Boiling Point (K)	Specific Gravity
Bromine pentafluoride (BrF_5)	174.9	211.8	313.6	2.47 @ 298 K
Chlorine pentafluoride (ClF_5)	130.4	170.2	259.4	1.78 @ 298 K
Chlorine trifluoride (ClF_3)	92.4	195.9	284.9	1.81 @ 298 K
Florox (ClF_3O)	108.4	207.1	302.6	1.85 @ 298 K
Fluorine (F_2)	38.0	54.1	85.0	1.50 @ NBP[b]
Fluorine/Oxygen (Flox) (70% F_2/30% O_2)	36.2	NA	86.5	1.24 @ NBP
Hydrogen peroxide (90% H_2O_2/10% H_2O)	32.4	261.6	414.3	1.39 @ 293 K

Appendix 16.D. Physical Properties of Liquid Oxidizers[11,34]*—Continued*

Oxidizer (formula)[a]	Molecular Weight (g/mol)	Freezing Point (K)	Normal Boiling Point (K)	Specific Gravity
Hydrogen peroxide (98% H_2O_2)	34.0	272.7	423.4	1.45 @ 293 K
MON-10 (10% NO/90% N_2O_4)	85.8	250.4	283.6	1.47 @ 273 K
IRFNA IIIA (14% NO_2, 83.4% HNO_3, 2% H_2O, 0.6% HF)	59.4	224.3	333	1.55 @ 298 K
IRFNA IV (44% NO_2, 57.4% HNO_3, 0.5% H_2O max, 0.6% HF)	55.0	235.9	297.9	1.62 @ 298 K
Nitrogen tetroxide (N_2O_4)	92.0	61.9	294.3	1.43 @ 293 K
Nitrogen trifluoride (NF_3)	71.0	65.9	143.7	1.54 @ NBP
Oxygen (O_2)	32.0	54.7	90.2	1.15 @ NBP
Oxygen difluoride (OF_2)	54.0	49.3	128.4	1.52 @ NBP
Ozone (O_3)	48.0	79.8	160.9	1.61 @ 78 K
Perchloryl fluoride (ClO_3F)	102.4	127	226.3	1.39 @ 298 K
Tetrafluorohydrazine (N_2F_4)	104.0	110.2	200.2	1.56 @ 173 K
Bromine pentafluoride (BrF_5)	174.9	211.8	313.6	2.47 @ 298 K

[a] See Appendix 16.C for nitric-acid compositions.
[b] Normal boiling point (NBP).

16.14 References

1. B. Siegel and L. Schieler, *Energetics of Propellant Chemistry* (Wiley, New York, 1964).
2. J. A. Barnard and J.N. Bradley, *Flame and Combustion* (Chapman and Hall, New York, 1985).
3. G. P. Sutton, *Rocket Propulsion Elements*, 6th ed. (Wiley, New York, 1992).
4. W. T. Thomson, *Introduction to Space Dynamics* (Dover, New York, 1986), p. 242.
5. *Theoretical Isp Program*, C. Selph and R. Hall (Air Force Rocket Propulsion Laboratory), adapted for microcomputers by C. Beckman, R. Acree, and T. Magee (Air Force Phillips Laboratory, Propulsion Directorate). This is a one-dimensional adiabatic isentropic equilibrium program, with full two-variable grid search capabilities.
6. M. W. Chase, Jr., C. A Davies, J. R. Downey, Jr., D. J Frurip,. R. A McDonald, and A. N. Syverud, "JANAF Thermochemical Tables," 3rd ed., *J. Phy. Chem. Ref. Data* **14**, S1 (1985).
7. S. J. Isakowitz, *International Reference Guide to Space Launch Systems* (AIAA, Washington D.C., 1991).
8. N. J. Barter, *TRW Space Data Book*, 4th ed. (TRW Space & Technology Group, 1992).
9. C. D. Brown, *Spacecraft Propulsion*, AIAA Education Series, 1995.
10. S. Janson, "Spacecraft As an Assembly of ASIMs," in *Microengineering Technology for Space Systems*, The Aerospace Corp. Report no. ATR-95(8168)-2 (1995).

11. F. S. Forbes, "Liquid Rocket Propellants," in *Encyclopedia of Physical Science and Technology*, Vol. 7 (Academic Press, 1987).
12. N. A. Holmberg, R. P. Faust, and H. M. Holt, *Viking '75 Spacecraft Design and Test Summary*, NASA Ref. Pub. 1027 (1980).
13. *LANDSTAT 3 Reference Manual*, General Electric Space Division, Philadelphia, PA (1978).
14. N. N. Greenwood and A. Earnshaw, *Chemistry of the Elements* (Pergamon, New York, 1984).
15. R. E. Dueber and D.S. McKnight, *Chemical Principles Applied to Spacecraft Operations* (Krieger Publishing, 1993).
16. J. D. Clark, *Ignition* (Rutgers University Press, New Jersey, 1972).
17. R. T. Morrison and R.N. Boyd, *Organic Chemistry* (Allyn and Bacon, Mass., 1975).
18. T. W. Hawkins, "Progress of Solid and Liquid Propellant Development at Phillips Laboratory," *Proceedings of the 1997 High Energy Density Materials Contractors Conference*, (Chantilly, VA), in press; PL-TR-97-3057, Air Force Research Laboratory (1997).
19. I. R. Beattie, *Mellor's Comprehensive Treatise on Inorganic and Theoretical Chemistry* (Longmans, London, 1967).
20. M. F. Winthrop and R.B. Cotta, *Options for Application of Nontoxic Propellants in the Second Stage Delta II*, PL-TR-94-3009, AF Phillips Laboratory (1994).
21. R. Nichols and S. L. Rodgers, *High Energy Rocket Propellant*, U.S. Patent No. 5616882 (1 April 1997).
22. H. D. Beckhaus, C. Riichardt, S.I. Kozhushkov, V.N. Belov, S.P. Verevkin, and A. de Meijere; "Strain Energies in [n]Triangulanes and Spirocyclopropanated Cyclobutanes: An Experimental Study," *J. Am. Chem. Soc.* **117**, 11,854–11,860 (1995).
23. P. E. Eaton, "Cubanes: Starting Materials for the Chemistry of the 1990s and the New Century," *Angew. Chem. Int. Ed. Eng. 31*, 1421–1436 (1992).
24. G. Mehta and S. Padma, "Synthesis of Prismanes," in *Carbocyclic Cage Compounds* (VCH Publishers, New York, 1992).
25. K. B. Wiberg, "Structures, Energies and Spectra of Cycloproppanes," in *The Chemistry of Cyclopropyl Group* (Wiley, New York, 1987).
26. *CPIA/M3 Solid Propellant Ingredients Manual*, CPIA, The Johns Hopkins University, GWC Whiting School of Engineering, Columbia, MD (September 1994).
27. J. Mueller, "Thruster Options for Microspacecraft: A Review and Evaluation of Existing Hardware and Emerging Technologies," Paper presented at the 33rd AIAA/ASME/SAE/ASEE Joint Propulsion Conference and Exhibit, Seattle, WA, 6–9 July 1997, AIAA 97-3058.
28. F. A. Cotton and G. Wilknison, *Advanced Inorganic Chemistry*, 4th ed. (Wiley, New York, 1980).
29. P. L. Marinkas, ed., *Organic Energetic Compounds* (Nova Science Publishers, 1996).
30. *Liquid Propellant Manual*, CPIA/M4, The Johns Hopkins University, GWC Whiting School of Engineering, Columbia, MD (June 1992).
31. S. W. Janson and H. Helvajian, "Batch-Fabricated Microthrusters: Initial Results," Paper presented at the 32nd AIAA/ASME/SAE/ASEE Joint Propulsion Conference and Exhibit, Lake Buena Vista, FL, 1–3 July 1996, AIAA-96-2988.
32. B. Miller, "Nano-Clay Particulates Create New Compounds," *Plastics Formulating and Compounding* **3** (3), 30–32 (1997).
33. J. J. Schwab, T. S. Haddad, J. D. Lichtenhan, P. Mather, and K. P. Chaffee, "Property Enhancements of Common Thermoplastics via Incorporation of Silicon Based Monomers: Polyhedral Oligomeric Silsesquioxane Macromers and Polymers," in *Proceedings of the Society of Plastics Engineers*, 54th ANTEC, 1817–1820 (1997).
34. D. R. Lide, *Handbook of Chemistry and Physics*, 73rd ed. (CRC Press, New York, 1992–1993).

17

Micropropulsion Systems for Aircraft and Spacecraft

S. Janson,* H. Helvajian,* and K. Breuer†

17.1 Introduction

Microelectromechanical systems (MEMS) initiated a revolution 10 years ago that brought the "muscle" (actuation) to silicon "chips." With MEMS and CMOS (complementary metal oxide semiconductor), it is now possible to endow an instrument with "eyes, ears, nose, muscle, and intelligence." Furthermore, by applying advanced microelectronics packaging solutions, the various elements of this instrument can be integrated into a compact system capable of autonomous action. Although not immediately obvious, all these technologies can be used to develop "smart" micropropulsion systems for general aerospace applications. Moreover, these micropropulsion systems can be fabricated in lots of hundreds or thousands, enabling the development of a new generation of smaller less-expensive spacecraft and miniature unmanned aircraft. Because micropropulsion systems require moving parts (e.g., valves and propellers) and must make the most efficient use of the available mass, volume, and limited quantity of propellant fluid, MEMS and microfabrication technology play key roles in their development and operation. Recent reviews on micropropulsion technology can be found in Mueller,[1] Janson,[2] and deGroot and Oleson.[3] The application of these micropropulsion systems to space and terrestrial use can be found in Chapter 2 and in Gallington *et al.*,[4] and McMichael and Francis.[5] In the United States, micropropulsion system development is currently under way at The Aerospace Corporation (Aerospace), the Air Force Research Laboratories (AFRL), NASA Jet Propulsion Laboratory (JPL), NASA Lewis Research Center, Massachusetts Institute of Technology (MIT), California Institute of Technology (Cal Tech), and TRW, Inc.[1,6–9]

The term microthruster was originally used decades ago for any thruster that produced less than 1 lbf (pound force or 4.45 N) thrust. In the mid-1960s, a review of microrocket technology stated that "a new and rapidly growing branch of rocket technology has emerged recently which may be called microthrust or microrocket technology."[10] The term "micro" simply meant small; microprocessors and submicron semiconductor fabrication were still in the distant future. Today, one can buy small off-the-shelf thrusters in the 10–1000-mlbf (millipound-force) range (or 45–4500-mN range) that are no longer considered micro. For this chapter and specifically for space applications, we define "micropropulsion" as propulsion systems with thrust levels in the micro-to-millinewton range that are based on micron-to-millimeter scale structures. In contrast, micropropulsion for "air-breathing" systems is defined as that required by a micro unmanned air vehicle (<15 cm size) that is designed to operate at low Reynolds number ($<10^4$). In both applications, the systems must be small, preferably made of low-mass materials, and packaged as an integrated unit.

This chapter is designed to accommodate a reader not familiar with rocket or air-breathing propulsion systems. Basic concepts and equations used in comparing propulsion systems and efficiencies are given in the introduction. The chapter is organized into the following sections.

---

*Center for Microtechnology, The Aerospace Corporation

†Department of Aeronautics and Astronautics, Massachusetts Institute of Technology

- Introduction and overview of micropropulsion
- Micronozzles, experimental data on thrust efficiencies and the modeling of the gas/fluid dynamics
- Discussion of two fluidic actuators—the synthetic jet and the microjet engine
- Presentation of a 3-dimensional (3D) material-processing technique for microfabrication of glass/ceramic materials
- Process flow and fabrication of three microthrusters—a pulsed cold gas thruster, a single-shot "digital" thruster, and a microresistojet thruster
- Experimental results from the cold gas and the digital thruster
- Discussion of a concept for utilizing the product waste gas from a bioorganic digester to provide supplemental fuel for a space resistojet thruster—concept for a "self-consuming" thruster
- Conclusions

17.1.1 Air-Breathing Engines

Dimensionless numbers are used in fluid mechanics to identify regimes of operation and relevant physical phenomena. Mach number Ma, for example, is the local flow velocity divided by the local speed of sound (U/s, where U is the flow speed and s is the speed of sound). Subsonic ($Ma < 1$) and supersonic ($Ma > 1$) flows have completely different characters and are usually modeled using different equations and/or integration schemes. Reynolds number, defined as Ud/ν where d is a characteristic length (tube diameter, wing chord, etc.) and ν is the kinematic viscosity of the fluid, is the ratio of inertial forces to viscous forces in the fluid. Low Reynolds number flows, particularly below 2000, are dominated by viscous effects, which usually produce stable flow. High Reynolds number flows usually produce turbulence and low drag along surfaces. Finally, the Knudsen number is the ratio between the mean-free path in the fluid and the characteristic device length. Knudsen numbers greater than 1 indicate free molecular flow conditions; collisions between fluid molecules are unimportant compared with molecule-surface collisions. Knudsen numbers below 0.01 indicate that continuum flow prevails (the range between 0.01 and 1 is the transition region).

Air-breathing engines accelerate portions of the surrounding atmosphere to generate thrust. They may carry their own fuel to generate energy. Fluid actuators, in particular, control the flow or pressure of a working fluid; aircraft propellers and jet engines represent just one important subset of this broad category.

Conventional air-breathing turbomachinery systems operate a thermodynamic cycle in which air passes through a compressor, after which heat is added (usually by burning fuel) and then expanded through a turbine. Some of the work delivered by the turbine is used to drive the compressor (thus closing the thermodynamic cycle), and the remaining available work can be used to generate electric power (using a generator) or to generate thrust directly by exhausting the gas at high speed out of the engine.

As is common for MEMS devices, the chief attraction for micromachined air-breathing propulsion systems derives from the cube-square law. Air-breathing propulsion systems (such as gas turbine engines and actuator disks like propellers or synthetic jets) generate thrust by drawing in low-momentum air on one face and expelling the air with high momentum from the other side. The thrust generated by the device is thus proportional to the exit area of the device. In contrast, the weight of the machine scales with its volume. Thus, the thrust-to-weight ratio *increases* as the device gets smaller, and this would seem to favor MEMS propulsion systems.

The challenges in making microturbomachinery can be divided into the fundamental physical difficulties and the engineering challenges. In the first, the decrease in component efficiency as a result of viscous losses is the chief difficulty with designing a MEMS microengine. In the second

category, many engineering difficulties are evident, including the impact of the limitations of current MEMS technology in the design and fabrication of compressors and turbines, the difficulty of supporting and lubricating high-speed rotating micromachinery, and the need for high-temperature materials that can also be microfabricated.

However, there are aspects of MEMS devices that can also be exploited. In conventional turbomachinery, the size of the rotating components is often limited by the fracture limit of the material. Since silicon has an exceptional strength-to-weight ratio, the rotating parts can actually be proportionately larger than their macroscopic counterparts, which promises an improved machine performance. Some of these design trade-offs are discussed in Jacobson and Piekos *et al.*[11]

17.1.2 Rocket Science

The basic figures of merit for rockets are thrust, minimum impulse bit, and specific impulse (Isp) (see Chapter 2). In general, 1-kg class satellites require 10–1000-μN thrusters, and 10-kg class satellites require 0.1–10-mN thrusters for typical on-orbit operations such as attitude control and orbit maintenance. Impulse bit is defined as the time integral of thrust, and the minimum impulse bit is the smallest value that a given propulsion system can deliver. The minimum impulse bit required is highly mission dependent; it can range from less than 1 μN-s for 1-kg-mass spacecraft performing station-keeping during an optical interferometry mission to greater than 1 N-s for a 10-kg spacecraft performing a significant orbit-raising maneuver. Smaller is generally better, especially for propulsion systems used for a variety of different tasks. Specific impulse is defined as the thrust divided by the mass-flow-rate of propellant through the thruster, and is a function of propellant and thruster type. It can range from about 50 s for a simple cold-gas thruster to greater than 5000 s for an electric thruster such as an ion engine. Higher I_{sp} means that less propellant mass is needed to perform a given mission, so higher is generally better. The rocket equation, which relates propellant usage to mission requirements and specific impulse, and propulsion requirements for representative space missions are summarized in Chapter 2, Sec. 2.3.

Spacecraft thrusters are classified as either cold gas, chemical, electric, nuclear, or solar thermal. With the exception of some forms of electric propulsion (e.g., ion engines, Hall-effect thrusters, and magnetoplasmadynamic thrusters), most of these devices use a converging/diverging nozzle to expand propellant in a plenum at pressure P_1 and temperature T (so-called "stagnation conditions") to much lower ambient pressure P_2. The nozzle converts propellant enthalpy into directed kinetic energy and hence thrust; the propellant expands, accelerates, and cools while exiting the nozzle. The converging section accelerates the flow until the flow velocity reaches the local sound speed, at which point a diverging section is required for continued expansion. The theoretical specific impulse for these gas-dynamic thrusters is approximately given by:

$$I_{sp} \cong \left(\frac{1}{g_o}\right)\left\{\left[\frac{2kR'}{2k-1}\right]\left[\frac{T}{M}\right]\left[1-\left(\frac{P_2}{P_1}\right)^{(k-1)/k}\right]\right\}^{1/2} \tag{17.1}$$

where g_o is the gravitational acceleration at the Earth's surface, k is the ratio of specific heats for the propellant in the plenum (reaction chamber), R' is the universal gas constant (8.314 J/mol-K), M is the mean molecular weight of the exhaust gas, and P_2 is the pressure at the exit plane. Note that Eq. (17.1) is purely thermodynamic; physical scaling does not enter into the simple theoretical calculation of specific impulse.

Cold gas thrusters use nonreacting propellants at ambient temperature, so T in Eq. (17.1) is roughly the temperature of the stored propellant. Plenum temperature T can be increased, M can be decreased, and/or the ratio P_1/P_2 can be increased to boost specific impulse. In practice, T is usually increased using chemical reactions and/or electric, nuclear-powered, or focused-sunlight

heaters. Low M propellants such as hydrogen, water, or ammonia are usually used for externally powered (nonchemical) thrusters. Table 17.1 gives the theoretical maximum specific impulse for a number of gases with various stagnation temperatures based on an ideal expansion to zero pressure. The calculation for ammonia includes estimates of molecular dissociation into H_2 and N_2; dissociation levels for molecular propellants depend on maximum temperatures encountered and the catalytic behavior of materials in the flow path. Chapter 16 gives an overview of rocket propellant chemistry for various chemical propulsion systems.

