VLSI Electronics
Microstructure Science

A Treatise Edited by

Norman G. Einspruch
College of Engineering
University of Miami
Coral Gables, Florida

VLSI Electronics Microstructure Science

Volume 19

Advanced CMOS Process Technology

J. M. Pimbley
Department of Mathematical Sciences
Rensselaer Polytechnic Institute
Troy, New York

M. Ghezzo
H. G. Parks
D. M. Brown
General Electric Corporate
Research and Development Center
Schenectady, New York

ACADEMIC PRESS, INC.
Harcourt Brace Jovanovich, Publishers
San Diego New York Berkeley Boston
London Sydney Tokyo Toronto

Academic Press, Inc.
San Diego, California 92101

United Kingdom Edition published by
Academic Press Limited
24–28 Oval Road, London NW1 7DX

Library of Congress Cataloging in Publication Data

Microstructure science

(VLSI electronics ; v. 19)
Includes index.
1. Metal oxide semiconductors, Complementary
--Design and construction. I. Pimbley,
Joseph M. II. Series.
TK7874.V56 vol. 19 621.381'73 s 88-16691
[TK7871.99.M44] [621.3815's]
ISBN 0-12-234119-8 (alk. paper)

Printed in the United States of America
89 90 91 92 9 8 7 6 5 4 3 2 1

Preface

The microelectronics revolution persists. Technical innovations abound, and the performance-to-cost ratios for semiconductor devices, integrated circuits, and systems continue to grow. Although it might be entertaining, a historical account of the past three decades of microelectronics would provide little direct benefit to the industry. This volume of the *VLSI Electronics* series, titled *Advanced CMOS Process Technology,* provides a current snapshot of one highly pertinent domain of microelectronics. For reasons discussed within the text, CMOS (complementary metal–oxide–semiconductor) technology plays a leading role in present and future electronic systems.

In choosing appropriate material for this monograph, we specified two selection criteria. First, we sought topics of primary importance to the present and future state of the art of CMOS process technology. Second, where constraints of space and time imposed on the number of topics to cover, we focused on issues with the least amount of coverage in other forums. Aside from introductory comments and background on CMOS device and circuit considerations, we narrowed our topic list to metallization, isolation techniques, reliability, and yield. The reader should not infer that omitted areas, including lithography and etching techniques, are of inferior rank. Rather, such topics have enjoyed a good deal of explicit scrutiny in, for example, earlier volumes of this *VLSI Electronics* series.

Finally, we note that it has been and is our goal to contribute to the global microelectronics industry by reporting as clearly as possible the present status of the CMOS process technology issues we chose to communicate. Furthermore, we attempted to project as accurately as possible expected future evolution. This contribution is transitory. We expect the industry to surpass the technical content of this monograph through innovation, invention, and, à la Thomas Edison, perspiration. In fact, we dedicate this volume to the engineers, scientists, and technical managers who will render obsolete many of the technical concerns we voice.

We express our deep gratitude to the General Electric Corporate Research and Development Center for many forms of support. Beyond merely countenancing our preparation of this manuscript as a tolerable extracurricular exercise, we felt a strongly supportive and encouraging environment. In particular, we thank William R. Cady and Kirby G. Vosburgh. The excellent and gracious secretarial skills of Elizabeth A. Harris and Marcia E. Vinick provided great assistance. One of us (JMP) acknowledges the continual encouragement of Harold J. Raveché (formerly Dean of Science at the Rensselaer Polytechnic Institute and now President of the Stevens Institute of Technology) and conversations with Igor Bol of the Xerox Corporation.

J. M. Pimbley
M. Ghezzo
H. G. Parks
D. M. Brown

Chapter 1

Introduction

We contemplate the microelectronics revolution of the twentieth century from a broad perspective. Current theories of evolution argue that the human race evolved from other animal species in a quantized manner. That is, evolution does not proceed at a uniform rate. Relatively short bursts of mutation punctuate longer periods of evolutionary dormancy. Such an understanding is consistent with the importance of fluctuations, deviations from the mean behavior, in nature. On a larger scale, distribution of matter in the universe is exceedingly nonuniform.

Advances in technology and human society also arrive in discrete steps. Consider the improvement of the standard of living in the past 200 years. The Industrial Revolution of the late nineteenth century yielded drastic productivity gains. With new inventions and manufacturing capability for farm equipment, agricultural production exploded as the need for human resources declined. Though many contemporary politicians would likely decry such a situation, society benefitted enormously. With the sudden drop in the human and economic price of agricultural production, both resources became available for the pursuit of "higher order" human needs such as medicine, manual labor reduction, and scientific research. The Industrial Revolution enabled this century's exponential population growth by loosening the shackle of one basic need.

It is also fair to say that the Industrial Revolution was a necessary precursor for all of the world's more recent technological gains. These gains include advances in medical techniques, transportation, electronics, microelectronics (one portion of which is the focus of this monograph), and biotechnology. Would any wise person have predicted most of these benefits in the initial stages of the Industrial Revolution? We think not. The point of this question is that advances in technology will likely precipitate

many gains in society beyond those that are easily recognized. We must therefore list anticipated benefits of any technology, as we will later for CMOS microelectronics, with humility.

Electronics, and particularly microelectronics, spawned the field of numerical computation. It is difficult to imagine the lack of such capability as recently as 40 years ago. Richard Feynman and R. Leighton recounted the method of computation of this era [1]. To evaluate a mathematical expression expressing a result of theoretical physics required a roomful of technicians and an assembly line operation. Such tasks today require inexpensive calculators. Computers in the worlds of finance, accounting, marketing, retailing, etc., greatly expedite previously expensive tasks. Beyond this tremendous increase in productivity arose the new methodology of numerical computation for solving previously intractable scientific, engineering, and mathematical problems.

Attempts to enumerate a representative, let alone complete, list of benefits of microelectronics to the development of technology and society are futile. Telecommunications has enjoyed revolutionary expansion. Numerical solution of mathematical models relevant to fluid flow, semiconductor physics, and optimization have drastically reduced the design costs of airplanes and automobiles, semiconductor devices, and sundry manufacturing processes. Many consumer products, including automobiles, now contain microprocessors for improved reliability and performance. Noninvasive medical imaging techniques, such as the computer-assisted tomography (CAT) scan, magnetic resonance, and ultrasound imaging, require sophisticated computational capability. Productivity in the office environment and publishing industries has surged forward with the advent of computer-assisted typing and text processing. On a "slow" day, the New York Stock Exchange trades two hundred million shares of stock with current automated trading systems. Ironically, some interested observers blame computers for the near collapse of the financial system on October 19–20, 1987. Such an assessment is inaccurate because the computer system merely executed the instructions of several of the largest pension funds and brokerage houses.