Table 17.2 lists relevant properties and possible scaling issues for a number of thrusters that are used on orbit today. The first six are thermochemical or electrothermal devices, and their ideal

Table 17.1. Theoretical Maximum I_{sp} for an Ideal Expansion to Zero Pressure for Different Stagnation Temperatures.

Gas	273 K (s)	400 K (s)	600 K (s)	800 K (s)	1000 K (s)
Hydrogen	284	343	421	486	543
Helium	172	208	254	294	328
Methane	112	135	165	191	213
Ammonia[a]	107	130	167[b]	203[b]	244[b]
Nitrogen	77	93	114	132	147
Neon	77	93	114	131	146

[a]Ammonia includes dissociation effects at a stagnation pressure of 1 atm.
[b]Partial dissociation into H_2 and N_2 included

Table 17.2. Propellants, Specific Impulse, and Scaling Issues (to Small Size) for Typical On-Orbit Thrusters.

Thruster Type	Class	Specific Impulse (s)	Typical Propellants	Microscaling Issues
Cold gas	Chemical	40–80	Freons, N_2, Ar	Drag losses at low Reynolds numbers
Monopropellant	Chemical	180–220	N_2H_4, H_2O_2	Drag losses at low Reynolds numbers
Bipropellant	Chemical	300–450	N_2H_4+ N_2O_4, $H_2 + O_2$	Propellant mixing, high-flame temperatures
Solid	Chemical	100–290	Nitrocellulose + nitroglycerine	Ignition
Resistojet	Electric	150–330	H_2O, NH_3, N_2H_4	Drag losses at low Reynolds numbers
Arcjet	Electric	400–900	NH_3, N_2H_4	Electrode erosion
Hall effect thruster	Electric	1400–2000	Xe, Kr	Plasma containment
Ion engine	Electric	1600–5000	Xe, Kr	Plasma containment

specific impulse can be estimated using Eq. (17.1). All these devices will suffer from drag or viscous losses as internal Reynolds numbers decrease below 1500. The last two are electrostatic thrusters that use space-charge or externally applied electric fields to accelerate positive ions in a plasma to energies of several hundred electron volts (eV) and higher. An electron source is used for both species ionization and the subsequent charge neutralization of the ion beam.

Not all conventional thrusters can be scaled to small size or thrust level. Cold gas, monopropellant, solid, and resistojet thrusters are readily scalable to millinewton thrust levels albeit at a cost in performance as a result of viscous losses. Bipropellant thrusters with millinewton thrust levels can be built, but there are physical limits on how small combustion chambers can be. The chamber volume must be large enough to allow for evaporation (i.e., for liquid propellants), gas mixing, and complete combustion. Some combination of premixing propellants, use of highly reactive propellants, micromachined injectors, and high-pressure operation can be used to accelerate the combustion, but the fundamental bimolecular reaction rates dictate the necessary reaction zone and thus the volume.

Conventional Hall-effect thrusters and ion engines are also difficult to scale to small size because magnetic confinement does not scale linearly with size. An understanding of scaling issues related to plasma containment can be obtained by studying the change in the density of an instantaneously generated plasma that has no sources of electrons or ions. Plasmas constantly lose ions and electrons through diffusion across magnetic field lines, and most plasma devices are governed by the Bohm diffusion law,[12] where the diffusion coefficient D_B is given by:

$$D_B = \frac{KT_e}{16eB} \tag{17.2}$$

where K is Boltzmann's constant, T_e is the electron temperature, e is the charge of an electron, and B is magnetic field strength. For a plasma cylinder of radius R, the plasma density decays exponentially in time, where the time constant τ is given by

$$\tau = \frac{R^2}{2D_B} \tag{17.3}$$

Substituting Eq. (17.2) into Eq. (17.3) shows that the time constant is proportional to R^2B for a constant electron temperature. To maintain a constant decay time, the magnetic field strength must vary inversely with the square of radius. This is not a favorable scaling relationship as dimensions decrease and indicates that a nonplasma-type generator of ions may be required for microion engines and related electrostatic thrusters. Liquid-metal and field ionization sources do not require a plasma for proper operation.

The field emission electric propulsion (FEEP) thruster is a micronewton class electrostatic thruster that operates at specific impulses in the range of 6000–10,000+ s.[13] This thruster uses a high electric field perpendicular to a conducting fluid surface to induce fluid motion into sharp points and subsequent ionization at these points. In practice, FEEP thrusters are linear liquid-metal ion sources with cesium as the propellant and are manufactured using conventional machining techniques. An indium liquid-metal ion source was tested on the Japanese Geotail mission, so some flight experience with conventional liquid-metal ion sources is available.[14] The major advantage of these sources is that no propellant valves are required since capillary action provides propellant delivery. The disadvantage is that the propellant is a metal in a molten state. The possible contamination of spacecraft surfaces by metallic plume effluents drove conventional ion engine technology away from cesium and mercury propellants in the 1960s and 1970s to xenon by the 1980s.

The key to creating efficient field-ionization and field-emission-based ion engines is to generate electric field strengths on the order of 10^8 V/cm and higher using potentials on the order of 100 V. To generate these field strengths requires fabrication of submicron gaps between electrodes and/or fabrication of nanometer-scale edges for electric field enhancement. Field ionization of a gas-phase atom near the surface in a high electric field occurs through an induced electron tunneling process, where a bound electron in the atom has good probability of tunneling into the conduction band of the surface. The result is a positively charged gas-phase ion. The concept is a variant of the field-ion microscope, which is used to image surfaces with atomic resolution.[15] Equation (17.4) gives the probability P for tunneling of the electron in the atom, at a critical distance χ_c, through a maximum tunneling barrier at the surface:[16]

$$P(\chi_c,F)= Exp\left\{-\left[\frac{8m}{\hbar^2}\right]^{1/2}\frac{2}{3}[I-2(e^3F)^{1/2}]^{1/2}\left[\frac{I-\phi}{F}\right]\right\} \tag{17.4}$$

In Eq. (17.4), m is electron mass, e is the electron charge, I is the ionization potential of the atom, ϕ is the work function of the surface, and F is the applied field strength, usually given in units of volts/angstrom. Figure 17.1 shows the change in the electron barrier penetration probability as a function of field strength. The calculation shows that for greater than 2.5 V/Å (2.5×10^8 V/cm) the probability of field ionization is > 0.5 per electron scattering event. The ionization rate of an atom at χ_c is related to this penetration probability along with the electron scattering frequency and the gas flux impinging on the surface.

The gas flux impinging on the surface is not the simple gas kinetic flux usually given by $G = P_g/(2\pi MkT)^{1/2}$, where P_g is the gas pressure, M is the mass of the atom, k is the Boltzmann constant, and T is the temperature. In the high local electric field, the atoms are attracted to the surface by a polarization force given by $\alpha F(dF/dx)$, where α is the atom polarizability and x is the atom-surface distance. This attraction force enhances the number of atoms hitting the surface. The enhancement factor for a cylindrical emitter has been calculated and published.[17] Equation (17.5) gives the enhancement factor, E_f per unit length of emitter.

$$E_f = \frac{Z}{G(2\pi r_t)} = \left(\frac{2\alpha F^2}{\pi kT}\right)^{1/2} \tag{17.5}$$

For Xe gas[18] with a polarizability of 4×10^{-25} cm^3 and an electric field strength of 2×10^8 V/cm with $kT = 25$ meV (room temperature), the calculated enhancement factor is 1.6. The enhancement factor for a spherical emitter increases to 2.5 for the same conditions. Calculations show that the I-V characteristics of such a field-ion source emitter should increase exponentially with the applied field strength squared until the ionization rate exceeds the rate of near surface atom replenishment.

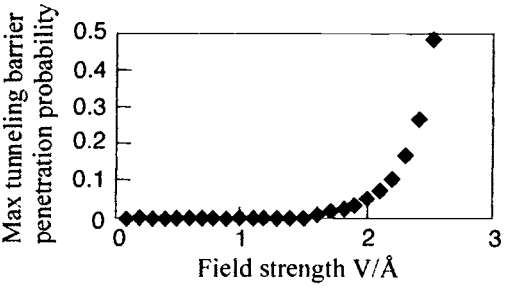


Fig. 17.1. Change in the electron barrier penetration probability as a function of field strength. For Xe over tungsten metal.

17.2 Micronozzles

Nozzles for gas-dynamic microthrusters have been fabricated using deep reactive ion etching (DRIE) of silicon or other materials, anisotropic etching of silicon, and laser-machining of glass/ceramic materials. An example of a DRIE etched nozzle is shown in Fig. 17.2, which shows scanning electron micrographs (SEM) of microfabricated supersonic nozzles for a space propulsion application.[19] Throat "diameters" varying from 12 to 30 μm have been fabricated with expansion ratios (exit plane area divided by the throat cross-sectional area) ranging from 5 to 20. DRIE technology allows very deep, extruded geometries to be fabricated (the nozzles above are 308 μm deep) while still maintaining excellent dimensional control. This fine dimensional control, such as the gradual contraction and expansion illustrated in Fig. 17.2, is essential for the fluid behavior to avoid separation, shock formation, and premature transition to turbulence. The DRIE technique permits fabrication of practical nozzle sizes with tapered dimensional control over two dimensions. With the DRIE technique the third dimension cannot easily be contoured and usually is a cylindrical extension of the two-dimensional (2D) pattern.

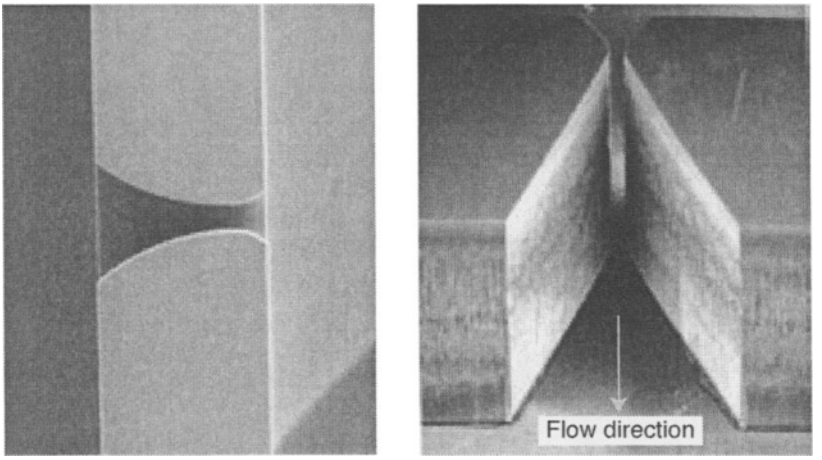


Fig. 17.2. SEMs of microfabricated supersonic nozzles for space propulsion applications, illustrating a top view of a trumpet nozzle (left) and an end-on view of a conical nozzle (right).

Three-dimensional axisymmetric nozzles are usually necessary to obtain the maximum efficiency. One approach is to use anisotropic etching of silicon, which can create pyramidal shaped pits that form converging or diverging nozzles of square or rectangular cross section. An alternative approach is to use laser-processing techniques that can fabricate true 3D axisymmetric nozzles of arbitrary contour in a variety of materials. Figure 17.3 shows two optical microscope views of a nozzle with a throat diameter of 100 μm and an expansion ratio of 10:1, fabricated in a photosensitive glass/ceramic material (Foturan™ manufactured by Schott Glassworks of Germany).[20] Glass/ceramic nozzles offer a "see-through" substrate and significantly reduce thermal conduction losses. The material is also considerably harder than silicon (for processed Foturan: Modulus of Rupture is 150 N/mm^2, and the Knoop hardness is 4600–5200 N/mm^2) and has good electrical isolation and no porosity. Furthermore, the material-processing approach allows the design and fabrication of the nozzle shapes and expansion ratios to be under user control.

Figure 17.4 shows experimental thrust measurements for an anisotropically etched converging-diverging silicon nozzle that was part of the pre-1996 EG&G IC Sensors Corp. Model 4425-15 MEMS microvalve (newer versions use a more complicated geometry that precludes their use

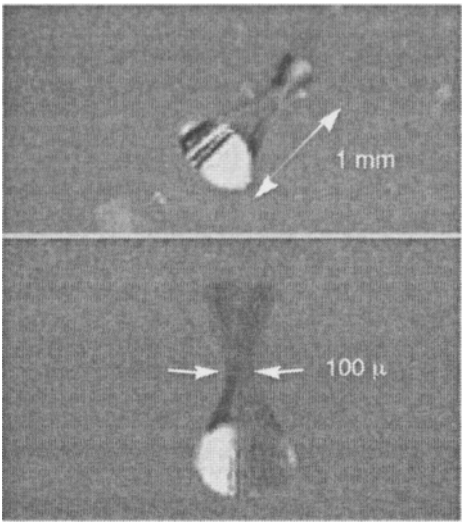


Fig. 17.3. Optical microphotographs of a laser-machined converging/diverging nozzle in a transparent photo-ceramic material. The fabrication effort was done from one side only.

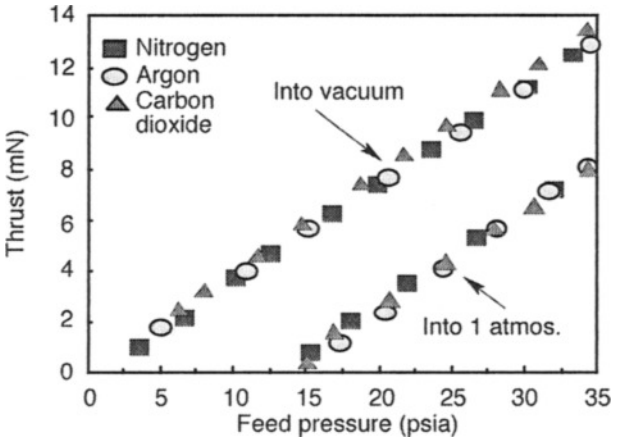


Fig. 17.4. Thrust vs feed pressure for the exit nozzle in the EG&G IC Sensors Model 4425-15 microvalve. The ambient vacuum pressure is between 0.1 and 0.2 torr. Thrust errors are ±0.2 mN, and feed pressure errors are ±0.2 psia.

as thruster nozzles). The nozzle has a 200-μm-sq throat and a 10:1 area expansion ratio. The exit velocity is a function of expansion ratio, mean molecular weight of the propellant, ratio of specific heats of the propellant, and stagnation temperature. Theoretical exit velocity is shown in Fig. 17.5 for nitrogen, argon, and carbon dioxide with a stagnation temperature (the temperature of the propellant before expansion) of 300 K. The ratio of exit-plane pressure to stagnation pressure can also be calculated as a function of area ratio and is shown in Fig. 17.6 for an ideal gas with specific heat ratio of 1.2 (some polyatomic gases), 1.4 (diatomic gases), and 1.67 (monatomic gases). Figure 17.6 assumes a stagnation temperature of 300 K. Note that an area ratio of 10 or greater is required to approach the maximum exit velocity, and these area ratios require that the exit pressure be less than 1% of the stagnation pressure.

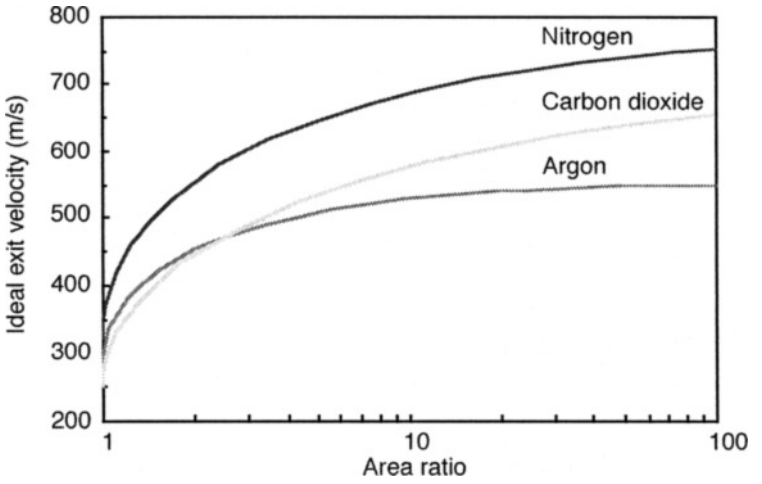


Fig. 17.5. Ideal (one-dimensional [1D] isentropic expansion) exit velocity as a function of area ratio for nitrogen and argon propellants initially at 300 K. This assumes that the exit pressure matches the ambient pressure.

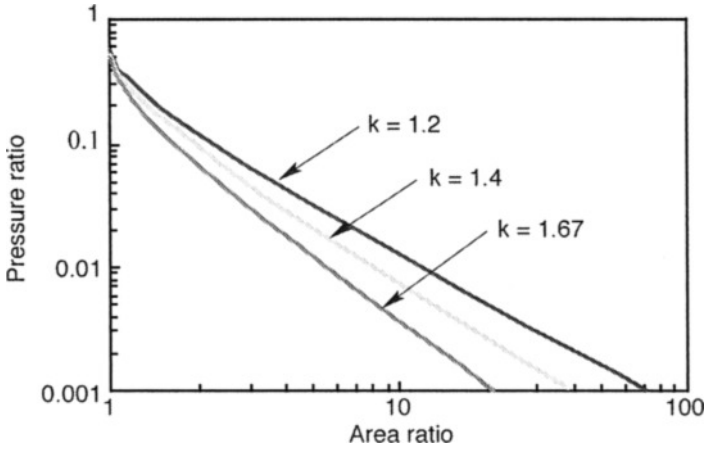


Fig. 17.6.Pressure ratio as a function of area ratio for the 1D, supersonic, isentropic expansion of an ideal gas with specific heat ratios of 1.2, 1.4, and 1.67.