As the reader suspects by now, the two key benefits of technological advances are increased productivity (for established tasks) and spontaneous creation of new capabilities. Improvement in the standard of living of the human race can arise only from increased productivity [2]. Government decree, legislation, or regulation that does not benefit productivity cannot improve the economic well-being of citizens. Furthermore, watershed events such as the Industrial Revolution and microelectronics revolution, which precipitate gigantic leaps in the standard of living, can only germinate, blossom, and grow in the fertile ground of political, personal,

and economic freedom. The justification of this statement lies in the nature of technological breakthroughs. Such breakthroughs begin as accidental experimental findings, ingenious ideas, straightforward adaptation of existing technologies, or some combination of these elements. It is never clear in these early stages if any useful product or process will eventually emerge. In fact, it is most likely that nothing useful will emerge. For every great invention we celebrate exist thousands that failed to match its inventor's hopes.

Frankly, there is no person or group of people smart enough to reliably predict which emerging ideas will succeed (i.e., benefit society) and which will not. Only society can decide by its reaction to the new product in the free market. Who brings the product to market? Assuming the existence of political and personal freedom, the inventor is entitled to do so. (In the absence of such freedom, the oppressive government may appropriate the idea in the unlikely circumstance that the inventor chooses to make this disclosure.) To bring his/her product to market, the inventor must invest the necessary capital. Thus, the inventor assumes financial risk and stands to gain or lose in concert with the technical or market success of the product. Where innovation is required, this system functions well since the inventor has great motivation (i.e., his/her own financial well-being) to invent and bring to market a product in a form that best meets society's needs. Furthermore, the inventor is disinclined to pour capital into an idea that, as is most likely, is failing because of technical or market reasons. Government intervention, even in a predominantly free market, dampens this process and depresses the standard of living. An apparently innocuous government program of lessening the risk to the inventor (by subsidy, for example) is still detrimental since the inventor is likely to waste more money before abandoning his/her project. The inventor understandably considers the taxpayer's money as disposable.

The relationship between political, personal, and economic freedom and microelectronics motivates our brief foray into economic issues. As much or more so than other technological areas, microelectronics requires a great deal of innovation. The technical challenges have been, are, and will be formidable. Continued progress feeds on brilliant ideas of scientists and engineers of many disciplines and flexibility and calculated risk-taking of corporate managers. Freedom is essential to progress in microelectronics. A quick survey of world microelectronics finds the dominant activity in North America (United States and Canada), Southeast Asia (Japan and South Korea), and western Europe. With the possible exceptions of Hong Kong and Taiwan, this list corresponds to that of the regions of greatest economic freedom. Western Europe lags North America and Southeast Asia primarily because of restrictions on economic freedom of its citizens.

Similarly, the technical backwardness of socialist countries in eastern Europe, Asia, and the Third World does not stem from any inherent deficiency of the citizens of these countries. Rather, the absence of economic freedom intrinsic to socialism and the loss of political freedom, which generally accompanies socialism, do not permit the necessary innovation. Of course, the problems of the Third World go far beyond a failure to foster microelectronics growth. The standard justification for barriers to free trade in the Third World is the need to "protect" the fragile economy. It is this protection and other socialist measures that have impoverished a large fraction of the human race. The continuing tragedy of starvation in Ethiopia is abetted by the counter-productive, dogmatic policies of the socialist government.

Having explored the social context of technological advances in general and microelectronics in particular, we will focus now, and for the remainder of this monograph, on technical aspects. Microelectronics is the discipline of designing and fabricating electrical circuits from discrete components integrated into a single semiconductor sample. This field owes its existence to many scientific and engineering accomplishments. Two of these are the inventions of the metal–oxide–semiconductor field-effect transistor (MOSFET) in the 1930s and the bipolar transistor in the 1940s. Electrical circuits previously constructed with discrete components wired together or with vacuum tubes as circuit elements are much less expensive, and hence more readily employed, with the advent of microelectronics. Furthermore, microelectronics allows the construction of previously impractical or impossible circuits.

Requirements of a particular application generally dictate the choice between MOS and bipolar technologies. The main advantages of bipolar circuits are fast switching and high current drive, while MOS enjoys higher input impedance, lower power dissipation, and greater packing density. The MOS technology has gradually displaced bipolar technology in many applications but will not do so completely. There will always exist situations, such as central processor units, in which high speed is critical. We will focus on the most promising aspect of MOS technology. Though beyond the scope of this manuscript, we alert the reader to recent attempts to merge MOS and bipolar technologies within the same substrate [3].

The three categories of metal–oxide–semiconductor (MOS) technology are NMOS (*n* channel), PMOS (*p* channel), and CMOS. Our choice of MOS technology implies a restriction to silicon as the semiconductor material because of its ability to form uniform oxide layers of high quality. The semiconductor substrate of *n*-channel devices is *p*-type, while that of *p*-channel devices is *n*-type. Electrons are the dominant charge carrier for current flow in *n*-channel devices, while holes pay a similar role in *p*-chan-

nel technology. Complementary metal–oxide–semiconductor (CMOS) technology employs both *n*-channel and *p*-channel devices, while NMOS (PMOS) is exclusively *n* channel (*p* channel).

PMOS served as the primary technology for the first MOS circuits in the late 1960s. Since the silicon–silicon dioxide interface tends to develop a positive charge [4], the PMOS choice allows effortless fabrication of enhancement-mode (negative threshold voltage) devices, and this property ingratiated PMOS with the semiconductor world. Improved methods of impurity diffusion and ion implantation soon allowed the fabrication of NMOS enhancement-mode devices. NMOS quickly displaced PMOS in view of the inherently higher mobility of electrons relative to holes [5]. The higher mobility allows greater performance (i.e., current drive, switching speed) for equal circuit size or smaller circuit size for equal performance with NMOS as opposed to PMOS. Inherent advantages of PMOS include relative insensitivity to ionizing radiation and reduced hot carrier instability. This latter observation coupled with the reduced disparity between electron and hole mobilities in the high electric fields of short-channel MOSFETs even prompted the suggestion that PMOS may be the superior technology as MOSFET channel lengths decrease [6]. This prediction will likely never be realized since PMOS possesses other disadvantages in terms of short-channel device fabrication. It is difficult to fabricate shallow source–drain junctions with conventional *p*-type dopants in silicon. For other material reasons, it is also more challenging to fabricate *p*+ polysilicon gates for PMOS than it is to fabricate *n*+ polysilicon gates for NMOS.

The debate between NMOS and PMOS superiority in the short-channel limit is not of great importance since there exist overwhelming advantages for CMOS technology. In fact, all submicron MOS technology will be CMOS so that one might have easily entitled this book "Advanced MOS Process Technology." CMOS circuit design combines *n*-channel and *p*-channel devices in such a manner to reduce the standby current by orders of magnitude relative to pure NMOS or PMOS implementations. We discuss this justification of CMOS at length in Chapter 2. Suffice it to say for now that the heat removal problem associated with power dissipation in NMOS and PMOS circuits poses a formidable barrier. Other advantages of CMOS include system complexity, circuit considerations, and device reliability issues. The greatest CMOS disadvantages arise from the design and processing layout for closely spaced *n*-channel and *p*-channel devices with the ever present constraint of latch-up immunity. Again, Chapter 2 will focus on these issues.

The microelectronics industry has struggled, and continues to struggle, to reduce the physical dimensions of individual devices. In CMOS fabrication, such a goal implies the realization of short-channel MOS field-effect

transistors (*n* channel and *p* channel) as well as adequate, high-density methods for interconnecting these devices. Diminution of circuits and whole systems results from the scale reduction of individual devices. Economic forces drive this miniaturization. Size reduction permits fabrication of a greater number of chips per wafer with generally a higher percentage yield of defect-free products. For reasons directly and indirectly tied to device size, circuit and system performances improve as devices shrink. Competition requires manufacturers to strive continually to upgrade their product through miniaturization and pass the manufacturing cost savings onto the consumer. The introduction and dramatic technical and economic metamorphosis of the hand-held calculator in the 1970s serves as an excellent example of the evolution of semiconductor manufacturing. The personal computer of the 1980s distributed and decentralized an enormous amount of computational power. Engineering work stations employ the latest technology to pack the memory and power of a superminicomputer into a desktop unit. Industry executives and analysts now (early 1988) discuss plans for desktop supercomputers.

Successful definition and development of a CMOS fabrication process with reduced design rules require the integration of many individual efforts. One must develop the basic tools of metallization, isolation techniques, lithography, etching, and impurity profile adjustment. Device size reduction exacerbates several device reliability mechanisms and hence necessitates device and circuit design to assure reliability. Finally, one must identify applications and design circuits to capitalize on the miniaturized fabrication process. This last task is not, as one might suspect, trivial. Certainly there do exist some functions, such as memory, for which the motives and design methods of miniaturization are evident. Identification of new and previously impractical circuit *and* system concepts rendered feasible by advances in process technology is essential. Such identification requires great ingenuity and can provide the raison d'être for investment in advanced process development.

Convenient forums for definition of past, present, and future technology are two ubiquitous circuits known as the dynamic random access memory (DRAM) and static random access memory (SRAM). Both the DRAM and SRAM are two-dimensional arrays of memory cells. The SRAM cell consists of two inverters with additional FETs for array addressing. The DRAM cell requires only a FET and capacitor per cell and hence is more compact. Memory retention in the DRAM is, however, unstable (i.e., dynamic), so that the peripheral circuitry and system must bear the burden of periodically refreshing the stored information. Both SRAMs and DRAMs find wide application in computational systems. Since the SRAM

and DRAM cell designs have changed very little in the past 15 years, advances in process technology propagate quickly to these memory chips. Device reduction allows more memory per area of silicon.

Armed with a 12-μm design rule manufacturing process, Intel marketed a SRAM with 256 bits (cells) of memory in 1969. Toshiba announced a 256-K SRAM in 1984 with a 1.2-μm process [7]. The SRAM performance as measured by cell access time improved by a factor of 20, while the storage capacity ballooned by three orders of magnitude. Figure 1.1 portrays this trend graphically [8]. From similarly humble beginnings, 1-M (one megabit) DRAMs have emerged and are now available on the open market. Quite symbolic is the restriction of the "old" and still pervasive personal computer operating system MS-DOS to 640 kbytes of dynamic memory. The hardware outstripped the software to the point that this operating system cannot handle one bank of dynamic memory chips. SRAMs with 1-M capacity and DRAMs at 4 M will require a 0.8-μm design rule process with some increase in chip size. Recent reports discuss technology development [7] and reduction to practice in the laboratory [9] for this high-density DRAM. With innovative solutions for, or elegant sidestepping of, alpha particle immunity, reliability concerns, and subthreshold leakage problems, continued device reduction will produce DRAMs in the 16–64-M range.

We have documented the benefit of reduced feature size in CMOS

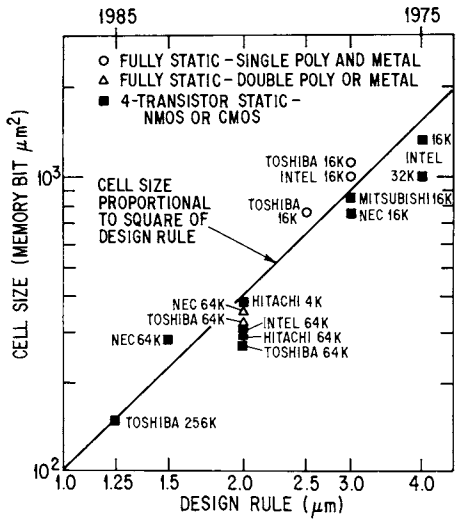


Fig. 1.1. We depict continued SRAM miniaturization as originally presented by Brown *et al.* [8]. © 1986 IEEE.

microelectronics. Given the enormous investment required to investigate and develop high-resolution fabrication processes, contemplation and prediction of ultimate limits become invaluable guides in planning technical goals. Fundamental and practical limits will conspire to frustrate miniaturization at some as yet unspecified level. Examples of fundamental limits are the thermal voltage fluctuations of the order of kT/q and the turn-on slope of a forward-biased diode. Practical limits are far more vexing and include identifying metals with a highly restrictive set of properties and the formation of ultrashallow source–drain junctions. We direct the reader to review Banerjee and Bordelon [10] for an excellent and philosophical discussion of this subject.