Thrust can be calculated using

$$F = V_2\dot{m} + p_2A_2 \tag{17.6}$$

where V_2 is the exit velocity, $\dot{m}$ is the propellant mass flow rate, p_2 is the pressure at the exit plane, and A_2 is the nozzle cross-sectional area at the exit plane. Equation (17.6) is valid for expansion into a perfect vacuum; this is approximately valid for ambient pressures below 0.2 torr and expansion ratios below 10. By using measured or estimated mass flow rates, the calculated exit-plane velocities given in Fig. 17.5, the calculated exit-plane pressures given in Fig. 17.6, and the geometric area ratios for various thrusters, one can calculate the expected thrust level. Dividing this thrust by the mass flow rate will produce the ideal specific impulse. For cold gas thrusters operating at about 300 K, the gases nitrogen, argon, and carbon dioxide produce specific impulses of about 70 s, 50 s, and 60 s, respectively.

Table 17.3 presents the ideal I_{sp}, the measured I_{sp}, the percent of ideal I_{sp}, the discharge coefficient C_D, the thrust coefficient C_F, and the ideal thrust coefficient for the silicon microvalve nozzle expanding into a ~ 0.2 torr vacuum. Discharge coefficient C_D is defined as the actual (measured) mass flow rate divided by the ideal (theoretical) mass flow rate. Thrust coefficient C_F is defined as

$$C_F = \frac{F}{(A_t P_o)}, \tag{17.7}$$

where A_t is the cross-sectional area of the throat and P_o is stagnation or feed pressure (~35 psia for Table 17.3).

Table 17.3. Performance Characteristics of the Batch-Fabricated Silicon Nozzle. These Results are for ~0.2 torr Ambient Pressure.

Gas	Expansion Ratio	Ideal I_{sp} (s)	Actual I_{sp} (s)	Percent of Ideal I_{sp}	C_D	C_F	C_F Ideal
Nitrogen	10:1	73	56	77	0.96	1.21	1.65
Argon	10:1	54	45	83	0.94	1.22	1.57
Carbon dioxide	10:1	61	47	77	1.00	1.28	1.70

One question associated with small fluidic nozzles is the effect of the small scale on the flow. Note the reduction in measured specific impulse and thrust coefficient compared to ideal values in Table 17.3. Also, as one would expect, device performance is degraded as a result of viscous losses. This is illustrated in Fig. 17.7, which shows experimentally measured and numerically predicted values of the mass flow and thrust efficiencies versus Reynolds numbers for nozzles similar to those shown in Fig. 17.2. The results indicate that the mass flow maintains a high efficiency, even at low Reynolds numbers, while the thrust (indicated by the I_{sp} efficiency) drops off dramatically below a Reynolds number of approximately 1500. This discrepancy is explained by the fact that the mass flow is set by the choked (sonic) conditions at the throat where the boundary

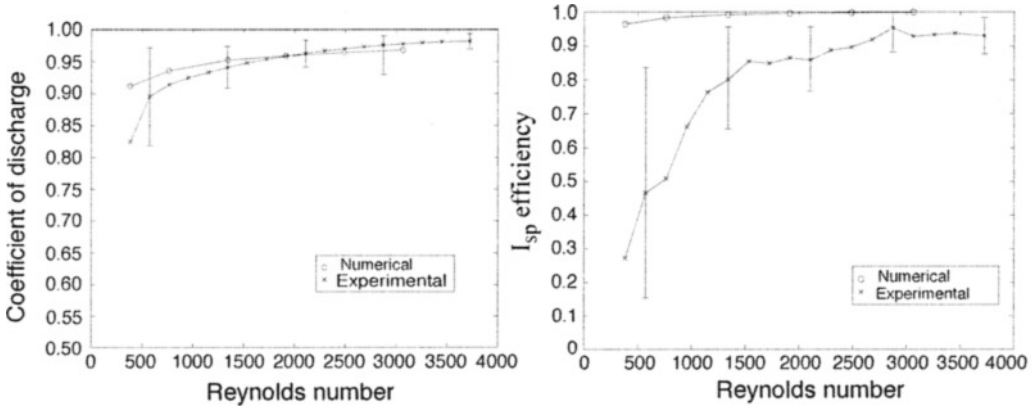


Fig. 17.7.Calculated and measured coefficient of discharge (left) and I_{sp} efficiency (right) vs Reynolds number for a 2D micronozzle with a throat diameter of 37.5 μm, an area ratio of 17:1, and a depth of 308 μm. (Citation used with permission[21]).

layers are still thin and viscous effects have not had a chance to degrade performance. In contrast, the thrust (and thus, I_{sp}) is set at the exit of the nozzle where the boundary layers have grown substantially. Indeed, at the low Reynolds numbers, it is doubtful that supersonic flow is still present. The numerical results, based on 2D compressible Navier-Stokes computations[21] show good agreement with the mass flow (again, because this is set by the throat conditions). The I_{sp} performance agrees less well because of the contamination by the upper and lower walls, which were not modeled in the CFD (continuum fluid dynamics) computations.[21]

17.2.1 Microfluid Modeling

The example of the micromachined nozzles shown in Fig. 17.2 illustrates the crucial need for accurate modeling of the fluid behavior in micromachined geometries. The small dimensions of the system imply that the devices operate in a somewhat unfamiliar parameter regime. The most obvious example is the pervasive low-Reynolds number, which for flows such as nozzle flows and turbomachinery is quite different from its counterpart in macroscopic machinery.

In addition to the relatively low Reynolds number, most applications require that the flow velocity be as high as possible in order to extract as much useful work from the fluid as possible. This combination of high velocity (or Mach number *Ma*) and low Reynolds number (because of the small scale) places the MEMS device in a regime usually found in low-pressure, high-altitude devices, where gas rarefaction can occur as a result of elevated Knudsen numbers, (*Kn*) where

$$Kn = \frac{\ell}{H} \tag{17.8}$$

ℓ is the typical mean-free-path of the gas, and H is the characteristic device length scale. For ideal gases in thermodynamic equilibrium, the Reynolds, Mach, and Knudsen numbers are related by

$$Kn \propto \frac{Ma}{Re} \tag{17.9}$$

which illustrates that, for even moderate values of the Mach number, the low Reynolds number dictated by the small scale can result in appreciable Knudsen numbers in MEMS devices. As an example, air in a 1-μm gap at standard temperature and pressure corresponds to a Knudsen number of approximately 0.07. For *Kn* below 0.01, standard continuum fluid models are quite sufficient. However, as the *Kn* rises, the first nonequilibrium effect that needs to be considered is the effect of the slip layer at a solid surface, which results in nonzero velocities (or a slip flow) at wall boundaries. In the case of a nonisothermal flow, there is also a temperature jump between the surface and the fluid. These are quite familiar effects in any rarefied flow and can be incorporated into continuum (i.e., Navier-Stokes) computations by simply replacing the standard "no-slip" boundary conditions at the wall with a Maxwellian "slip-flow" boundary condition for the tangential velocity. Equation 17.10 describes this, where u is the tangential velocity, σ is the coefficient of tangential momentum accommodation, *Kn* is the Knudsen number, and y is the distance from the wall.

$$u|_{y=0} = \frac{2-\sigma}{\sigma} Kn \left.\frac{\partial u}{\partial y}\right|_{y=0} \tag{17.10}$$

Other, higher-order boundary conditions have also been proposed,[22] although available experiment results in microchannels at small to moderate Knudsen numbers indicate that the first-order condition appears sufficient.[23,24,25]

Although the effects of high Knudsen number are at first glance no different from conventional (high-altitude, low-pressure) rarefied gas dynamics, MEMS devices have two aspects that are unique to the small-dimensions: the effects of the surface roughness and surface accommodation.

Unlike standard low-pressure rarefied flows, the surface conditions in MEMS can vary considerably from atomically smooth surfaces (in the case of crystalline silicon surfaces) to surfaces with appreciable roughness. These surface conditions are highly process-dependent and may well have a significant effect on the overall flow in the device. In particular, for rarefied flows, the momentum and energy accommodation coefficients may not be unity (as is usually assumed in macroscopic flows). Recent experiments of flows in micromachined channels have explored this[25] and have found that in addition to the expected gas rarefaction, the tangential accommodation coefficient between the gas and the silicon substrate was measured to be approximately 0.8. Although this is not a huge departure from the standard engineering value of 1, it is significant, particularly when careful simulations of viscous effects on the performance of MEMS devices are required. This is currently a subject of active investigation.

As the Knudsen number increases further, modified Navier-Stokes equations are no longer appropriate, and molecular-based computations become necessary. The most popular approach is the Direct Simulation Monte Carlo (DSMC) technique, developed and popularized by Bird[26] in the context of high-atmospheric and hypersonic flows. The application of DSMC to MEMS flows is straightforward and has been demonstrated in a number of relatively simple geometries such as long 2D channels.[22,27] Some complications peculiar to the simulation of MEMS devices do naturally arise. Unlike typical hypersonic and reentry problems, MEMS devices operating at low or moderate Mach numbers are characterized by large molecular thermal velocities compared with typical molecular drift velocities. This means that random statistical variations that are inherent in DSMC (and other molecular-based numerical schemes) are large compared with the flow velocities, and thus long averaging times are required to obtain statistically converged simulations (recall that the average of a normally distributed random variable converges as $N^{-1/2}$, where N is the number of samples). This makes simulations of flows in MEMS very computationally intensive and to be avoided unless the flow regime absolutely requires this approach. A second difficulty that arises with low-Mach-number DSMC computations is the treatment of inflow-outflow boundary conditions. At high Mach numbers, molecules entering the simulation domain can be assumed to be independent of the flow conditions inside the domain. Similarly, molecules leaving the domain at high Mach numbers do so with effectively no regard to the flow conditions downstream (i.e., outside the computational domain). At low Mach numbers, however, the propagation characteristics of the flow change and the DSMC boundary conditions must be modified to account for this. An approach based on CFD-style Riemann characteristics has been successfully demonstrated.[27]

Clearly, the most promising approach in the future is to use molecular schemes, such as DSMC, only in the small regions of the flow where they are absolutely necessary, and to use continuum-based schemes in the regions of the flow where the Knudsen number is acceptably low. As with the low-Mach-number boundary conditions, the matching of these two approaches presents some challenges.[28]

17.3 Fluid Actuators

17.3.1 Synthetic Jet Actuators

One simple illustration of solid-state propulsion suitable for manufacture using MEMS technology is the "synthetic jet," illustrated in Fig. 17.8. This device, first presented by Coe *et al.*,[29,30] is based on a simple concept of pumping fluid in an oscillatory motion and taking advantage of the nonreversible nature of the cycle introduced by the fluid viscosity. The actuator is composed of a cavity with a small hole; the volume of the cavity is modulated in an oscillatory fashion. In principle, any mechanical device could be used to drive the cavity volume, although in practical

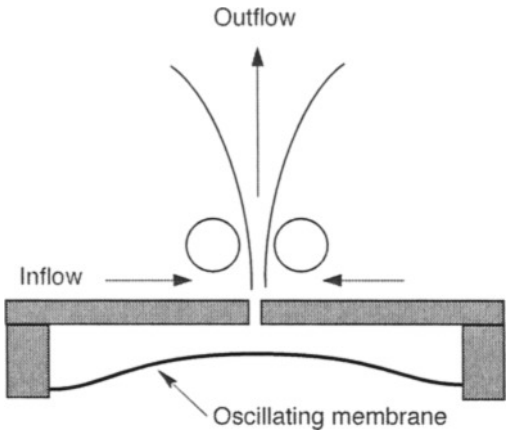


Fig. 17.8. Schematic of a zero-net-mass-flux jet (or "synthetic jet").

applications (particularly MEMS-fabricated devices), a vibrating membrane or a cantilever beam operating at its resonant frequency is used. By operating the structure at its resonant frequency, moderate driving amplitudes can be achieved (1–5 μm in a MEMS device).

During the "downstroke" of the membrane, the cavity volume increases, sucking fluid in from the surroundings. To the fluid outside the cavity, this appears like a point sink, and the air is drawn in uniformly from all directions. On the membrane's "upstroke," the volume of the cavity decreases, and the resultant rise in pressure squirts the fluid through the hole. However, the flow separates at the sharp etch of the hole (assuming the Reynolds number is greater than about 50), and in contrast to the downstroke, the fluid is thus directed upward in a concentrated jet of fluid. The net effect, integrated over many cycles of the membrane motion, is that fluid is drawn in from the sides with low momentum and expelled upward with high momentum.

The concept of the fluid rectifying an oscillatory mechanical motion through viscosity is not new and was first recognized by Ingard[31] in the context of high-amplitude acoustics. In recent years it has been rediscovered and used to great effect in flow-control applications.[32–35] It has a natural appeal in MEMS devices because the mechanical motion is easy to achieve and can be driven at high frequency. Jet velocities of 1–20 m/s have been reported (depending on the fluid and the device size). It should be noted, however, that as a whole, the device is not very efficient since the total momentum imparted to the jet is small compared with the energy expended in the unsteady acceleration of the fluid in and out of the actuator cavity.

Most resonant jet devices presented have relied on a trial-and-error design procedure, and researchers have noted that the device performance varies wildly when new actuator dimensions or materials are used.[36] Accurate modeling and design of such devices require a coupling between the structural characteristics of the actuator membrane and the fluid in the cavity and exit hole.[35]

The membrane dynamics are straightforward and are controlled by the membrane dimensions (diameter, thickness) and material (Young's modulus and Poison Ratio). With these specified, the (unloaded) resonant frequency of the device is easily determined according to classical theory. The membrane motion drives the fluid in the cavity in a number of possible ways. It can compress the fluid, adding a compression stiffness to the membrane (raising its resonance frequency). In addition, the unsteady acceleration of fluid in and out of the actuator hole results in a virtual mass and a damping term. All three of these couple to the membrane dynamics and can result in changed resonant frequency and damping ratios for the system.

Fabrication of synthetic jets using MEMS techniques has two attractions. First, device uniformity is high, and losses associated with mechanical inefficiencies are minimized by the typically high-Q performance of MEMS devices (particularly if the device is manufactured using single crystal silicon as its mechanical material). Second, MEMS can be used to construct arrays of actuators to increase the thrust of the system.

17.3.2 Micromachined Arrays of Synthetic Jets

Arrays of synthetic jets have been demonstrated at Georgia Institute of Technology (Georgia Tech), which has pioneered much of the synthetic jet work. In addition to showing how two or more jets can be put together, Georgia Tech also showed that the character of the overall jet can be modified by altering the phase of the driving voltage between adjacent devices. This is illustrated in Fig. 17.9, which shows how by varying the phase between two jets, one can vector the resultant flow as much as 90 deg away from the primary jet direction.

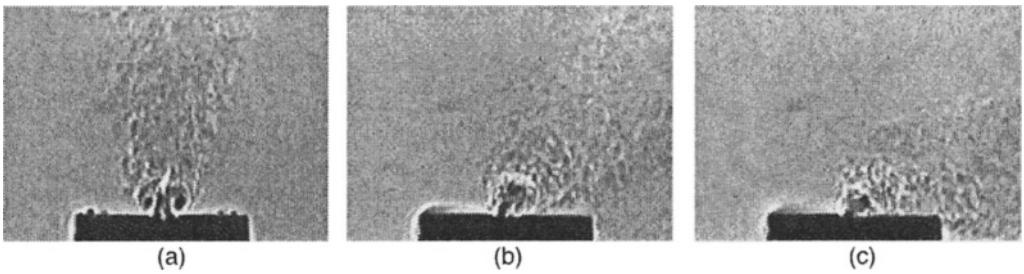


Fig. 17.9. Two synthetic jets placed side by side. The forcing signal is (a) in phase, (b) 60-deg shifted, and (c) 150-deg shifted. Note how the direction of the resultant jet varies with the phase shift (courtesy B. Smith and A. Glezer. © *Physics of Fluids*[32]).

17.3.3 Micromachined Actuator Disks Using Synthetic Jets

From the example in Fig. 17.9, it is clear that close arrangements of synthetic jets can result in complex interactions between the jet flows. In particular, it is easy to see that in a large array, actuators in the interior of the array can become "starved" for air as they are competing for air with their adjacent actuators. This starvation can be alleviated by providing "feed holes" between each device to supply low-momentum fluid to each actuator without stealing it from an adjacent device. This concept is demonstrated in Fig. 17.10, which shows a cross section of an actuator disk constructed using arrays of resonant jet actuators. A photograph of a micromachined actuator disk

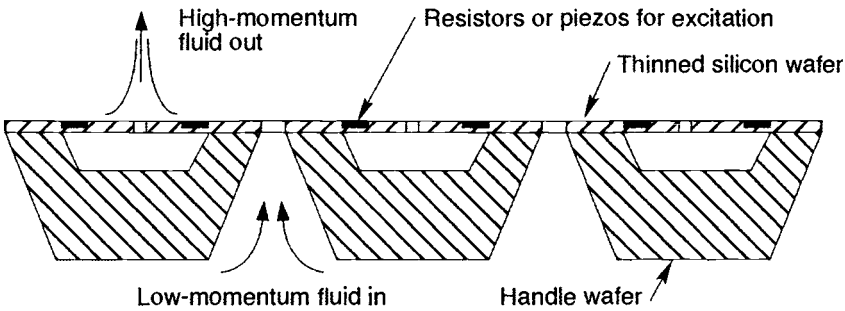


Fig. 17.10. Schematic of an array of synthetic jets with feed holes forming a solid-state actuator disk. The feed holes supply fluid from the bottom of the disk to the actuators, which propel the fluid with high momentum upward.

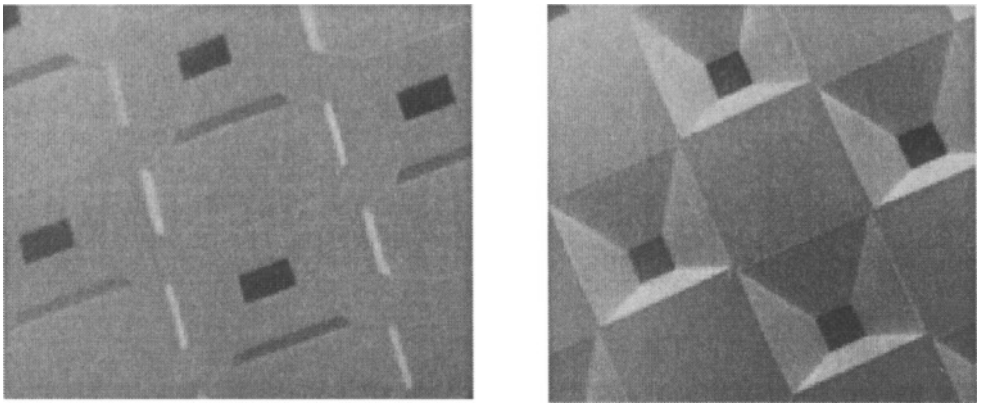


Fig. 17.11. SEMS of micromachined actuator disk. Left photograph shows view of the top-side without the cover plate, illustrating the cavities and feed holes from the back side. Right photograph shows view from below, showing the anisotropically etched feed holes.[37]

is shown in Fig. 17.11. Here the feed holes are fabricated using a regular array of anisotropically etched holes. The actuator cavities are shallow-etched on the front side and then covered with the actuator membrane (using a wafer-bonding and etch-back process). The membrane can be driven by any means, either electrostatically or thermally, or with active materials such as piezoceramics.