Furthermore, the ultimate minimum device size will certainly be a function of the specific design philosophy and application [11–13]. Consideration of noise immunity permits shorter channel lengths in digital logic circuits implemented in enhancement/enhancement CMOS, for example, as compared to enhancement/depletion CMOS [11]. Logic gates in turn are relatively insensitive to subthreshold leakage. This leakage contributes mainly to undesirable power dissipation in this implementation. In transmission gates, however, and particularly in dynamic memory cells, subthreshold conduction can destroy the desired function. Present theoretical estimates find that silicon MOSFETs with effective channel lengths in the area of 0.1 μm and exceedingly low subthreshold conduction at room temperature are manufacturable given the appropriate and presently unavailable technology [14]. A recent research effort has successfully produced a 0.1-μm channel length MOSFET for low-temperature operation [15].

In this monograph, we seek to describe the present state of the art in the technical areas most relevant and unique to advanced CMOS process technology. We give a brief background discussion on device and circuit considerations in Chapter 2. Our hope is that the reader will find this section useful for subsequent appeals to fundamental knowledge in later chapters. Chapter 3 discusses choices and techniques for metallization based on numerous practical constraints and considerations. Interconnection of short-channel devices is, of course, crucial to full realization of density gains in miniaturization. Dimensional reduction of metal lines is fraught with difficulties related to etching, reliability, and performance issues. We also focus on the improved technique of unframed contacts for density enhancement. Chapter 4 reviews the most important and promising isolation techniques. Such techniques are of central importance to present and future CMOS technology since the requirement of latch-up immunity opposes the goal of high packing density. We devote an extended discussion to reliability concerns (hot carriers, electromigration,

and oxide wear-out) in Chapter 5. One must recognize reliability as an essential ingredient of technological development in the earliest phase. An exposition of yield in present and future advanced CMOS processes in Chapter 6 concludes this volume.

REFERENCES

1. R. P. Feynman and R. Leighton, "Surely You're Joking, Mr. Feynman." W. W. Norton and Company, New York, 1985.
2. A. Smith, "Wealth of Nations," Harvard Classics version. P. F. Collier and Son, New York, 1909.
3. See, for example, T. Ikeda, A. Watanabe, Y. Nishio, I. Masuda, N. Tamba, M. Odaka, and K. Ogiue, High-speed biCMOS technology with a buried twin well structure, *IEEE Trans. Elec. Dev.* **ED-34,** 1304, 1987.
4. See, for example, E. H. Nicollian and J. R. Brews, "MOS Physics and Technology." Wiley, New York, 1982.
5. See, for example, S. M. Sze, "Physics of Semiconductor Devices." Wiley, New York, 1981.
6. E. Takeda, Y. Nakagome, H. Kume, N. Suzuki, and S. Asai, Comparison of characteristics of n-channel and p-channel MOSFETs for VLSI, *IEEE Trans. Elec. Dev.* **ED-30,** 675, 1983.
7. M. Isobe, J. Matsunaga, T. Sakurai, T. Ohtani, K. Sawada, H. Nozawa, T. Iizuka, and S. Kohyama, A 46 ns 256 K CMOS RAM, *IEEE Int. Solid State Circ. Conf. Tech. Dig.,* 214,1984.
8. D. M. Brown, M. Ghezzo, and J. M. Pimbley, Trends in advanced process technology —submicrometer CMOS device design and process requirements, *Proc. IEEE* **74,** 1678, 1986.
9. P.-L. Chen, A. Selcuk, and D. Erb, A double-epitaxial process for high-density DRAM trench-capacitor isolation, *IEEE Elec. Dev. Lett.* **EDL-8,** 550, 1987.
10. S. Banerjee and D. M. Bordelon, A model for the trench transistor, *IEEE Trans. Elec. Dev.* **ED-34,** 2485, 1987.
11. J. D. Meindl, Theoretical, practical and analogical limits in ULSI, *1983 IEDM Tech. Dig.,* 8, 1983.
12. J. R. Pfiester, J. D. Shott, and J. D. Meindl, Performance limits of CMOS ULSI, *IEEE Trans. Elec. Dev.* **ED-32,** 333, 1985.
13. L. L. Lewyn and J. D. Meindl, Physical limits of VLIS DRAMs, *IEEE Trans. Elec. Dev.* **ED-32,** 311, 1985.
14. J. M. Pimbley and J. D. Meindl, MOSFET scaling limits, to be published in *IEEE Trans. Elec. Dev.,* 1988.
15. G. A. Sai-Halasz, M. R. Wordeman, D. P. Kern, E. Ganin, S. Rishton, H. Y. Ng, D. S. Zicherman, D. Moy, T. H. P. Chang, and R. H. Dennard, Experimental technology and characterization of self-aligned 0.1μm-gate-length low-temperature operation NMOS devices, *1987 IEDM Tech. Dig.,* 397, 1987.

Chapter 2

CMOS Device and Circuit Background

Fabrication of the MOSFET is arguably the central theme of the CMOS process flow. Virtually all manufacturing steps seek to define the MOSFET structure, isolate adjacent MOSFETs, and interconnect and passivate MOSFETs. The preponderance of this basic device in CMOS digital and analog circuits necessitates this processing focus.

When we speak of CMOS miniaturization and design rule reduction, the first and foremost consideration is the fabrication of smaller MOSFETs with improved isolation and interconnection techniques. Physical delineation of ever smaller structures by advances in lithography, etching, ion implantation, and metallization is certainly a major challenge. These issues spring easily to mind in any reflection on advanced CMOS processing techniques.

But there is much more to the subject of processing techniques. Mere fabrication of small structures ignores the driving motivation of producing the next generation of low-cost, superior performance integrated circuits. Device performance considerations must assume great importance during conception and definition of the CMOS process flow. In addition, low cost requires assurance of device and circuit reliability and high manufacturing yield. Thus, one must view superior performance, reliability, and high yield as specifications of the process flow on an equal footing with more conventional requirements, such as nominal MOSFET gate length and minimum lithographic feature size.

Subsequent chapters will address the physical delineation issues of metallization and device isolation as well as reliability and yield. This chapter will first discuss device physics relevant to the formulation of an advanced CMOS fabrication process. Our goal is to provide a complete discourse on

this specialized topic to which the reader may refer while studying other chapters of this text. We also recommend the more expansive books by Sze [1], Grove [2], Ghandhi [3], and Nicollian and Brews [4]. We then address CMOS circuit design considerations and compare the CMOS and NMOS technologies. A description of the CMOS latch-up mechanism follows.

While it is reasonable to consider the MOSFET as the most fundamental device in any CMOS process, one's physical understanding is facilitated by considering the junction diode and the MOS capacitor. The next two sections discuss these basic devices. We follow with a description of the MOSFET. Beyond the description of MOSFET operation, we analyze the concepts of scaling theory and the related issue of short-channel effects.

I. JUNCTION DIODE

The great utility of semiconducting materials stems from the ease with which one may alter material conduction properties. Substitution of impurities such as phosphorus, arsenic, and antimony from column V of the periodic table for the host atoms of an elemental (column IV) semiconductor, even in small concentrations, drastically enhances the density of electrons in the semiconductor conduction band. This impurity "doping" renders the electron density far greater than the hole density since column V (donor) impurities essentially donate electrons to the host semiconductor. A hole is simply an unfilled electron energy state in the semiconductor valence band. Introducing impurities such as boron and gallium from column III of the periodic table has precisely the opposite effect. These column III (acceptor) impurities essentially remove electrons from the valence band and leave behind holes.

Consider the thought experiment in which we produce two semiconductor samples. In one sample, we introduce a uniform concentration of donor impurities, while, in the other, we specify a uniform concentration of acceptor impurities. In both specimens, the total charge density is zero. While the *n*-type (donor impurity) sample has many more electrons than holes, the total negative charge of this electron-to-hole excess is exactly compensated by the positively charged donor impurities. Similarly, the *p*-type (acceptor impurity) sample excess hole positive charge is exactly compensated by the negatively charged acceptor impurities.

Suppose we bring the *n*-type and *p*-type silicon samples into intimate contact. What happens? The conduction band electrons and valence band holes are free to move by drift or diffusion while the charged impurities are immobile. Electrons and holes will quickly begin diffusing into the *p* region and *n* region, respectively, because of the large concentration gradients.

That is, there will begin a net diffusion of electrons from the *n*-type silicon to the *p*-type silicon simply because there are many more electrons in the *n*-type sample than in the *p*-type sample. As electrons leave from the *n*-type region, holes pour into this region from the *p*-type side. Clearly, the loss of negative charge and the influx of positive charge will impart a net positive charge to the initially neutral *n*-type region. Conversely, the *p*-type region becomes negatively charged.

The magnitude of the positive charge of the *n* region balances exactly the negative charge of the *p* region. An electric field develops in concert that tends to inhibit further out-diffusion of mobile charge. Thus, this system of an *n*-type semiconductor sample and a *p*-type sample in contact reaches a steady state in which the *n*-type region is at a positive potential relative to the *p*-type region and there is zero net current flow of either mobile carrier (electrons or holes). Figures 2.1(a) and (b) sketch the impurity concentration and energy bands (showing the potential difference), respectively, of this structure (known as a junction diode).

It may seem that we have progressed with excessive caution through a description of the one-dimensional junction diode with no voltage applied.

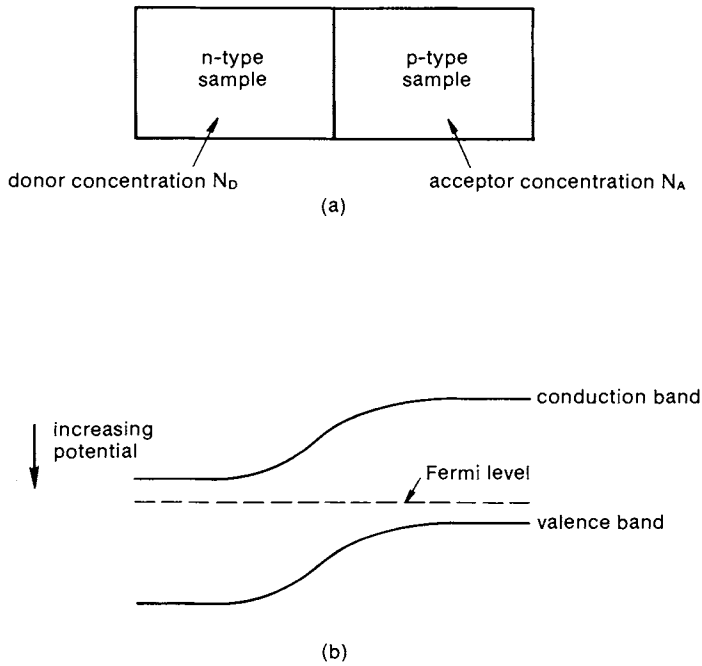


Fig. 2.1. We sketch the impurity concentration (a) and energy bands (b) of the one-dimensional junction diode.

Clearly, this must be the simplest imaginable semiconductor device. But the concept of a potential difference between the p and n regions due to diffusion of mobile charge is of central importance in the modeling and understanding of all devices. This "built-in" potential difference exists independent of any applied bias to the two ends of the diode.

Unfortunately, it is too easy to develop the wrong perception of the built-in potential. A common belief is that this potential is in some sense not real. Quite often we are shown an energy band diagram such as that of Fig. 2.1(b) and are told that the potential across the diode is zero when we apply equal external potentials to the p and n regions. However, the story is not complete. If we place metal contacts on the two ends of the diode and short the two contacts by means of an external lead connecting the contacts, we will indeed have zero potential from one contact to the other. However, in this case, there will also be band "bending," or curvature within the silicon at both metal contacts. The physical origin of the electric field at the metal contact is equivalent to that at the junction between the p region and n region. There is net charge diffusion in one direction until the diffusion is balanced by the induced potential differences. In the case of the shorted diode, then, the sum of the three voltage drops ($p-n$ junction, metal to n region, and metal to p region) is precisely zero.

We now focus on the important electrical characteristics of the junction diode: voltage-dependent current flow and capacitance. As we described above, applying zero potential difference to the two ends of the diode (by "shorting" these ends with an external lead) elicits zero net current. The built-in potential of the junction is an energy barrier that inhibits the diffusion of electrons and holes from their respective high-concentration regions. Decreasing this barrier height with the application of a positive potential to the p-type silicon region ignites a mass exodus of electrons and holes from the n-type and p-type regions, respectively. Current increases exponentially with this barrier height reduction. Thus, the forward (positive bias on p-type region) current–voltage characteristic shows an exponential rise until series resistance, internal and external to the diode, limits the fraction of applied bias translated into barrier height lowering.

Placing a negative bias on the p-type region (reverse bias) produces a very different result. The reverse bias increases the barrier height for diffusion of electrons and holes from the respective high-concentration regions. Current flow with this bias polarity, in the idealized and unrealizable case of zero charge generation, results from the net removal of electrons from the p-type region and holes from the n-type region. Since the minority charge carrier concentration is exceedingly small, the diode reverse current is correspondingly miniscule. In silicon diodes, the reverse current is, in

fact, dominated by the effect of spontaneous charge generation precisely because of the low intrinsic (generation-independent) reverse current.