As shown, this device is a true actuator disk (a solid-state propeller)—it draws in low-momentum fluid from one side and blows out high-momentum fluid on the other side. As with other MEMS propulsion devices, the scaling physics favor miniaturization since the thrust scales with the active area (L^2), while the weight scales with the volume (L^3). Thus the thrust-to-weight ratio should increase as 1/L as the device shrinks. Of course, inefficiencies, particularly due to viscous effects in the hole, will serve to degrade the device performance as it shrinks.

17.3.4 Micro-Jet Engines

An ambitious and aggressive example of micromachined actuators for both propulsion and energy conversion is illustrated in Fig. 17.12, which shows the cross section of the MIT microengine, a complete gas turbine on a chip.[38] This device, currently in development, is composed of a radial compressor, combustion chamber, and radial turbine. Detailed studies indicate that the device, using hydrogen as its fuel, will produce 0.2 N thrust, with a thrust-to-weight ratio comparable to that found in conventionally sized gas turbine engines. The addition of an electrostatic induction generator (mounted on the compressor shroud) converts the device into a turbogenerator, estimated to generate approximately 20 W of electrical power. A SEM of the turbine is shown in Fig. 17.13, illustrating the complex geometric design of the device. Note that the gap between the stator and rotor blades is approximately 10 μm wide and 300 μm deep. This forms the gas-lubricated journal bearing that supports the rotor (in the radial direction) as it rotates (at 2.4 million RPM). The high-aspect ratio of this device, machined using reactive ion etching (RIE), is an example of the rapidly advancing state of MEMS technology, which has enabled the design of new high-performance micromachined devices that would not have been possible only 3 years ago.

17.4 Laser-Based Processing of Glass/Ceramic Materials

This section briefly presents the basic steps of a nonsilicon processing technique developed at Aerospace for microthruster applications. Development of this technique was dictated by the need for true 3D microstructures in materials as hard or harder than silicon. Details of the process chemistry can be found in Chapter 5.

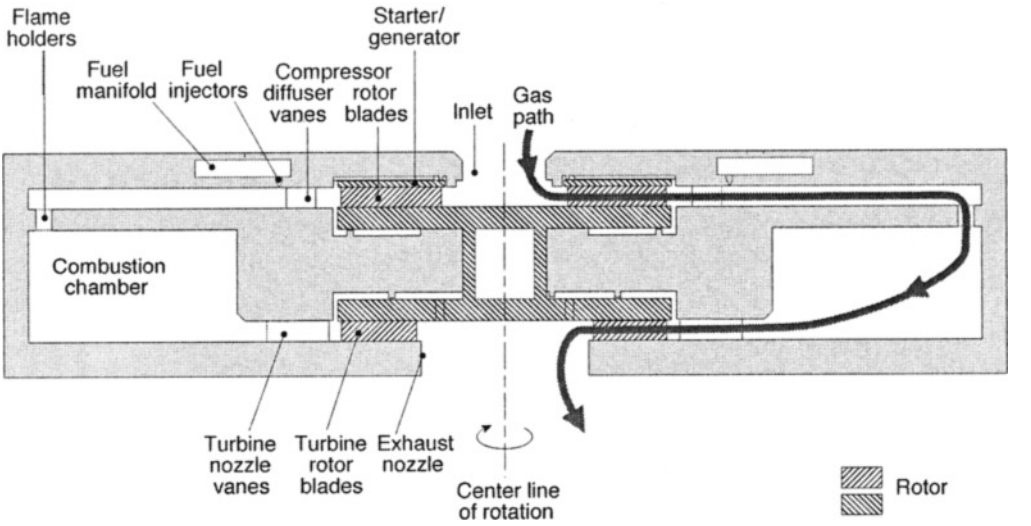


Fig. 17.12. Schematic of the MIT micromachined gas turbine engine. The overall dimension of the device is 1 cm. The shaded portion contains the rotating components and includes a radial outflow compressor, radial inflow turbine, and a motor/generator (mounted above the compressor).[38,39]

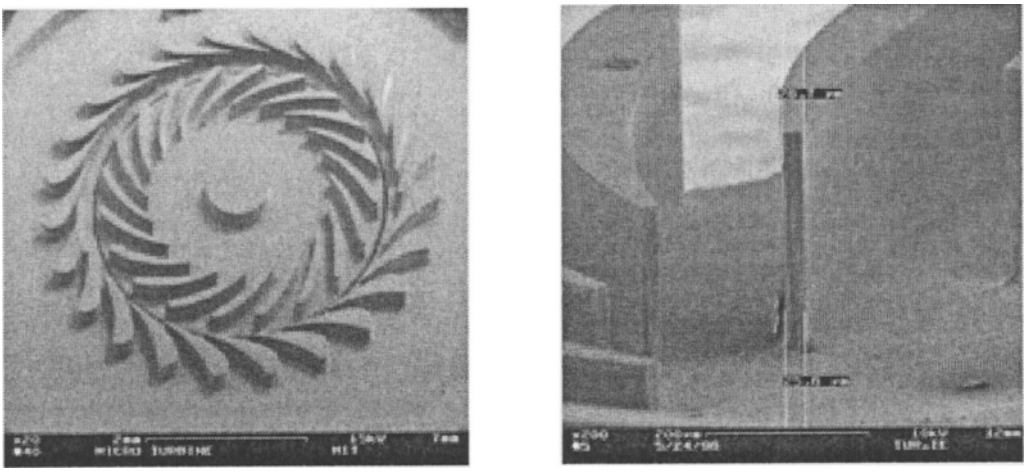


Fig. 17.13. SEMs of the MIT microengine, showing the 100-W (mechanical power that is used to drive the compressor and power generator and to overcome system friction) turbine (left) and a close-up of the trailing edge of the turbine rotor (right). The blades are 300 μm high, and the piece is fabricated using RIE.

Photocerams, photositalls, and pyrocerams are a class of materials with a long heritage similar to silicon, but with less microfabrication-processing development. These interesting materials can undergo controlled devitrification, resulting in the formation of microcrystalline structures[40] and can best be described as glass that can be converted to tough ceramic by high-temperature baking. They differ from glass in that they are basically crystalline, and differ from true ceramic in their much smaller grain size. There are nearly 5000 formulations of this class of materials with properties advantageous to microthruster development. Certain formulations can have specific physical properties (Knoop hardness 4600 N/mm^2, Young's modulus 7.8 × 104 N/mm^2, Poisson's

ratio = 0.22), can be made nonporous, and can be "engineered" to sustain wide thermal cycling (~900° C). Because the material is essentially a glass, it also has a strong resistance to chemical attack. An additional advantage is that to some degree the material strength and brittleness can be tailored by controlling the ceramization phase (i.e., the temperature bake). However, its major feature is that it can be patterned by masking and broad-area exposure to ultraviolet (UV) light (i.e., creating 2D+ (2D outline to a variable depth)) or by direct-write focused UV laser exposure (i.e., creating true 3D shapes).[41] The ability to transfer a 2D or 3D image/pattern in the material is made possible by a single photon excitation photolytic process. A photon with energy in the UV ($\lambda < 360$ nm) induces a charge transfer and neutralization between two stabilized metallic-ion species. Equations (17.11)–(17.13) present the essential chemical details of the patterning in one formulation of this photoceram material, Foturan.

$$\text{Illumination: } Ce^{3+} + Ag^{+} >> Ce^{4+} + Ag^{o} \tag{17.11}$$

$$\text{Ceramization: } nAg^{o} >> (Ag^{o})_{n} - Li_2SiO_3 \tag{17.12}$$

$$\text{Isotropic etching: } Li_2SiO_3 + 3HF >> 2LiF + H_2SiF_6 + 3H_2O \tag{17.13}$$

Figure 17.14. shows the three-step process used to fabricate microstructures in Foturan. In the first step, a specific volume of material is exposed to UV light. The depth of exposure is a strong function of wavelength: 355 nm light goes completely through a 1 mm-thick wafer, while 248 nm light penetrates a few hundred microns. For the second step, the material is baked using a specific time-temperature profile. This causes the previously exposed regions to form crystals that are much more sensitive to hydrofluoric acid attack than the unexposed material. The exposed and baked material turns brown, while the unexposed and baked material remains clear. In the third step, the exposed regions are preferentially etched in a 40° C, 5% HF solution. Figure 17.15 shows an example 3D microstructure fabricated by the three-step process shown in Fig. 17.14. The two SEMs show the same array of micropyramids from two perspectives.

Other processing and material attributes of Foturan include

- The processed material is stronger. Preliminary tests have shown that the modulus of rupture (MOR) for Foturan increases by 63% upon partial ceramization (98 N/mm^2). The vendor[42] gives the MOR for a fully ceramicized part as 150 N/mm^2. However, for microthruster applications, partial ceramization (i.e., ceramization/crystallization within a "glass" amorphous host) has some advantages such as resistance to mechanical shock.
- Foturan can be metallized, and at Aerospace we have successfully deposited gold and platinum by standard radio-frequency sputtering and laser direct-write chemical-vapor-deposition techniques. In both cases a surface preparation process is required to get the metal to adhere.
- Tall microstructures (~1 mm) can be fabricated using the laser processing technique at 355-nm wavelength.

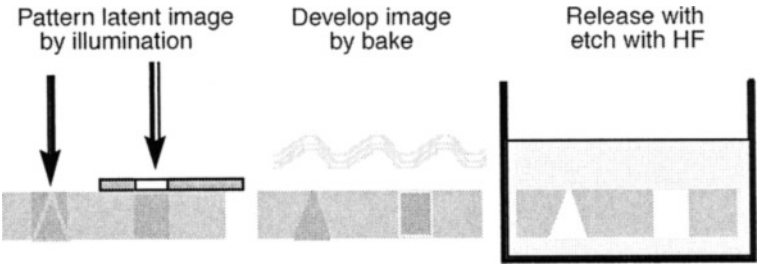


Fig. 17.14. The three-step Foturan patterning process.

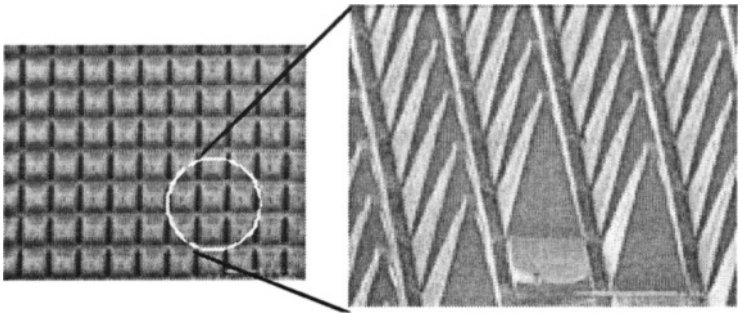


Fig. 17.15. SEM of array of glass/ceramic pyramids. The white bars at the lower right are 100 μm wide.

- The surface texture/finish of the fabricated microstructures can be somewhat controlled from a 1–5 μm smooth finish (smallest grain size possible in Foturan) to a multifaceted scalloped surface. Figure 17.16 shows an array of spires displaying a faceted surface texture.
- Processed Foturan wafers/coupons can be fusion-bonded to each other using a low-bulk temperature process (< 350 C) developed at Aerospace. Initial experiments show that the fusion bond can withstand pressures of 1.035 MPa (over a 1-mm-diam area). Actual fusion bond strength measurements are now under way. The ability to bond multiple wafers permits the assembly of very complex fluidic devices. Internal feed lines, actuators, and electronics all can be incorporated within a stacked-layer assembly.

17.5 Microthruster Fabrication

Microthrusters typically have mesoscopic (mm-scale) dimensions and microscopic (micron-scale) structures. Some specialized thrusters such as field-ionization and field-emission-based ion engines may even require nanoscopic (nanometer-scale) structures. The required accuracy and the large dynamic range in dimension over which a material must be processed place a constraint on the fabrication tools used in the development of an *integrated* microthruster. The tools must be able to process material over a large dynamic range of scale. For microthruster systems, propellant tank sizes are usually in the centimeter range, while valve openings and field-emission or ionization sources must be fabricated with micron or submicron accuracy. Cofabrication of propellant tanks and micromachined ion engines can impose complex constraints on the accuracy of the fabrication tools employed. In general, integration or cofabrication becomes easier when the number of materials utilized is limited, as in the case of the silicon-based complex microjet engine

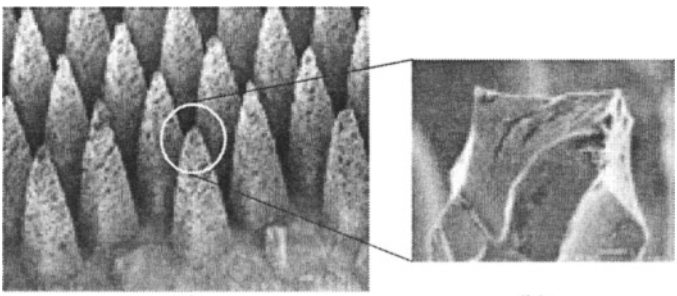


Fig. 17.16. SEM of array of spires. The white bars are scales for the SEMs: (a) 100 μm, (b) 1 μm.

shown in Fig. 17.13. But it is not always possible to do this and maintain a high level of performance in the components. Micromachined valve seats that incorporate polymeric or pliant metals provide much lower leak rates than silicon-silicon or silicon-glass sealing surfaces; ceramics can provide much higher operating temperatures and chemical resistance to exotic propellants in comparison with silicon.

17.5.1 Cold Gas Microthruster Fabrication

Four types of microthrusters are under development at Aerospace: cold gas microthrusters that utilize MEMS valves, arrays of single-shot solid fuel microthrusters for a Defense Advanced Research Projects Agency (DARPA)-sponsored "digital propulsion" program, complementary metal oxide semiconductor (CMOS)-fabrication-compatible microresistojets, and field-emission-based ion engines. In all cases Foturan plays a prominent role.

Figure 17.17 is a photograph of a bidirectional cold gas thruster with a partial complement of the piece parts: the MEMS valve die (EG&G IC Sensors model no. 4425-15), a Foturan ceramic plenum chamber, and a 3D axis-symmetric micronozzle with a 10:1 nozzle area expansion ratio. Using a more integrated packaging scheme, the microthruster in Fig. 17.17 can be reduced in size by a factor of 2 to 4. Key aspects of this thruster are a gas plenum 1 mm deep by 36 mm^2, the use of two silicon microvalves (one for each direction) in this plenum, and the use of laser-machined micronozzles previously shown in Fig. 17.3. The hour-glass nozzle shape is 1 mm thick (~100 μm diam throat) and is fabricated from one side only, without trepanning, by exposure to several pulses from a 355-nm-wavelength laser.

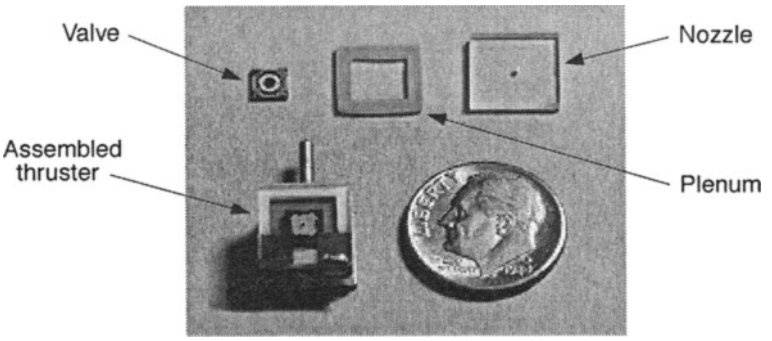


Fig. 17.17. A 1-mN bidirectional thruster module and its major components.

To fabricate an efficient nozzle, the hour-glass waist must be precisely positioned within the 1-mm-thick sample. This process entails more than just mechanically setting the focal point of a laser beam into the sample; it requires knowledge about the material optical index and its effect on focal position. Figure 17.18 schematically shows the expected change in the position of the waist as a consequence of focusing from a medium of index N1(λ) (air) into a medium with index N2(λ), where it is assumed N2 is greater than N1. The inset in the figure also displays the pertinent relationships. Figure 17.19 shows the calculated change in the distance to focus into the medium as a function of the distance of the focusing element (i.e., microscope objective) above the sample. More crucial to the focal position alignment is the fact that the ultimate focal position also depends on the input aperture setting as well (variable "a" in Fig. 17.18). Figure 17.20 shows this nonlinear dependence on the "error" or change in focus as a function of input aperture size. Other properties of the focal region, such as the waist radius (i.e., throat dimension), the depth of field

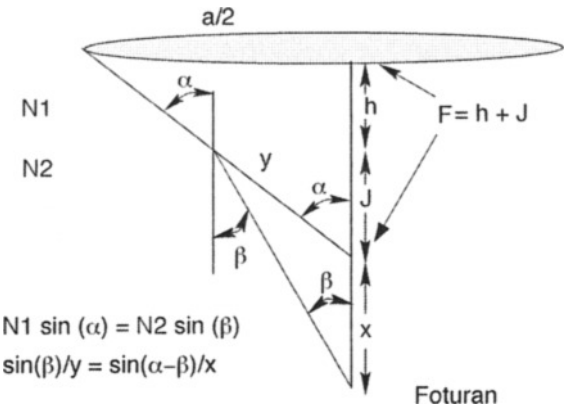


Fig. 17.18. Optical schematic showing the change in the focal position as a result of focusing from a medium of lower index into a medium of higher index.