We digress briefly to explain charge generation. Reference to electron concentration is reserved only for conduction electrons. We ignore electrons tightly bound to a silicon atom since such bound carriers contribute nothing to current flow. That is, they do not move in response to an electric field. The most loosely bound of these immobile electrons are in the outer (valence) shell of the silicon atom. The binding energy of these valence electrons is, in a semiconductor, the bandgap energy. At the absolute zero of temperature, there are no conduction electrons in a semiconductor. At any nonzero temperature, some valence electrons will be excited to the conduction band. In fact, there will develop a continuous generation process in which valence electrons spontaneously receive energy and "jump" to the conduction band as well as a recombination process in which conduction electrons may emit energy and "fall" to an empty valence electron state. The inevitable existence of impurities in the semiconductor lattice, even in vanishingly small concentrations, drastically enhances the rates of generation and recombination by introducing additional electron energy states intermediate to those of the valence and conduction bands. We refer the reader to the excellent pioneering work of Hall [5] and Shockley and Read [6].

Another important charge generation mechanism is impact (avalanche) ionization. Electrons and holes tend to gain energy from an electric field and lose the acquired energy through collisions with the silicon lattice. At very high electric fields, such collisions cannot dissipate the excess charge carrier energy efficiently. Consequently, charge carrier energy can increase to the point where it is possible for this energetic carrier to exchange its energy for the promotion of an electron from the valence band to the conduction band. The classical picture is that of a conduction electron, or valence hole, colliding with a valence electron and imparting the necessary binding (bandgap) energy to liberate the valence electron. This avalanche generation process is exponentially dependent on the electric field strength and is most common in reverse-biased junction diodes. At some large reverse bias, a diode will abruptly begin conducting a great deal of current with a small increase in bias. This "breakdown" voltage represents the maximum voltage at which one may operate the diode.

In equilibrium, the processes of generation and recombination balance exactly so that the *net* generation is zero. Application of a reverse diode bias, however, depletes the minority carrier concentration (i.e., electrons in the p region and holes in the n region) and thus renders generation dominant over recombination. Thus, the net generation produces electrons and

holes spontaneously. These generated carriers drift with the electric field and form the bulk of the reverse current.

At this point, we introduce the approximate, yet important, concept of the depletion region. Even though the one-dimensional junction diode is the simplest semiconductor device, we cannot solve analytically the relevant equations governing the device physics. Taking ψ, n, and p as the potential, electron, and hole concentrations, respectively, we can specify [7,8]

$$d^2\psi/dx^2 = q/\epsilon_s[n(x) - p(x) - N_d(x) + N_a(x)], \tag{2.1}$$

$$J_n = \mu_n[kT(dn/dx) - qn(x)(d\psi/dx)], \tag{2.2a}$$

$$dJ_n/dx = -g(x, n, p, \psi), \tag{2.2b}$$

$$J_p = -\mu_p[kT(dp/dx) + qp(x)(d\psi/dx)], \tag{2.3a}$$

and

$$dJ_p/dx = +g(x, n, p, \psi). \tag{2.3b}$$

Equation (2.1) is Poisson's equation and contains the electronic charge q and the silicon dielectric permittivity ϵ_s. The continuity equations for electron current and hole current, respectively, are represented by Eqs. (2.2b) and (2.3b) with the net generation function $g(x, n, p, \psi)$ discussed previously. Equations (2.2a) and (2.3a) define the current density in terms of electron and hole concentrations and the electrostatic potential. The symbol μ represents charge carrier mobility. These equations are one dimensional to reflect the simple problem we have chosen.

There exists only one instance in which we can even approach an analytical solution to these strongly coupled and nonlinear equations. With zero applied bias, the diode will be in equilibrium so that the net generation function is zero everywhere. Thus, the current densities must be constant, and, because of the equilibrium restriction, are themselves zero. With this simplification, we may solve Eqs. (2.2b) and (2.3b) to get the electron and hole concentrations as exponential functions of the electrostatic potential. Plugging this result into Eq. (2.1) yields one nonlinear, second-order differential equation for the potential. If we further restrict the impurity density to be piecewise constant, we may solve for the derivative of the potential. But we still cannot derive the potential explicitly.

The structure of the solution is clear, though. The electron concentration is greatest and nearly uniform in the n-type region far removed from the junction. The electron density decreases monotonically as we approach and traverse the junction and eventually saturates at an exceedingly low value within the p-type region. The hole density behaves in a similar

manner in that it is greatest in the p-type region and decreases monotonically to a low value in the n-type region. In the spatial region close to and encompassing the junction, both the electron and hole concentrations are several orders of magnitude less than their respective maximum values. The charge density at any point near the junction, then, is completely dominated by the immobile ionized impurity charge. Far from the junction, the mobile charge carrier density and the ionized impurity density compensate to high precision so that the total charge density, and thus the Laplacian of the potential, vanishes.

These qualitative observations suggest the well-known depletion approximation in which we consider discrete blocks of silicon to be either electrically neutral (i.e., to possess zero total charge density) or fully depleted of mobile carriers. In this latter case, the charge density is the ionized impurity concentration. In the junction diode, a region of undetermined width encompassing the junction will be designated a depleted region while the remaining regions on both sides of the depleted region will be neutral. With this depletion approximation, we need only solve Poisson's equation for the potential under the assumption of $n = p = 0$ in the designated depletion region and match the solution of this depletion region appropriately with the solution in the neighboring neutral regions. It is customary to take the potential as constant within neutral regions even though, in principle, the depletion approximation will accommodate any potential function with a vanishing Laplacian. In the depleted regions, the potential varies quadratically with position if the problem is one dimensional and the impurity concentration is uniform. The depletion region width is determined uniquely by requiring a continuous potential and electric field across the boundaries between neutral and depleted regions. Figure 2.2 sketches the depletion approximation as applied to a junction diode. We

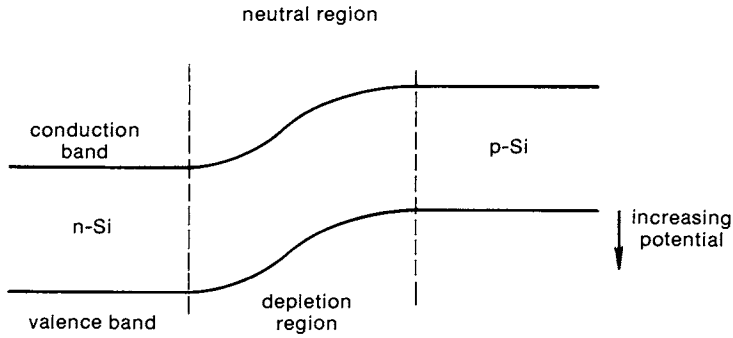


Fig. 2.2. We sketch the depletion approximation as applied to the one-dimensional junction diode with piecewise constant doping.

should note that the depletion approximation must be revised when either the *n*-region donor concentration or the *p*-region acceptor concentration is nonuniform [9,10].

The depletion approximation is useful since it allows us to derive (approximate) analytical results and properly reflects the physics of charge separation at a junction. Application of a reverse bias to the junction widens the depletion region since a negative bias on the *p* region pushes electrons within the *n* region and holes within the *p* region further from the junction. In fact, we note that one should expect a high net generation function within the depletion region of a reverse-biased junction diode since, with vanishing mobile charge carriers, there is precisely no recombination to balance the generation process. In silicon diodes near room temperature, the dominant contribution to reverse-bias current is thermal generation of charge within the depletion region.

Our description of the junction built-in potential noted the *n* region of a diode is positively charged while the *p* region is negatively charged. The depletion approximation certainly preserves this property since ionized donor impurities in the depletion region of the *n*-type terminal of the diode are positively charged while ionized *p*-type acceptor impurities are negatively charged. Since application of a reverse bias widens the depletion region, the positive charge of the *n* region and the negative charge of the *p* region increase. Thus, the diode acts as a capacitor since an equal and opposite voltage-dependent charge is stored on both diode terminals in the form of uncompensated ionized impurities. Defining the differential junction capacitance as the derivative of the magnitude of stored charge on each terminal with respect to the applied bias, an inspection of the depletion approximation shows that the capacitance decreases as the reverse bias increases and as the impurity concentration within either the *p* region or the *n* region decreases [1].

II. METAL–OXIDE–SEMICONDUCTOR CAPACITOR

The MOS capacitor consists of a conductive gate material separated from the semiconductor substrate by an insulating dielectric. Either a metal such as aluminum or molybdenum or heavily doped polycrystalline silicon generally serves as the conductive gate material. In silicon technology, the dielectric is either a thermally grown silicon dioxide film or an oxide film combined with a silicon nitride film. Charge storage in devices such as the DRAM and diverse image sensors is the dominant application of the MOS capacitor. But, one may also consider the capacitor as one

component of the MOS field-effect transistor and, for this reason alone, detailed understanding of the MOS capacitor is mandatory.

The MOS capacitor differs from its more conventional counterpart in that the semiconductor terminal is not of high conductivity. That is, when both terminals of a capacitor are conductive, then a potential difference applied to the two terminals yields an electric field within the dielectric, and this field is zero outside the dielectric. This "metal-dielectric-metal" capacitor stores charge at the interfaces of the respective metal electrodes with the dielectric.

The situation can change drastically with the semiconductor as the terminal of the MOS capacitor. If we specify the semiconductor as *p* type, then application of a negative bias to the gate with respect to the substrate will induce a positive charge on the substrate and a negative charge on the gate. The high conductivity of the gate requires that the electric field within the dielectric terminate at the dielectric-gate interface. The charge on which the field lines terminate is the charge stored on the gate. The semiconductor must supply positive charge to the substrate-dielectric interface. This task is readily accomplished because the mobile (majority carrier) holes are in great abundance and will collect at the interface in response to a negative bias on the gate. Thus, the negative bias case of the *p*-type silicon MOS capacitor is equivalent to that of the metal-dielectric-metal capacitor in that the insulator electric field is terminated on surface charges. This example, as well as that of a positive gate bias applied to an *n*-type silicon MOS capacitor, defines "accumulation" of the majority carriers at the semiconductor surface.

The MOS capacitor distinction appears with the opposite polarity of gate bias. A positive gate potential with the *p*-type MOS capacitor imparts a positive charge to the gate and a negative charge to the semiconductor substrate. In this case, there are very few mobile negative charge carriers (electrons) in the *p*-type substrate. Thus, the substrate cannot easily supply the negative charge. The positive gate bias repels the majority carrier holes from the substrate-dielectric interface and forms a depletion region stretching from the interface into the substrate. Evacuation of the mobile carriers (holes and electrons) from this surface region leaves this region negatively charged because of the immobile, ionized acceptor impurities. This negative "background" charge becomes the negative charge induced by the positive gate potential. But, the semiconductor terminal charge of the MOS capacitor is no longer confined to the surface. The charge extends to the depth of the depletion region, and this added distance increases the "effective" dielectric thickness and thus decreases the capacitance.

The low density of conduction electrons in *p*-type silicon arises from the relatively large energy difference separating the conduction electron energy

states and the Fermi level. (The Fermi energy may be roughly interpreted as the energy in a material above which no electrons will be found at the absolute zero of temperature.) The positive bias applied to the gate of a p-type MOS capacitor will, as noted, repel holes from the surface to form a surface depletion region. The positive bias also tends to lower the energy of conduction electron states relative to the Fermi level throughout the depletion region. Thus, electron density near the surface increases rapidly, exponentially at first, with increasing positive bias. At small positive bias, the electron concentration remains small compared to the acceptor impurity concentration so that the concept of a depletion region (i.e., region of essentially zero mobile charge density) is reasonable. As the positive gate bias increases beyond a critical value known as the *threshold voltage,* a narrow layer of high electron concentration forms at the silicon–dielectric interface. The electron concentration in this layer grows only linearly, as opposed to exponentially, above the threshold voltage since the energy level occupation probability ceases to behave exponentially with the energy separation from the Fermi level. This high electron density layer is called the "inversion" layer since the dominant charge carrier is opposite to that of the bulk of the semiconductor.

Charge storage in a semiconductor dynamic RAM cell or charge-coupled device manifests itself as inversion layer charge. Thus, we typically apply positive bias to p-type MOS capacitors and negative bias to n-type MOS capacitors. The threshold voltage, bias at which the silicon surface becomes inverted, is of great importance. Because of the initial exponential increase of electron concentration with gate bias, the inversion layer "appears" fairly abruptly with increasing gate bias. The accepted convention specifies that the electron concentration at the surface equals the ionized impurity concentration at the onset of inversion. Thus, the requisite surface potential for inversion relative to the substrate potential is twice the separation of the bulk (substrate) Fermi level from the midgap energy. See Fig. 2.3 for the band diagrams in accumulation, weak depletion, and inversion. The threshold voltage is then the gate bias required to produce this critical surface potential. We compute this quantity by solution of Laplace's equation in the oxide dielectric coupled with Poisson's equation in the semiconductor with the aid of the depletion approximation. The threshold voltage V_t is

$$V_t = 2\psi_b + \frac{t_{ox}}{\epsilon_{ox}} \sqrt{2\epsilon_s q N_a (2\psi_b)} + V_{FB}. \tag{2.4}$$

In Eq. (2.4), ϵ_{ox} and ϵ_s are the oxide and silicon permittivities, respectively. The oxide thickness t_{ox} and ionized impurity concentration N_a also appear. The bulk potential ψ_b is a weak, logarithmic function of the impurity concentration N_a.