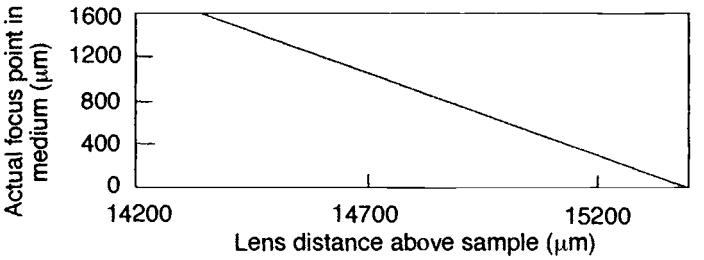


Fig. 17.19. Calculation showing the change in the focus distance within the higher index medium as a function of the microscope objective distance above the surface. (y = –1.5174x + 23348)

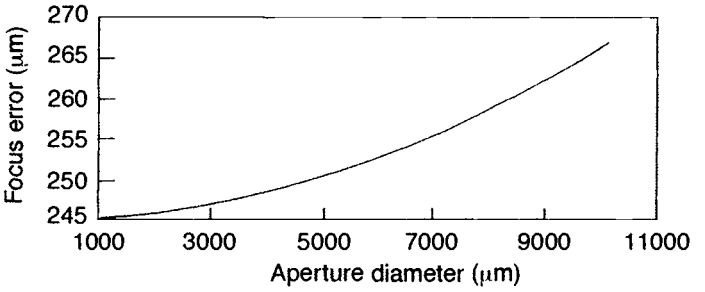


Fig. 17.20. Calculation showing the change in the focal position within a medium of index n >1 as a function of the incident aperture diameter.

(i.e., length of throat region), and the optical divergence that sets the exit-hole diameter can be precisely defined for a Gaussian optical wave traversing a diffraction limited optical setup. A Gaussian beam can be described by the radius function $w(z)$ and the wavefront curvature function $R(z)$ along the propagation direction z. The functions $w(z)$ and $R(z)$ are given by the Eqs. (17.14)–(17.16).[16]

$$\omega^2(z) = \omega_o^2\left[1 + \left(\frac{2z}{b}\right)^2\right] \tag{17.14}$$

$$R(z) = z\left[1 + \left(\frac{b}{2z}\right)^2\right] \tag{17.15}$$

$$b = \frac{2\pi\omega_o^2}{\lambda} \tag{17.16}$$

where λ is the wavelength, ω_o is the beam radius at the waist, and b as defined in Eq. (17.16) is the confocal parameter (i.e., distance within which the diameter of a focused beam remains nearly constant: $-b/2 < z < +b/2$). The Gaussian beam contracts to a minimum diameter $2\,\omega_o$ at the beam waist where the phase front is a plane wave. At large distances from the beam waist, the radius/diameter of the beam expands linearly with distance. Eq. (17.17) shows this dependence, and one can deduce a constant divergence cone angle given in Eq. (17.18).

$$\omega(z) \sim \frac{\lambda z}{\pi\omega_o} \tag{17.17}$$

$$\theta \sim \frac{\lambda}{\pi\omega_o} \tag{17.18}$$

As a rule, it is possible to fix the expansion ratio of a micronozzle fabricated by the laser exposure technique. The throat diameter and exit nozzle diameter are fixed to the required specifications, using Eqs. (17.17) and (17.18), and the length of the nozzle is used as the accommodating variable. In general a diffraction-limited optical setup is difficult and costly to achieve, but one can generate Gaussian beams and can use their properties to advantage. The Aerospace laser-processing facility uses a diode-seeded Nd-Yag solid-state laser system with a focusing capability approaching 0.5 μm at 266 nm (4th harmonic of Nd-Yag lasing).

The performance of cold gas thruster systems depends not only on nozzle design but also on propellant choice and thruster valve operating characteristics. Several MEMS valves are commercially available and can be integrated into low flow-rate cold-gas, chemical, and electric thrusters. In the past, these devices were electrothermally actuated and used silicon-glass or silicon-silicon sealing surfaces. As a result, operating powers were in the several-hundred to several-thousand milliwatt range, response times were in the tenths-to-1 second range, and leak rates were about 0.02 sccm (standard cubic centimeters per minute) and higher. Maximum flow rates were about 1500 sccm and lower. While leak rates of 0.02 sccm would be unacceptable for continuously operating long-term missions (i.e., geosynchronous communications satellites and interplanetary missions), they would be perfectly adequate for short-term missions ranging from a few days to a few weeks in duration. The use of elastomeric seals in the current generation of valves (e.g., Redwood Microsystems Corp.) has significantly decreased leak rates to the point where year-long missions using MEMS valves are now possible.

The microvalve used in the bidirectional thruster (Fig. 17.17) had an advertised response time of 100 ms. This normally closed valve opens with 3–5 VDC at 80–100 mA and can be operated by 5-V logic circuits capable of driving a 50-W load. We determined that the valve required a minimum of 20 ms to open and a minimum of 30 ms to close under no-load (no pressure drop) conditions. These results were better than the advertised values, but may degrade when a pressure differential is applied.

17.5.2 Solid Microthruster Array Fabrication

DARPA funded TRW, Aerospace, and Cal Tech to develop and fabricate a microsatellite "digital thruster" system as part of its MEMS program. Digital propulsion, conceived in the United States by D. Lewis at TRW and E. Antonnson at Cal Tech, consists of an array of single-shot thrusters that individually produce only one impulse each; spacecraft maneuvers are performed by firing unused thrusters at the right locations at the right times. Microfabrication enables the creation of large arrays of addressable thrusters, for example, 10,000 on a 10-cm-sq surface, using 1-mm, center-to-center spacing. The digital thruster system is planar, scalable in area, does not require separate propellant tanks or plumbing, does not suffer from microvalve leakage, and can also function as structure. Digital thruster arrays (DTA) can be used for orbit maintenance, orbit adjust, and attitude control on micro- and nanosatellites. Figure 2.17 (Chapter 2) shows a schematic cross section of the basic concept.

The Aerospace effort on the DARPA program has been to design and fabricate thruster components, test various propellants, and characterize thruster performance. Figures 17.21 (a)–(c) show examples of the three micromachined layers that are assembled into a digital propulsion chip (DPC), and Fig. 17.21(d) shows the final product. The current DPC consists of only 15 thrusters in a 3-by-5 array and is used to evaluate thruster performance, repeatability, and reliability. A 24-pin ceramic dual in-line package (DIP) serves as a convenient thruster system carrier during testing and evaluation phases.

The DPC or "rocket chip" is composed of:

- A 400-μm-thick top silicon wafer, which has a 0.5-μm-thick coating of low-stress silicon nitride on top and bottom surfaces. Openings in the top nitride are patterned using laser ablation, and the bulk silicon is anisotropically etched using potassium hydroxide (KOH). This leaves a 70.6-deg expansion nozzle of square cross section with a 0.5-μm-thick silicon nitride diaphragm or burst disk on the bottom. Figure 17.21(c) shows the top view of a processed wafer that contains multiple dice. The die on the left contains 300-μm-sq diaphragms, the middle die contains 400-μm-sq diaphragms, and the die on the right contains 500-μm-sq diaphragms. Each die contains a 3-by-5 array of nozzles on 1-mm centers plus four extra outboard nozzles that serve as alignment ports for die assembly. Chemically assisted laser micromachining could also be used to fabricate these dice. This would enable more freedom in nozzle design, fabrication of circular cross-section nozzles, arbitrary expansion profiles, and circular diaphragms.
- A middle Foturan wafer (typically 1-mm thick), which is direct-write laser-exposed, baked to a semiceramic state, and then etched in a 5% HF solution in water at 40° C to leave cylindrical cavities. Figure 17.21(b) shows a propellant storage die that is 4.5 mm wide × 6 mm high × 1 mm thick. We have fabricated cavities ranging from 500 to 900 μm in diameter with little difficulty. Noncircular cross sections (hexagonal, square, star, etc.) can also be fabricated to improve propellant packing efficiency or to control burn characteristics.
- A bottom silicon wafer that contains polysilicon igniters. Figure 17.21(a) shows a prototype 5-mm-sq die that contained a 4-by-6 array of polysilicon resistors. This die was manufactured using the Multi-User MEMS Processes Service (MUMPS) process at MCNC. Aerospace now fabricates its own 3-by-5 thruster igniter dice in house. The patterned dice consist of a 6-mm-sq silicon substrate, a 3-μm-thick layer of silicon dioxide, a 0.5 or 1-μm-thick patterned polysilicon layer, and a lift-off metal layer on top. The silicon dioxide serves as a transient thermal insulator; the polysilicon traces are 100-Ω resistors that can also serve as exploding bridge wires, and the metal traces serve as electrical interconnects. Future versions could have

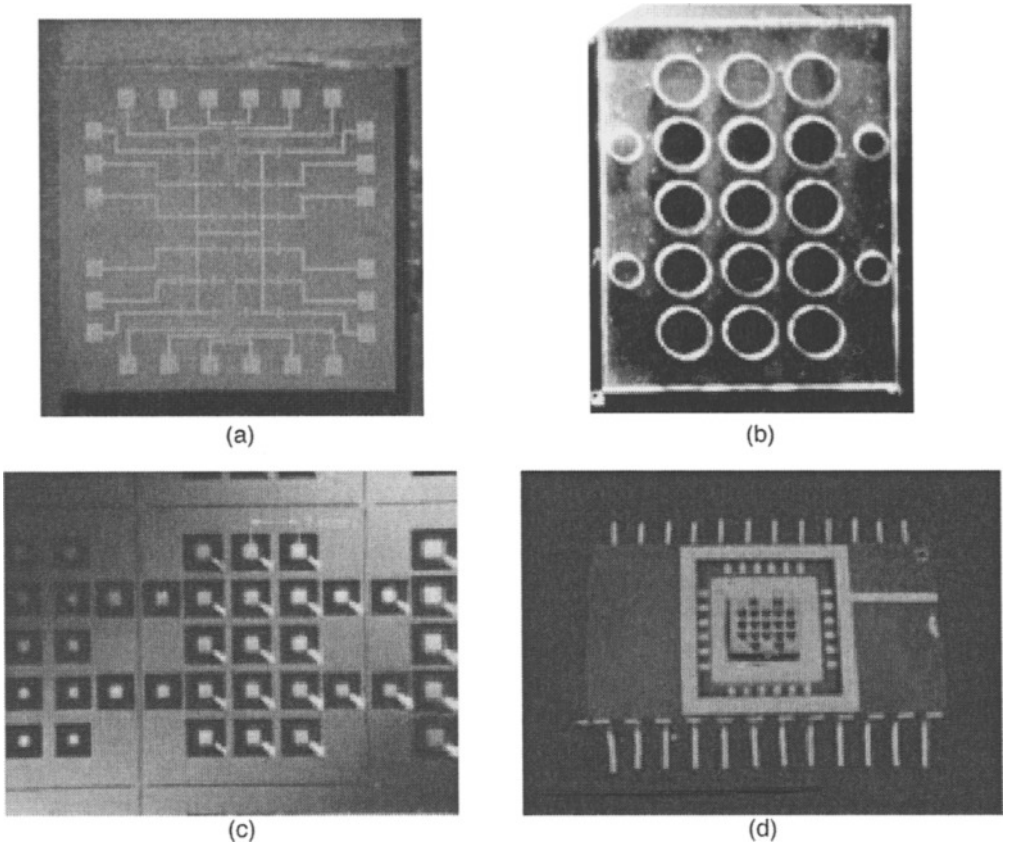


Fig. 17.21. (a) Bottom silicon die contains polysilicon initiators (heaters or exploding bridge wires). (b) Middle Foturan die filled with propellant. (c) Top view of the top silicon layer; the dark regions are etched silicon and the bright squares within them are the silicon nitride diaphragms on the bottom. (d) An assembled "rocket chip" in a 24-pin ceramic carrier.

integrated transistors and/or diodes to provide simplified addressing capability for arrays with a large number of individual thrusters.

Each microthruster is a sealed propellant cavity that is fired by passing a current through the appropriate polysilicon resistor located on the bottom die. We are currently using lead styphnate as the propellant for solid-chemical microthrusters and paradichlorobenzine as the propellant for microresistojets. A voltage of 100 V applied to a 100-Ω polysilicon resistor causes it to promptly vaporize, thus generating a pressure wave that ignites lead styphnate. A voltage of 30 V heats the resistor to just below its melting point, which is used to thermally vaporize paradichlorobenzine over the course of a few seconds. In either case, pressure builds until the silicon nitride diaphragm bursts (this can be adjusted from 220 psi to 1000 psi by correctly sizing the diaphragm area), thus creating an impulse bit as the propellant rapidly leaves. The membrane also protects the propellant from exposure to the environment.

Microthruster fabrication begins with the patterning/processing of 4-in. wafers of silicon and Foturan. Individual "die" are then diced, stacked, fueled, and fused. Other fabrication and assembly issues that were addressed in the development of the DPC are given below.

- Use of the thin silicon rather than silicon nitride as a membrane material

- Use of a circular membrane design rather than a square, to reduce stress-concentration effects normally found in vertex corners
- Alignment and fusing of the three layers to better than 30-μm X-Y tolerance
- Injection of the propellant and removal of air pockets
- Tailoring of the resistor valve and drive circuit for minimum power ignition
- Thermal isolation strategies for the polysilicon heaters
- Static pressure testing of the nitride membrane burst pressure
- Fracture pattern of the diaphragm and its effect on net thrust produced
- Fusing the three wafers and evaluating potential debonding at other nonignited adjacent cells

17.5.3 CMOS Resistojet Fabrication

CMOS circuits are ubiquitous: they are used in computers, wireless telephones, most "smart" appliances, automobiles, and many other electronic devices. A large number of CMOS "foundries" exist, and a fair number are available for custom circuit fabrication. Also, CMOS processes can be used to produce bulk-micromachined MEMS if the dice are designed such that selected regions of the silicon substrate are exposed after fabrication. Liquid etchants such as HNO_3 and HF in water or acetic acid attack all exposed surfaces equally and can be used to undercut masked areas. Etch rates vary from about 1 to 10 μm per min. Liquid etchants such as KOH-water,[43,44] hydrazine-water,[45,46] ethylenediamine-pyrocatechol-water (EDP),[47,48] or tetramethyl-ammonium hydroxide (TMAH)[49,50] attack the <111> planes of silicon at a much slower rate than the other planes to produce "chiseled" cuts as shown in Fig. 17.21(c). These cuts follow the <111> planes, which form a 54.7-deg angle with the surface for a <100>-oriented silicon substrate. Anisotropic etch rates are typically about 1 μm per min. In practice, EDP and TMAH are typically used for bulk micromachining of silicon die that include CMOS or other electronics. These etchants remove silicon at high rate, but not the passivation layers or exposed metals (e.g., bond pads).

The key to cost-effective initial development is to utilize existing prototyping services such as the DARPA-supported Metal Oxide Semiconductor Implementation Service (MOSIS) and MUMPS.[51,52] The Orbit Semiconductor "Foresight" foundry also offers its services directly to users if more rapid turn-around time is required. MOSIS can provide user-designed CMOS die for under $1000 (Tiny Chips: 4 or 5 copies of roughly 2.2 × 2.2 mm die) and MUMPS can provide about 14 copies of user-designed 1-cm-sq surface-micromachined die for under $3000. One can also use other fabrication services such as Sandia's SUMMiT service[53] or have entire wafers fabricated by a custom foundry (on the order of $100,000 per wafer).

Figure 17.22 shows a photograph of our first CMOS resistojet die, fabricated using a 2-μm CMOS process offered by Orbit Semiconductor through MOSIS. This "Tiny Chip" is 2.3 mm on a side. The polysilicon layer, normally used as the gate structure in CMOS transistors, was patterned to provide a resistive heater element that is sandwiched between two patterned passivation (glass) layers. This basic concept is an extension of the readily-available pixel-160 × 160 microheater design developed by the National Institute of Standards and Technology (NIST).[54] Two NIST microheaters were included on the die (center and lower part of Fig. 17.22) to test our post-processing technique. Two square openings, one above each microheater in Fig. 17.22, were added as an etch-depth indicator. When the square openings become pyramidal pits whose walls meet at a point, the etch process is complete. Resistojet 2 is simply the NIST design with an added input channel and an expansion nozzle, while Resistojet 1 is a larger design with a heftier heating element. Complete resistojets are created by etching the die to create flow channels, sectioning the die to separate the four resistojet subdie, and bonding paired components along their broad surfaces.

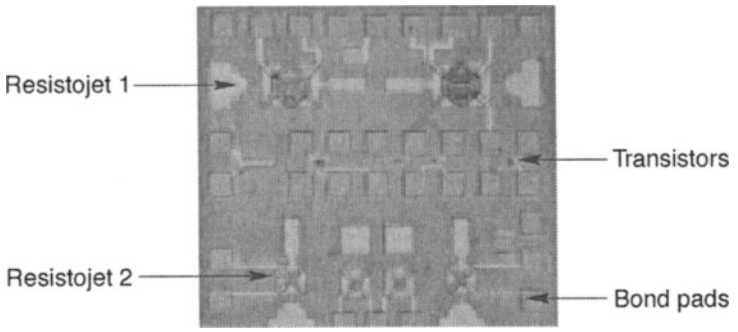


Fig. 17.22. Unetched silicon die with resistojet and electronic structures as received from MOSIS.

Figure 17.23 shows a SEM of part of the etched CMOS silicon die that contains the NIST microheaters on the right and the etch-depth indicators on the left. The polysilicon heating elements are suspended above the pyramidal pits by glass and metal stringers; metal traces exit the central heating element and connect to bond pads on the far right.

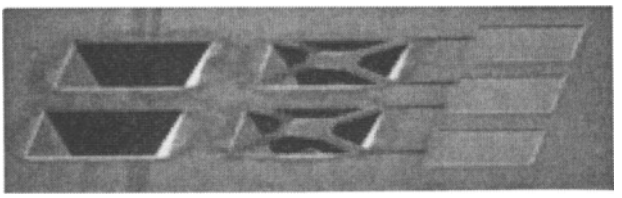


Fig. 17.23. SEM photograph of EDP-etched structures in a CMOS-processed silicon die.

Figure 17.24 shows a close-up of the resistojet-1 subdie from the upper left corner of Fig. 17.22. The expansion nozzle is on the left, the heating element is near the middle, and the propellant inlet is on the right. More advanced designs, which include integrated flow sensors, temperature sensors, and power electronics, are currently under fabrication.[55]

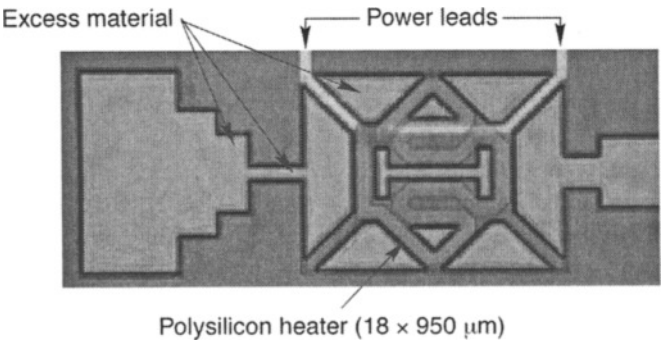


Fig. 17.24. Close-up of the resistojet-1 design from Fig. 17.22. This die has not undergone an EDP post-process etch.

17.5.4 Process Flow for Synthetic Jets

The synthetic jets manufactured at MIT are fabricated as follows.

- The resonant cavity geometry is etched 20 μm into a silicon handle wafer, using a medium-depth etch process such as KOH or, in more recent times, DRIE. The depth of this etch will depend on the design of the cavity and membrane characteristics.
- The handle wafer is turned over and the feed holes are etched from the rear using anisotropic KOH etch. This yields the characteristic pyramid shape of the feed holes (Fig. 17.11). The handle wafer is fusion-bonded to the "device" wafer, which consists of a silicon-on-insulator (SOI) wafer. The top layer of the SOI will define the drive membrane.
- The device wafer is thinned back to the imbedded oxide layer by a combination of grinding and KOH etching. The oxide layer is then stripped.
- The metal electrodes are patterned onto the top surface. These form the resistive heaters, which are used to excite the membrane into motion. At this stage, alternative techniques for excitation could be used, including the use of a solgel piezoceramic film.
- As the last step, a shallow plasma etch is used to punch a hole through the top membrane to expose the cavity and feed holes.

The following process flow is used for the Georgia Tech synthetic jets.[29,30]

- The device wafer is anisotropically etched, using silicon dioxide as a mask, from both sides to form the orifice and actuator cavity.
- A second, shallow anisotropic etch is then performed on the backside of the wafer to create a shallow recess, which forms the electrostatic actuator.
- The wafer is then reoxidized and a layer of aluminum is sputtered onto the back side of the wafer to create the actuation electrode.
- The actuation membrane is formed from a polyimide film, bonded to the back side. This film is then coated with aluminum to create the second actuation electrode.

The resulting devices reported had orifices of 175 μm with membrane diameter of 3 mm. In operation, the device exhibited a resonance frequency of about 1 kHz, and fluid velocities as high as 17 m/s were measured close to the actuator exit.

17.5.5 Microengine Process Flow

The MIT microengine process is defined by a lengthy and extremely complex series of fabrication steps. The success of the microengine depends on its manufacture to very tight tolerances, and this greatly complicates the fabrication and process details. As an intermediate step to the full engine, a microturbine/bearing rig, consisting of a free-floating radial turbine supported by thrust and journal bearings, has been fabricated.[56] This device contains most of the necessary process steps required for the full engine and also provides a crucial test platform for rotor and bearing designs. The details of the fabrication and initial testing of this bearing rig are too lengthy to repeat in detail here, but are outlined in Lin, *et al.*[57] The bearing test rig is composed of five silicon wafers fusion-bonded to form a 2.5-mm stack. The outermost wafers (wafers 1 and 5) are "foundation plates" that contain fluid interconnects to the internal device. Moving inward, wafers 2 and 4 are the aft and forward thrust plates, which contain thrust bearings and running gaps on the inner surfaces, and fluid distribution channels etched into the outer surfaces. Finally, the innermost wafer, wafer 3, contains the rotor and stator as well as the journal bearing. The journal bearing represents one of the more challenging features of the microengine design and is defined by a 300-μm-deep, 12-μm-wide DRIE etch. Future builds will increase this dimension to 800 μm deep.

The microengine process makes heavy use of DRIE (Bosch process) etching technology, which is used to define the structural components of the engine (turbine blades, combustion

chamber, etc.), the fluidic channels connecting each component, and the lubrication systems (journal and thrust bearings). These deep etches are coupled to a series of shallow etches used to define finer, more subtle features of the device, such as blade clearances. Last, more specialized processes are included for the electrical machinery.

17.5.6 Ion Thruster Fabrication

Miniaturized liquid metal ion sources (MILMIS) that could be fabricated using microfabrication technology were championed by J. Mitterauer during the early 1990s.[58,59] These devices would be mechanically similar to the Spindt "microvolcano" shown schematically in Fig. 17.25.[60] The Spindt microvolcano relies on solid electrodes, *not* on conductive fluids, that are electrostatically shaped into sharp tips; it can directly field-ionize gases. Figure 17.26 shows electrostatic potential contours in 10-V steps for a potential drop of 100 V between the upper and lower electrodes. Electric fields of ~5×10^8 V/m (50 V per 0.1 μm) are developed near the rim of the volcano orifice. These fields are high enough to generate ions by field-ionization and can readily produce molecular ions without fragmentation. In general, electric fields of ~10^9 V/m (order of magnitude) will produce copious electron emission from metals, and fields of ~10^{10} V/m (order of magnitude) will enable efficient field-ionization.[61]

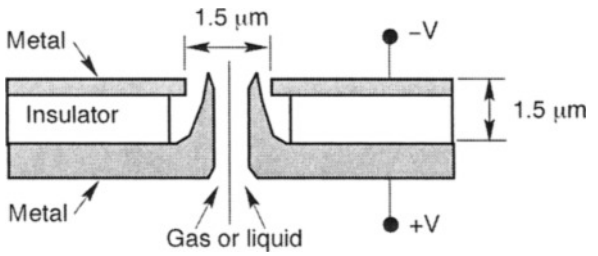


Fig. 17.25. Schematic cross section of the Spindt microvolcano field emission ionizer.

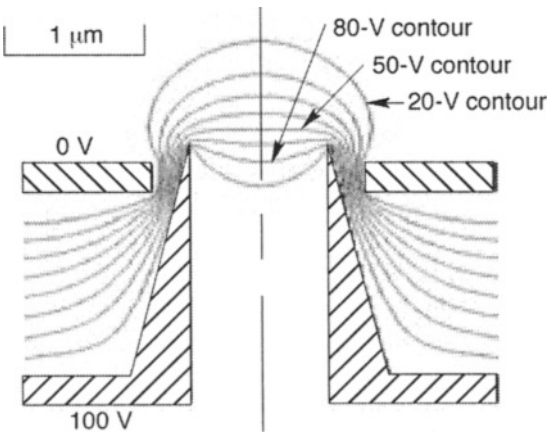


Fig. 17.26. Schematic cross section of the Spindt microvolcano field emission ionizer showing electrostatic contours in 10-V steps.

From the standpoint of fabricating a practical ion thruster based on the above concepts, the most important criteria is developing microstructure gaps or tips that can generate very high electric fields. An approach being considered at Aerospace is the nanotexturing of surfaces whereby natural gaps, facets, or separations can be exploited for field ionization. Initial emission-current experiments from a microstructure similar to that shown in Fig. 17.16 found that the emission characteristics exceeded expectations. Further experiments are currently under way on similar structures.

17.6 Microthruster Performance

A brief review of various methods used to characterize electric thrusters can be found in Crofton.[62] Many of the techniques used to measure the millinewton-to-newton thrust levels produced by electric thrusters can be directly applied to both chemical and electric microthrusters. Improvements in sensitivity, however, are required to support micronewton thrust levels produced by some electric microthrusters, for example, micro-ion engines operating at 1–10 W power levels. Basic thruster performance is usually characterized using a thrust stand and a mass flow rate monitor. Thrust is measured directly, and specific impulse is calculated using thrust and mass flow rate. For pulsed thrusters, impulse bit is measured instead of steady-state thrust. We used these basic performance measurement techniques to characterize experimental cold-gas and solid-rocket microthrusters.

17.6.1 Cold Gas Thrusters

We measured the micronozzle performance, as shown in Sec. 17.2, and found that it is degraded under low Reynolds number conditions as a result of viscous losses. Spacecraft thrusters, however, are more than just nozzles; they usually require valves, filters, etc. A similar situation exists for flow through micromachined valves. The EG&G IC Sensors silicon valves used in the bidirectional thruster shown in Fig. 17.17 have a limited stroke; the boss-to-valve seat gap is typically about 20 μm in the open state, thus introducing a flow impedance upstream of the nozzle. In addition, these electrothermal valves are sensitive to operating temperature and are affected by heat transfer to the fluid that is being controlled.

We characterized the bidirectional thruster in a vacuum chamber at 5 millitorr (0.7 Pa) ambient pressure, using an inverted pendulum thrust stand based on a NASA-Lewis design.[63] We measured continuous thrust to 0.1 mN accuracy and mass flow rate through the thruster to 5×10^{-8} kg/s accuracy. Figure 17.27 shows thrust as a function of feed pressure and valve voltage for this thruster, using room temperature argon as the propellant. The EG&G silicon valves are normally closed, and they begin opening at applied voltages greater than 4.5 V. For 5-V operation, the mass flow rate, and hence thrust, increase as feed pressure increases up to about 19 psi (1.3×10^5 Pa). At higher feed pressures, the thrust (and mass flow rate) drop because of valve cooling by the propellant. An almost linear thrust versus feed pressure relationship is obtained at 5.5-V operation. The maximum thrust of 1 mN at 24.5-psia (1.7×10^5 Pa) feed pressure is about an order of magnitude below what the nozzle itself can produce. For the range of operating voltages and feed pressures given in Fig. 17.27, the data reveal that the thrust level is limited by viscous losses in the silicon microvalve.

Specific impulse for the bidirectional thruster with argon propellant increased almost linearly from 16 s at 8.5-psia feed pressure to 27 s at 24.3-psia feed pressure. The highest measured specific impulse of 27 s is 50% of the ideal specific impulse for argon with a 10:1 expansion nozzle into a vacuum. Further optimization of nozzle size and microvalve operating characteristics is still required.

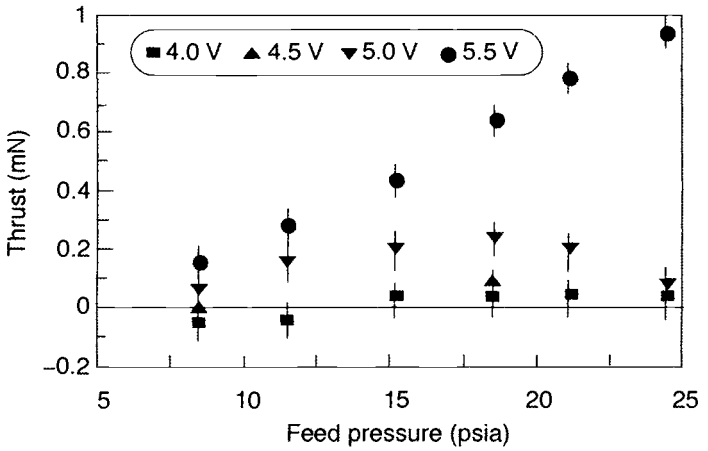


Fig. 17.27. Measured data for thrust vs feed pressure for a bidirectional mN thruster shown in Fig. 17.17.

17.6.2 Solid Rocket Microthrusters

Our solid rocket microthruster arrays use lead styphnate as the propellant. Lead styphnate is a shock-sensitive explosive that is typically used as an initiator; once triggered, it provides enough thermal energy to ignite a larger quantity of "secondary" propellant or explosive. Our preliminary propellant testing effort revealed that typical double-base solid-rocket propellants,[64] such as nitrocellulose plus nitroglycerin, could not be ignited using a simple polysilicon resistor heated to incandescence. Solid-rocket propellants and most secondary explosives are designed so that they are relatively safe under normal shipping and handling conditions and are therefore difficult to ignite. Successful ignition for our microthruster array required the use of a primary or initiator explosive (lead styphnate) and vaporization of the polysilicon resistor using a high-current pulse. This pulse creates a transient high-pressure shock wave that ignites the propellant in a single sealed cavity.

The polysilicon resistors are 0.5 or 1 μm thick, 20 μm wide, and 400 μm long. They are patterned as a planar double hair-pin structure on top of a 3-μm-thick oxide layer and are doped with phosphorous to provide a sheet resistance of 11 and 5.5 Ω per square, respectively, for 0.5-μm and 1-μm-thick layers. The bond pads and power buses are composed of 100-μm or wider polysilicon traces topped with a 0.35-μm-thick metal (0.05 μm Ti + 0.1 μm Ni + 0.2 μm Au) layer. Figure 17.28 shows several high-speed video frames of a 0.5-μm-thick polysilicon resistor that has 60 V applied across it starting somewhere between t = 49 and t = 74 μs. The optical flash is quite bright but short-lived. Instantaneous power is 16 W, but since it lasts for only about 100 μs, only 1.6 millijoules of energy are used. Most of the resistor survives, but it becomes an open circuit. We routinely apply 100 V across these resistors, which completely vaporizes the double-hairpin polysilicon trace. In open air, the polysilicon vaporization generates an audible "pop" that can be heard several feet away.

We measure the impulse bits produced by our digital microthrusters using a wireless ballistic pendulum. The current pendulum contains a knife-edge pivot on top, a 9-V battery, an infrared receiver for data communication, an on-board microcontroller, a 9-V to 100-V dc-to-dc step-up converter, a 1-μF energy storage capacitor, a zero-insertion-force socket for the 24-pin "rocket chip" shown in Fig. 17.21(d), a bank of silicon-controlled rectifiers (SCR) that switch power to individual microthrusters, a 5-mm-diam mirror, and a copper vane for Eddy-current damping. Firing commands are sent to the pendulum at 1200 baud by an infrared light emitting diode using

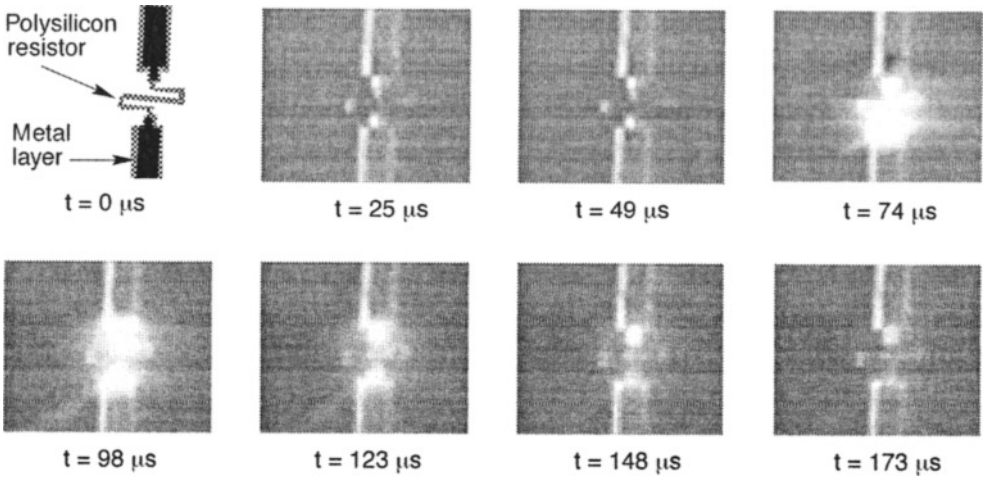


Fig. 17.28. Video frames of a polysilicon resistor operating as an explosive bridge wire.

on-off envelope modulation of a 38-kHz carrier. The current pendulum has a mass of 398 g, and its center-of-gravity is 11.8 cm from the knife-edge pivot.

Pendulum displacement or swing angle is measured by an optical interferometric displacement monitor: a rigidly fixed optical fiber launches monochromatic light toward the mirror on the pendulum, the light reflects off the mirror and is sent back through the fiber, and the interferometer measures the changes in free-space displacement between the fixed fiber and the pendulum mirror in real time. The interferometer measurement accuracy is better than 20 nm, but laboratory vibrations currently limit our instantaneous minimum resolution to about 0.5 μm. Our ballistic pendulum can measure impulse bits from 10^{-6} N-s to 10^{-3} N-s. Improvements in vibration isolation should enable measurement of even smaller impulse bits.

The pendulum is calibrated by rolling ball bearings of different mass down an inclined ramp and letting them impact the pendulum. A series of 20 remotely driven solenoids are used to launch individual bearings down the ramp. A high-speed video camera captures the bearing trajectory before and after impact, thus enabling calculation of momentum transfer to the pendulum.

Figure 17.29 shows measured impulse bits in air for eight successive thrusters on a single 15-thruster die. The error bars are a result of random pendulum vibrations caused by air currents, and the scatter in the data is probably due to variations in propellant mass between thrusters. Each thruster contains roughly 1 mg of propellant, and the expected impulse bit in air is 1.8 mN-s. The measured values are 5% of expected, and further investigation has revealed that most of the propellant is blown out of the thrusters in an unburned state. We are now investigating the use of smaller propellant particles and smaller diaphragm areas (burst pressures increase to 1000 psi) to promote faster propellant combustion. We are also investigating the use of igniters on the diaphragm side of the propellant cylinder to start combustion near the exit plane.

The solid microthrusters have a firing time of approximately 1 ms. Figure 17.30 shows a series of still frames from a high-speed video of a single thruster firing on the thrust stand. Ignition occurs at t = 0 s when the plume appears near the center of the frame. The ballistic pendulum is to the left of center, the thruster chip nozzles point to the right, the fixed optical fiber mount is on the bottom right, and the individual frames are 7.6 cm sq. The visible plume is longer than 1 cm and decays to the limit of visibility by t = 1.11 ms. Average thrust during this time is 80 mN, and the equivalent chemical power level is 60 W. Improved combustion efficiency should boost these numbers by an order of magnitude.

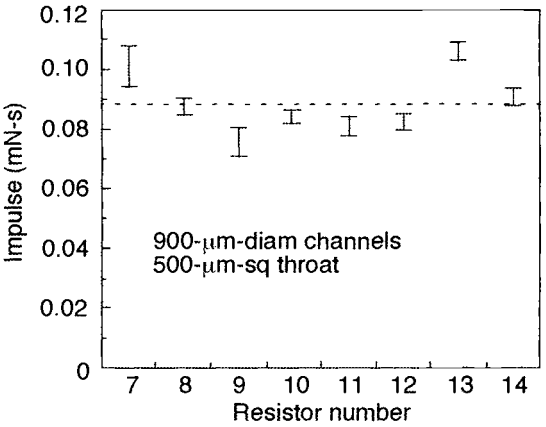


Fig. 17.29. Measured impulse bits generated by eight thrusters from a single thruster chip.

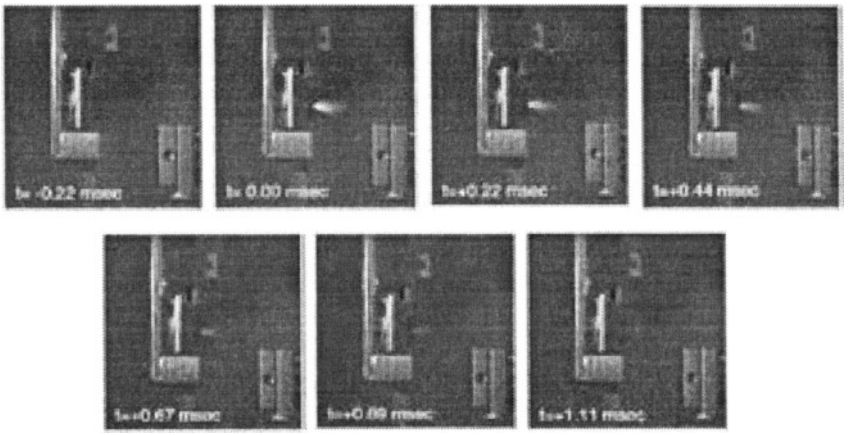


Fig. 17.30. Video frames of the solid microthruster firing on a thrust stand.

17.7 Micropropulsion via Bioenergetic Decay Processes

As satellites shrink in size, the desire to utilize the precious mass for multiple purposes increases. One intriguing possibility is to use mass for both structure and on-orbit propellant; spacecraft structures are usually designed to survive launch loads, but once on orbit, much less structural integrity is required. Satellite propulsion systems are designed to make efficient use of the onboard propellant to extend mission life. Consider the possibility of a self-consuming satellite whereby noncritical satellite structural components are used as sources for propulsion fuel. Evolutionary examples toward this end are the use of teflon rod structural members as feedstock for a pulsed plasma thruster or the use of energetic materials with innate strength/stiffness as fuel for a chemical or electric thruster. A more natural option, but yet novel to space systems, would be to implement biological processes to convert biofeedstock structures (e.g., proteins, sugars, polysaccharides from sugarcane, bagasse) to fuel (e.g., ammonia, methane, methanol, ethanol) by anaerobic digestion schemes (i.e., via microbacterial and enzymatic action). A simpler (i.e., better understood) approach to biofuel production would be by aerobic (i.e., combustion) digestion but would

entail carrying oxygen on board. Since these novel concepts use microbes to produce propellant, they are true "micro" thrusters.

At first glance, biodigestion schemes have more commonality with science fiction rather than aerospace engineering, but a brief review of the bioenergetics and other pertinent information gives a result that deserves mention and should abet further detailed study. It is understood that the very slow nature of biological processes would render such a system impractical as an on-demand fuel generator for a particular orbit maneuver. It is more likely that a biofuel generator would be continuously operational throughout the lifetime of the satellite, supplementing the on-board stored fuel and affecting an extension in the mission lifetime. For a biofuel generator source to be considered for space propulsion use, critical issues beyond that of the bioenergetics must also be analyzed. Some of these issues are listed here.

- Ability of biological feedstock materials to meet the mechanical strength requirements of space structural components
- Effects of vacuum on microorganism survivability
- Means for maintaining hydrolysis reactions in partial vacuum, at cold temperatures, and in near zero gravitational force environment
- Feasibility of fabricating and maintaining (e.g., feeding) miniature digesters—or the feasibility of fabricating structural members as biodegradable digester systems

Why bother obtaining fuel/energy from a bioconversion process? If such a system can be made for space propulsion applications then it naturally offers a self-consuming system with innate survival rules designed to continue delivering useful energy as the system degrades. In most biosystems, energy is initially derived from the hydrolysis of simple sugars (e.g., glucose, glycogen). However, as these food stores get scarce—that is, during starvation, energy is still derived albeit at less efficiency from the breakdown of functional organic material—less vital material is sacrificed to feed those that are urgently necessary for survival. These choices are naturally made in a closed biological system. For example, certain microfungus (e.g., saccharomyces cerevisiae—"yeast") can even automatically alter the digestion scheme going from full respiration mode (i.e., in the presence of oxygen) to a fermentation mode (i.e., absence of oxygen). By implementing biochemical processes and incorporating a variety of microbial genera, it is possible that an adaptable generator for propellant fuels can be developed to extract the maximum energy allowable. The development of such a generator has other obvious benefits: it could help sustain very long-duration manned missions.[65]

Most known biological systems have materials that have a structural function, as do space systems. In some biological systems pectins and hemicellulose (carbohydrates that yield a mixture of sugars and acids when hydrolyzed) serve that role. In other biosystems bioceramics (e.g., $CaCO_3$) play that role. Biological ceramics are inorganic biocompatible minerals (e.g., for human bone it is hydroxyapatite) reinforced with organic biopolymers (e.g., for human bone it is collagen, a triple helix fibrous protein structure with high tensile strength). The ceramic phase provides structural integrity (strength and hardness), while the polymer gives the structure toughness. A sintered hydroxyapatite bioceramic yields a fracture strength of 140 Mpa.[66] There are other biological hard tissues, like that of the mollusk shell, which contain a smaller fraction of organic material (< 5% weight as opposed to >40% for mammalian bone) and exhibit high flexural and compressive strengths (>300 Mpa)[67] and are very resistant to fracture (e.g., fracture toughness between 5–11 Mpa-$m^{1/2}$ has been reported for the nacreous shell,[68] 4–10 Mpa-$m^{1/2}$ for the abalone shell[69]). These biological ceramics are stronger than some high-technology ceramic and cermet materials (e.g., strength and toughness for Si_3N_4 ~190 MPa and ~5 Mpa-$m^{1/2}$, for SiC ~140 MPa and ~4 Mpa-$m^{1/2}$, and for ZrO_2 ~80 MPa and ~6 Mpa-$m^{1/2}$).[70] The issue is not that a biological

material with sufficient structural strength can be found, but can this material be made space-environment survivable, and more importantly can this material store sufficient biodegradable matter to yield a practical propulsion fuel? A structural member might be a thin-wall inorganic or metallic tube that is capable of withstanding the space environment and is filled with organic solid "waste" or nutrients. This type of an integrated structural member digester system would generate fuel until contamination or product waste inhibitors quench digestion.

The key to promoting and maintaining digestion is to establish a habitable environment (e.g., pH, temperature) for microorganisms (i.e., bacteria) and the digestive enzymes. Digestive enzymes are proteins that exhibit catalytic activity of a highly specific nature to help break down large molecules (e.g., proteins into simple sugars). The enzyme acts by reducing the activation barrier of a particular bimolecular reaction. Bacteria act by excreting digestive enzymes into the surrounding area. Digestion can proceed by aerobic or anaerobic processes, depending on the bacteria and enzymatic action. For space applications, anaerobic digestion seems the more practical. In nature, a heterogeneous population of bacteria will degrade organic matter to form methane (CH_4) and carbon dioxide (CO_2) gas. The biochemical processes involved in anaerobic digestion can be generally described as the breakdown of proteins by hydrolysis to yield amino acids and simple sugars (e.g., $C_5H_{10}O_5$); for methanogenic bacteria the digester further yields volatile acids, ammonia, and in the final stage, the formation of CH_4 by catalyzing the reaction ($8\ H + CO_2 >> CH_4 + 2\ H_2O$). Other anaerobic processes yield alcohols (e.g., ethanol, C_2H_5OH) or acids (e.g., lactic acid, CH3·CHOH·COOH) and energy. Experiments show that 50–75% of the fermentable solids can be converted to final products using mesophilic (35–40 C) or thermophilic (60 C) bacteria.[71] Maintaining a 60-C digester in space might require too much input power even though the reaction rate is faster by a factor of 2. Assuming the digester is not initiated on the ground and is allowed to cool, it will require some initial power to start the digestion process in space. The biochemical reactions release "waste" energy as heat, which can act to increase temperature. In addition, in low Earth orbit (LEO), sun light and "poor" heat convection schemes can be used to heat and maintain the required temperatures. Although enzymes can recover after being frozen to 0° C, they in general undergo irreversible molecular restructuring and loss of capability upon heating to 80° C. Some enzymes do work at temperatures up to 400° C, but these are commonly found near geothermal vents and are sulfur rather than carbon reducing.[72] Conceptually a digester designed for a space fuel generator application would need to be sealed with a semipermeable membrane for product (e.g., gases/liquid) pass-through into a secondary propellant container. The container could be partially pressurized (i.e., subatmosphere) enough to maintain hydrolysis reactions within the "feedstock" structural members. Microorganisms do live in reduced atmosphere environments, and experiments show that some bacterial spores and molds can, in fact, survive vacuum conditions ($5 \times 10^{-7} - 10^{-10}$ torr) for a significant period of time (~5 yr).[73] It is more likely that the propellant container will be under high pressure (e.g., >>10 atm) and possibly there will be a gas/liquid equilibrium. Clearly, microbial action does not cease at elevated pressures. There is a class of barophilic bacteria, found in deep sea environments, that are adapted to life at high pressures (e.g., methanococcus thermolithotrophicus functions at 50 Mpa = 500 atmospheres,[74] MT41 at 1200 atmospheres (temperature 2° C).[75] Microorganisms that do not come from high-pressure environments may still respond when placed at high pressure (e.g., E. coli at 530 atmospheres).[76] In general, elevated pressure favors those biochemical processes that tend to reduce system volume and inhibits those that cause an increase in volume. Furthermore, increase in hydrostatic pressure can either increase or decrease the upper temperature limits for bacterial growth; it does, however, narrow the pH ranges for growth.[77]

Although a case can be made for finding a microbial/biofeedstock system that can be made to provide both structural support and useful biodegradable products, the basic issue remains whether microbial digestion can produce sufficient quantities of product gas or liquid to be of any use for propulsion. To definitively address this issue a particular biosystem process and a specific type of propulsion system must be chosen. It is clear that a bipropellant system carrying oxygen can initiate aerobic or combustion processes, which with pure solid glucose ($C_6H_{12}O_6$) as feedstock, can at the theoretical limit (100% combustion to carbon dioxide and water) liberate 673 Kcals/mol (12.1 KJ/g) of heat.[78] Because of the complexity of designing bipropellant systems, we assume that such a "self-consuming" bipropellant propulsion system would not be the first design choice. An alternative approach is to consider anaerobic digestion schemes that can naturally degrade organic matter within a single container and provide supplemental fuel to a monopropellant cold gas or resistojet type thruster. Experiments show that the choice and concentrations of nutrients in the biofeedstock influence the product and the yields. Results from the Biomethane Project[79] show that feedstock with high levels of dry solids (i.e., 8–12% nutrients) can result in unstable operation of a laboratory digester—a 4–5% level of concentration was deemed more appropriate.[80] In one laboratory digester experiment methane production was measured as ~ 400 cc (STP) per gram of consumed organic matter (protein-rich dog food was used as feedstock); the total gas production (methane plus carbon dioxide) was higher reaching ~ 600 cc per gram of consumed organic matter.[81] The residence time for the nutrients to begin generating steady-state gas production can be chemically controlled, but left to its own was measured to require a minimum of 5 days. Using this simple example of a biogas generation process, one can apply it to a microthruster system. For this example,

- Assume a 10-kg class satellite has a single 1-mN monopropellant gas thruster (or resistojet that can heat the propellant to 800 K).
- Assume of this 10-kg class satellite, 20% of the mass (2 kg) is available for redesign as a combination digester/structural element system.
- Assume only 5% of this 2 kg matter is composed of a biodegradable organic matter (100 g); then roughly 40 L (STP) of methane or 60 L (methane + CO_2 at STP) of total volatile gases can potentially be produced. This conversion rate is roughly 67% by weight of the initial starting mass (100 g). Conversion rates of 75% have been reported.
- Further assume that only 10% of this net volume of gas is really extracted (i.e., 10% efficient in the space applications, therefore 4 L STP methane and 6 L methane plus carbon dioxide).

Figure 17.31 shows the supplemental propulsion firing time (seconds) that is gained from the biomass fuel generator as a function of propellant temperature. The results show that in excess of an hour is gained. For the more efficient design, whereby 100% of the potential generated gas is used (60 L total, 40 L methane), the additional fuel gives propulsion stores between 11 and 19 h. The traditional way to compare the usefulness of this expended "extra" fuel is to calculate the Δv (m/s) for the 10-kg satellite. Figure 17.32 gives this result for the 10% efficient system as a function of propellant temperature. The calculated Δv (m/s) should enable the 10-kg satellite to do a couple of maneuvers; for example, increase the altitude by a few kilometers (at ~700-km orbit), or move forward 10 km in a few hours while in a 700 km orbit. If the more efficient digester (100% capture of generated gas) design were implemented, then the Δv (m/s) is in the range ~5–10 m/s, and this enables 100-km altitude changes in GEO (geosynchronous orbit) and a faster rate of "move-forward" maneuver in LEO (~700km). In the given example, it is important to remember that the assumption is that the biomass fuel generator *does not penalize* the satellite with extra added mass but is actually a part of the satellite design weight perhaps integrated into a structural member.

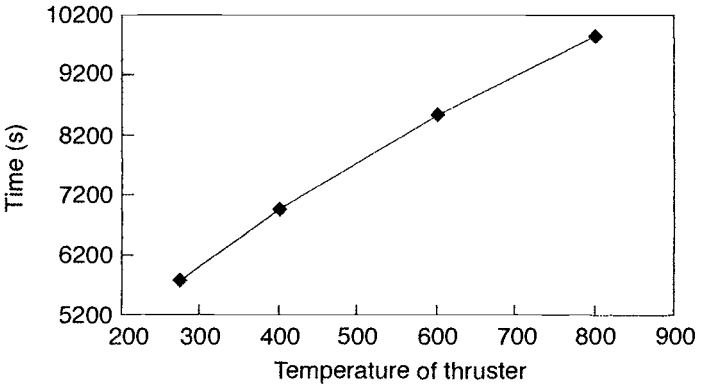


Fig. 17.31. Supplemental thrust time gained for a 1-mN thruster with the gas by-products from the anaerobic digestion of 100 g of protein feedstock, but only 10% of the generated gas is captured.

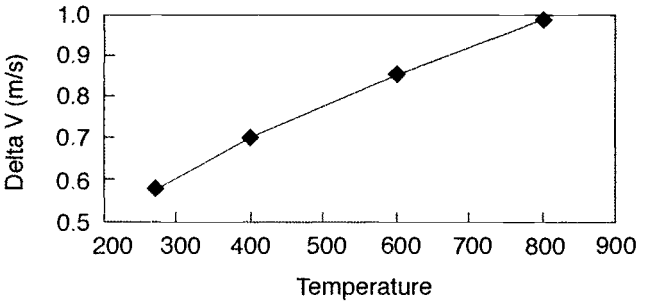


Fig. 17.32. Change in speed of 10-kg satellite (Δv [m/s] as a function of propellant temperature). Assumes 6 L of gas expended (2/3 methane, 1/3 CO_2 by volume).

The above anaerobic digester example is designed to produce methane for a monopropellant thruster. Anaerobic fermentation can also produce alcohols (e.g., ethanol, methanol). Ethanol, for example, has 30% more energy concentration when compared with methanol and represents nearly 80% of the energy contained in glucose (12.1 KJ/gm), but is half the weight when produced (i.e., 2 moles ethanol are produced per 1 mole of glucose). Unfortunately, the fermented fuel would need to be combusted to extract the energy (i.e., requiring an oxygen source).

17.7.1 Alternative Application

An alternative space application where anaerobic processing of biomass would be useful is the deployment and maintenance of very large (>>10 m) inflatable structures.[82] These structures can be used as solar concentrators or as antennas. The inflatable structures are typically fabricated of polymeric materials that are rigidized by maintaining a slight overpressure (10^{-4} torr) and/or by an optical curing scheme using the sun. Gas is necessary to initially inflate the "space-bag." Furthermore, during the lifetime of the mission (>1 yr), micrometeorites will puncture holes in the inflatable structure, and additional make-up gas will be required to offset the loss of gas because of leaks. The use of anaerobic biomass digestion could be a viable alternative to producing both the initial gas for inflation and that needed for make-up. Furthermore, it is conceptually possible that with control of the digestive chemistry, no gas-valving apparatus would be required. Gas could be produced at the rate required by setting the pH of the fermentation chemistry outside the

optimum range. Using the above example of biogas generation where 6 L (STP volume $6 \times 10^{-3} m^3$) of gas is be generated, expanding this volume of gas at STP to a pressure of 10^{-4} torr ($\sim 1.3 \times 10^{-7}$ atm) would fill a "space-bag" volume ~46,000 m^3, a balloon structure nearly 44 m in diameter.

17.7.2 Microengineering Biogas Generation and Outstanding Issues

Microengineering technology is uniquely suited for developing miniaturized digesters, which could be integrated into common satellite systems. MEMS valves, pumps, injectors, stirrers, and heaters are much more advanced for fluidic applications than they are for gas applications. MEMS temperature sensors and pH meters for fluidic applications and microelectrophoresis techniques are advancing, which will enable the controlled injection of enzymes. One can conceive of bioengineered rod-shape microdigesters that are patterned much like the rings of a tree, except the layers represent nutrients and contain microorganisms, and material is digested from the center out. However, there are several issues that must be addressed before biogas generators become useful in space. Some of these are listed here.

- The effect of microgravity on biological processes, microorganism behavior and growth
- Better understanding of the heterogeneous chemistry for biogas production
- Better understanding of the microfluid dynamics for space environment hydrolysis reactions
- Better understanding of the gas dynamics through micropore channels in microgravity environments
- Bioengineering of miniaturized, efficient, anaerobic "near" self-contained digester systems
- Radiation-tolerance of different microorganisms

Finally, we recognize that biomass propulsion and biomass gas generators will be used primarily in Earth orbit to prevent the introduction of Earth organisms onto other planetary bodies. In addition, adequate disposal mechanisms such as deorbit with complete reentry incineration may be required to prevent introduction of radiation-induced mutations to our biosphere.

17.8 Conclusions

Micropropulsion is a rapidly expanding discipline as a result of the recent emergence of micro/nanosatellites, micro air vehicles, and MEMS. Microtechnology offers micromachined versions of existing thruster designs and the opportunity to exploit new physical phenomena that become dominant at small scale. Field-ionization thrusters. for example, directly exploit quantum tunneling. We have presented some of our current work in this rapidly evolving discipline and expect many exciting developments in the years to come.

17.9 Acknowledgments

We gratefully acknowledge the support provided by The Aerospace Corporation through the Corporate Research Initiative Program. We also gratefully acknowledge the DARPA MEMS program for supporting the digital microthruster effort at TRW, Aerospace, and Cal Tech. We also acknowledge NASA JPL Microdevices Laboratory, the Air Force Lincoln Laboratory, Army Research Office, and DARPA for supporting work on the micronozzle synthetic jets and microturbine engine.

17.10 References

1. J. Mueller, "Thruster Options for Microspacecraft: A Review and Evaluation of Existing Hardware and Emerging Technologies," AIAA 97-3058 (Seattle, WA, July 1997).
2. S. W. Janson, "Chemical and Electric Micropropulsion Concepts for Nanosatellites," paper presented at the *30th AIAA/ASME/SAE/ASEE Joint Propulsion Conference* (Indianapolis, IN, June 1994).

3. W. A. deGroot and S. R. Oleson, "Chemical Microthruster Options," AIAA 96-2863 (Lake Buena Vista, FL, July 1996).
4. R. W. Gallington, H. Berman, J. Entzminger, M. S. Francis, P. Palmore and J. Stratakes, "Unmanned Aerial Vehicles," in A. K. Noor and S. L. Venneri, *Future Aeronautical and Space Systems*, Vol. 172, *Progress in Astronautics and Aeronautics* (AIAA Press, Reston, VA, 1997), p. 251.; D. A. Fulghum "Miniature Air Vehicles Fly into Army's Future," *Aviation Week and Space Technology* (9 November 1998), p. 37.
5. J. M. McMichael and M. S. Francis, "Micro Air Vehicles—Toward a New Dimension in Flight," http://web-ext2.darpa.mil/tto/mav/mav_auvsi.html
6. *Proceedings, JPL Micropropulsion Workshop* (JPL, Pasadena, CA, 7-9 April 1997).
7. R.L. Bayt, A. A. Ayon, and K. S. Breuer, "A Performance Evaluation of MEMS-based Micronozzles," AIAA 97-3169 (Seattle, WA, 1997).
8. G. Stix, "Little Bangs," *Scientific American* **279** (5), 50–51 (November 1998).
9. R. A. Spores and M. Birkan, "The USAF Electric Propulsion Program," *Proceedings of 34th AIAA Joint Propulsion Conference*, AIAA 98-3181 (Cleveland OH, 1998).
10. G. S. Sutherland and M. E. Maes, "A Review of MicrorocketTechnology: 10-6 to 1 lbf Thrust," *J. Spacecraft and Rockets* **3** (8), 1153–1165 (August 1966).
11. S. Jacobson, "Aerothermal Challenges in the Design of a Microfabricated Gas Turbine Engine," AIAA 98-2545 (Albuquerque, NM, June 1998); E. S. Piekos, D. J. Orr, S. A. Jacobson, F. F. Ehrich, and K. S. Breuer, "Design and Analysis of a Microfabricated High Speed Gas Journal Bearings," AIAA 97-1966 (Snowmass, CO, June 1997).
12. F. F. Chen, *Introduction to Plasma Physics*, 1st ed. (Plenum Press, New York, 1974), p. 169-173.
13. M. Andrenucci, S. Marcuccio, and A. Genovese, "The Use of FEEP Systems for Micronewton Thrust Level Missions," AIAA 93-2390, *29th Joint AIAA/SAE/ASME/ASEE Propulsion Conference* (Monterey, CA, June 1993.)
14. R. Schmidt, *et al.*, "A Novel Medium-Energy Ion Emitter for Active Spacecraft Potential Control," *Rev. Sci. Instrum.* **64** (8), 2579-2584 (August 1993).
15. T. T. Tsong, *Atom-Probe Field Ion Microscopy* (Cambridge University Press, 1990).
16. *Ibid.*, p. 13.
17. *Ibid.*, p. 15.
18. M. L. Klein and J. A. Venables, *Rare Gas Solids*, Vol. 1 (Academic Press, New York, 1976), p. 146.
19. R. Bayt, K. S. Breuer, and A. A. Ayon, "DRIE-Fabricated Nozzles for Generating Supersonic Flows in Micropropulsion Systems," *Proceedings of the Solid-State Sensor and Actuator Workshop* (Hilton Head, SC, June 1998).
20. D. Hülsenberg, R. Bruntsch, K. Schmidt, F. Reinhold, *Mikromechanische Bearbeitung von fotoempfindlichem Glas*, Vol. 41 (Silikattechnik, 1990), p. 364.
21. R. Bayt, and K. S. Breuer, "Viscous Effects in Supersonic MEMS-Fabricated Micronozzles," *Proceedings of the 3rd ASME Microfluids Symposium* (Anaheim, CA, November 1998).
22. A. Beskok and G. Karniadakis, "A Model for Flows in Channels, Pipes, and Ducts at Micro- and Nano-Scales," *J. Microscale Thermophysics and Engineering*. In press.
23. J. C. Harley, Y. Huang, H. H. Bau, and J. N. Zemel, "Gas Flow in Micro-Channels," *J. Fluid Mechanics* **284** (1995).
24. K. C. Pong, C. M. Ho, J. Lui, and Y. C. Tai, "Nonlinear Pressure Distribution in Uniform Micro-channels," In FED Vol. 197, *Applications of Microfabrication to Fluid Mechanics* (ASME, 1994).
25. E. B. Arkilic, M. A. Schmidt, and K. S. Breuer, "Gaseous Slip Flow in Long Microchannels," *J. MicroElectroMechanical Systems* **6** (2) (June 1997).
26. G. Bird, *Molecular Gas Dynamics and the Direct Simulation of Gas Flows* (Oxford University Press, 1994).
27. E. S. Piekos and K. S. Breuer, "DSMC Modeling of Micromechanical Devices," *J. Fluids Engineering* **118**, 464–469 (1996).

28. D. B. Hash and H. A. Hassan, "A Hybrid DSMC/Navier-Stokes Solver," AIAA 95-0410 (Reno, NV, 1995).
29. D. J. Coe, M. G. Allen, M. A. Trautman, and A. Glezer, "Micromachined Jets for Manipulation of Macroflows," *Solid-State Sensor and Actuator Workshop* (Hilton Head, NC, 1994).
30. D. J. Coe, M. G. Allen, B. L. Smith, and A. Glezer, "Addressable Micromachined Jet Arrays," *Technical Digest: Transducers '95* (Stockholm, Sweden, 1995).
31. U. Ingard and S. Labate, "Acoustic Circulation Effects and the Nonlinear Impedance of Orifices," *J. Acoust. Soc. Am.* **22**, 211 (1950).
32. B. L. Smith and A. Glezer, "The Formation and Evolution of Synthetic Jets," *Phys. Fluids* **10** (9) 2281–2297 (1998).
33. M. Amitay, A. Honohan, M. Trautman, and A. Glezer, "Modification of the Aerodynamic Characteristics of Bluff Bodies Using Fluidic Actuators," AIAA-97-2004, *28th AIAA Fluid Dynamics Conference* (1997).
34. S. A. Jacobson and W. C. Reynolds, "Active Control of Streamwise Vortices and Streaks in Boundary Layers," *J. Fluid Mech.* **360**, 179–211 (1998).
35. R. Rathnasingham and K. S. Breuer, "Coupled Fluid-Structural Characteristics of Actuators for Flow Control," *AIAA J.* **35** (5), 832–837 (May 1997).
36. J. T. Lachowicz, C. S. Yao, and R. W. Wlezien, "Scaling of an Oscillatory Flow Control Actuator," AIAA-98-0330, *36th Aerospace Sciences Meeting* (1998).
37. K. Breuer and A. Padmanabhan, "The Design and Fabrication of Micromachined Actuator Disk," MIT FDRL internal report (1996).
38. A. H. Epstein, S. D. Senturia, G. Anathasuresh, A. Ayon, K. Breuer, K-S. Chen, F. E. Ehrich, G. Gauba, R. Ghodssi, C. Groshenry, S. Jacobson, J. H. Lang, C-C. Lin, A. Mehra, J. O. Mur Miranda, S. Nagle, D. J. Orr, E. Piekos, M. A. Schmidt, G. Shirley, M. S. Spearing, C. S. Tan, Y-S. Tzeng, and I. A. Waitz, "Power MEMS and Microengines," paper presented at the *IEEE Transducers '97 Conference* (Chicago, IL, June 1997).
39. A. H. Epstein, S. D. Senturia, O. Al-Midani, G. Anathasuresh, A. Ayon, K. Breuer, K-S. Chen, F. E. Ehrich, E. Esteve, L. Frechette, G. Gauba, R. Ghodssi, C. Groshenry, S. Jacobson, J. L. Kerrebrock, J. H. Lang, C-C. Lin, A. London, J. Lopata, A. Mehra, J. O. Mur Miranda, S. Nagle, D. J. Orr, E. Piekos, M. A. Schmidt, G. Shirley, M. S. Spearing, C. S. Tan, Y-S. Tzeng, and I. A. Waitz, "Micro-Heat Engines, Gas Turbines, and Rocket Engines—the MIT Microengine Project," AIAA 97-1773 (Snowmass Village, CO, June 1997).
40. A. Berezhoi, *Glass-Ceramics and Photo-Sitalls* (Plenum Press, New York, 1970).
41. W. W. Hansen, S. W. Janson, and H. Helvajian, "Direct-Write UV Laser Microfabrication of 3D Structures in Lithium-AlumoSilicate Glass," *SPIE* **2991**, 104 (1997).
42. "Foturan—a Material for Microtechnology," Corporate brochure no. 10095 e 11951.0 (IMM Institute für Mikrotechnik GmbH and Schott Glaswerke, Mainz, Germany).
43. K. D. Bean, "Anisotropic Etching of Silicon," *IEEE Transactions on Electron Devices* ED-25, 1185 (1978).
44. A. I. Stoller, "The Etching of Deep, Vertical-Walled Patterns in Silicon," *RCA Review* **31**, 271 (1970).
45. D. B. Lee, "Anisotropic Etching of Silicon," *J. Appl. Phys.* **40**, 4569 (1969).
46. M. Declerq, L. Gerzberg, and J. Meinol, "Optimization of the Hydrazine-Water Solution for Anisotropic Etching Silicon in Integrated Circuit Technology," *J. Electrochem. Soc.* **122**, 545 (1975).
47. A. Reisman *et al.*, "The Controlled Etching of Silicon in Catalyzed Ethylenediamine-Pyrocatechol-Water Solutions," *J. Electrochem. Soc.* **126**, 1406 (1979).
48. M. P. Wu, Q. H. Wu, and W. H. Ko, "A Study on Deep Etching of Silicon Using Ethylenediamine-Pyrocatechol-Water," *Sensors and Actuators* **9**, 333 (1986).
49. O. Tabata *et al.*, *Technical Digest, Transducers '91; IEEE Int. Conference On Solid-State Sensors and Actuators* (1991), p. 811.
50. U. Schnakenberg, *Technical Digest, Transducers '91; IEEE Int. Conference On Solid-State Sensors and Actuator* (1991), p. 815.

51. "The MOSIS Service," http://www.isi.edu/mosis.
52. "MUMPS™ MEMS Technology Applications Center," http://mems.mcnc.org/mumps.html.
53. "Technologies: SUMMiT Technology," http://www.mdl.sandia.gov/Micromachine/trilevel.html.
54. J. Marshall *et al.*, "Realizing Suspended Structures on Chips Fabricated by CMOS Foundry Processes Through the MOSIS Service," NIST Report no. NISTIR 5402 (June 1994). Available at http://www.mosis.org/pubs.
55. S. W. Janson, "Batch-Fabricated Resistojets: Initial Results," paper presented at the *International Electric Propulsion Conference* (Cleveland, OH, September 1997).
56. C. C. Lin, R. Ghodssi, A. Ayon, D-Z Chen, S. Jacobson, K. Breuer, A. Epstein, and M. Schmidt, "Fabrication and Characterization of a Micro Turbine/Bearing Rig," paper presented at *MEMS '99* (Orlando, FL, January 1999).
57. *Ibid.*
58. J. Mitterauer, "Miniaturized Liquid Metal Ion Sources (MILMIS)," *IEEE Transactions on Plasma Science* **19** (5), 790–799 (October 1991).
59. J. Mitterauer, "Prospects of Liquid Metal Ion Thrusters for Electric Propulsion," IEPC paper 91-105, *22nd AIDAA/AIAA/DGLR/JSASS International Electric Propulsion Conference* (Viareggio, Italy, October 1991).
60. C.A. Spindt, "Microfabricated Field-Emission and Field-Ionization Sources," *Surface Sci.* **266**, 145–154 (1992).
61. E. Stuhlinger, *Ion Propulsion for Space Flight* (McGraw-Hill, New York, 1964) 257.
62. M. W. Crofton, "Evaluation of Electric Thrusters," Aerospace Corp. Report no. ATR-97(8201)-1 (April 1997).
63. T. W. Haag, "Thrust Stand for High-Power Electric Propulsion Devices," *Rev. Sci. Instrum.* **62** (5), 1186 (1991).
64. G. P. Sutton, *Rocket Propulsion Elements*, 6th ed. (John Wiley & Sons, New York, 1992), pp. 416–422.
65. A. Globus, "The Design and Visualization of a Space Biosphere," *10th Biennial Space Studies Institute/Princeton University Conference on Space Manufacturing* (Princeton University, 15 May 1991).
66. Z. Shaoxian, G. Jingkun, Y. Zhixiong, C. Jie, and C. Wanpeng, "Sub-micrometer Hydroxyapatite Bioceramics," *Mat. Res. Soc. Symp. Proceedings*, Vol. 292 (1993) p. 225.
67. V. J. Laraia and A. H. Heuer, "The Microindentation Behavior of Several Mollusk Shells," *Mat. Res. Soc. Symp. Proceedings,* Vol. 174, 125 (1990).
68. M. Sarikaya, K. Gunnison, and I. Aksay, "Seashells as a Natural Model to Study Ceramic-Polymer Composites," *Proceedings 47th Annual Meeting of the Electron Microscopy Society of America* (San Antonio, TX, 6–11 August 1989), p. 558.
69. M. Sarikaya, K. E. Gunnison, M. Yasrebi, and I. A. Aksay "Mechanical Property-Microstructural Relationships in Abalone Shell," *Mat. Res. Soc. Symp. Proceedings,* Vol. 174, 109 (1990).
70. *Ibid.*
71. E. S. Lipinsky, R. A. Nathan, W. J. Sheppard and J. L. Otis, "Systems Study of Fuels from Sugercane, Sweet Sorghum and Sugar Beets, Vol 3.: Conversion to Fuels and Chemical Feedstocks, Final Report, Task 77," U.S. Dept. of Commerce Report no. BMI-1957-V-3 (1976), p. 91.
72. J. C. Alt and W. C. Shanks III, "Sulfur in Serpentinized Oceanic Peridotites: Serpentinization Processes and Microbial Sulfate Reduction" *J. Geophysical Res.* **103**, 9917 (1998).
73. R. E. Cameron, F. A. Morelli, and H. P. Conrow, "Survival of Microorganisms in Desert Soil Exposed to Five Years of Continuous Very High Vacuum" NASA/JPL Technical Report no. 32-1454 (15 March 1970); P. J. Geiger, F. A. Morelli, and H. P. Conrow, "Effects of Ultrahigh Vacuum on Three Types of Microorganisms," JPL Space Programs Summary no. 37-27, Vol. IV (30 June 1964), p. 109.
74. R. Jaenicke, G. Bernhardt, H.-D. Ludemann, and K. O. Stetter, "Pressure-Induced Alterations in the Protein Pattern of the Thermophilic Archaebacterium Methanococcus Thermolithotrophicus" *Appl. Environ. Microbiol.* **54**, 2375 (1988).

75. A. A. Yayanos, "Evolutional and Ecological Implications of the Properties of Deep-Sea Barophilic Bacteria," *Proceedings Nat'l. Acad. U.S.A*, Vol. 83, 9542 (1986).
76. D. H. Bartlett, C. Kato, and K. Horikoshi, "High- Pressure Influences on Gene and Protein Expression," *Res. Microbiol.* **146**, 697 (1995).
77. D. H. Bartlett, "Microbial Life at High Pressures," *Sci. Progress Oxford* **76**, 479 (1992).
78. E. S. West and W. R. Todd, *Textbook of Biochemistry*, 3rd ed. (Macmillan Press, New York, 1961), p. 786.
79. "Technology for the Conversion of Solar Energy to Fuel Gas," Univ. of Pennsylvania National Center for Energy Management and Power, Philadelphia, Report no. NSF/RANN/SE/GI27976/72/4 (31 January 1973).
80. G. L. M. Christopher, "Biological Production of Methane from Organic Materials (Biomethane Project)," final report to Columbia Gas Service Corp. (1971), p. 84.
81. *Ibid.*
82. R. E. Freeland, G. D. Bilyeu, G. R. Veal, and M. M. Mikulas, "Inflatable Deployable Space Structures Technology Summary," *49th International Astronautical Congress* (Melbourne, Australia, 28 September – 2 October 1998).

Index

A

B

C

D

E

F

G

H

I

K

L

M

N

O

P

Q

R

S

T

U

V

W

Y