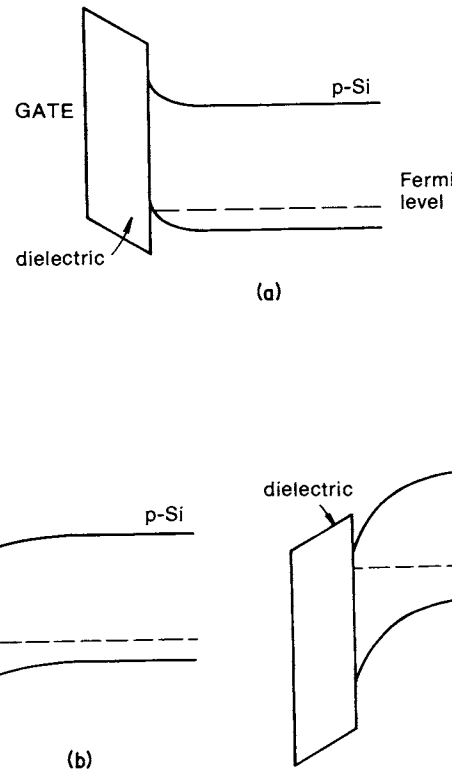


Fig. 2.3. We draw the band diagrams for the MOS capacitor in (a) accumulation, (b) weak depletion, and (c) inversion.

The "flatband" voltage, V_{FB}, also enters in the threshold voltage expression of Eq. (2.4). The flatband voltage is the gate bias one must apply in order to impart zero potential gradient to the semiconductor. One might reasonably expect this flatband voltage to be simply zero with respect to the substrate. That is, if we apply equal biases to the gate and substrate, one might expect zero potential gradient within the semiconductor (as well as the dielectric). There are two complications that mar this simplicity. First, there may exist charge (usually positive) within the gate dielectric. With nonzero charge in the dielectric, Laplace's equation is not valid, and one must apply a nonzero gate bias in order to leave a zero potential gradient in the semiconductor. The possible origins of dielectric charge include impurities, interface states, and broken silicon-oxygen bonds [11].

The remaining contributor to the flatband voltage is the gate-to-substrate work function difference. While the terminology is altered slightly, the physics of this work function difference is identical to that of the built-in potential of the junction diode discussed earlier in this chapter. If we

connect the gate of a MOS capacitor to the substrate with an external lead, then there will be a net flow of electrons and holes between the gate and substrate as long as electrons or holes in one material can "find" lower energy states in the other material. This net flow of current leaves equal and opposite charge on the two poles (gate and substrate). Thus, there will develop a potential barrier between the gate and substrate in the region of the external contact to oppose the net flow of charge. Note that this flow of charge did not occur through the dielectric. Yet this potential difference between the gate and substrate at the external lead demands, by Kirchhoff's law, a compensating potential drop as we circle the loop of gate–external lead–substrate–dielectric–gate. This compensating potential drop exists within the dielectric and, possibly in addition, at the semiconductor interface with the dielectric.

Thus, even when one "shorts" the gate to the substrate, there appears a potential difference within the gate–dielectric–substrate portion of the system. One can eliminate the potential drop in this portion by modifying the applied gate bias, and this modification must be reflected in the threshold voltage expression [Eq. (2.4)]. Note that this nonzero potential gradient would not exist from the gate to the substrate through the dielectric in the absence of an external contact of the gate to the substrate. This external contact allows the net charge flow that leaves the gate and substrate with nonzero net charge. Two "floating" masses with differing work functions separated by a dielectric of infinite resistivity would exhibit no potential gradient within the dielectric since these two masses would embody two isolated systems.

III. METAL–OXIDE–SEMICONDUCTOR FIELD-EFFECT TRANSISTOR

The MOSFET dominates MOS circuit designs. The primary goal of advanced process technologies is to reproducibly fabricate small-geometry, high-performance MOSFETs with excellent reliability. Defining and implementing advanced fabrication techniques require a thorough comprehension of the MOSFET impact. In this section, we discuss the underlying physics of MOSFET operation in the hope that we will shed light on the close coupling between MOSFET fabrication and performance.

Figure 2.4 sketches an n-channel MOSFET. Situated in a p-type silicon substrate are two separate surface $n+$ regions. In addition to these two junction diodes is a MOS capacitor. The side edges of the MOS capacitor gate coincide closely with one side of each $n+-p$ junction since the pres-

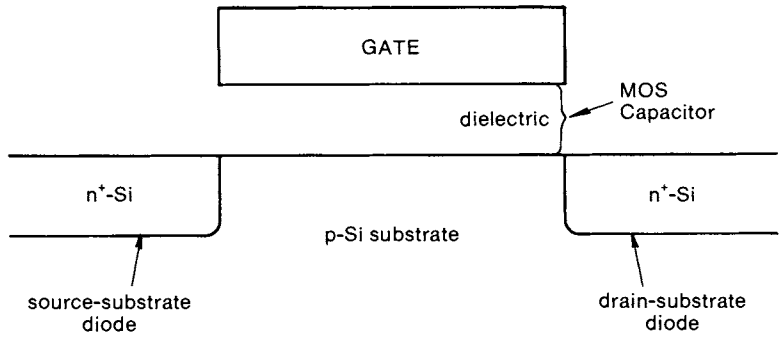


Fig. 2.4. We sketch a typical *n*-channel MOSFET in which the dimensions are not to scale.

ence of the gate precludes the ion implantation of donor impurities into the silicon substrate. The MOSFET appears as a conglomeration of two junction diodes and a MOS capacitor. The two $n+$ regions are known as the source and drain. The source, drain, and gate comprise the three important MOSFET terminals, while the potential of the *p*-type substrate is generally held at a fixed value. In typical operation, the source and substrate are at ground potential. The drain is raised to a positive bias such that the drain-substrate diode is reverse biased and the direct current (leakage) from drain to substrate is ideally small. Generally, the positive drain bias will not induce electrons to leave the $n+$ source since the electron energy barrier surrounding the source will, ideally, continue to prevent net electron flow. Application of a positive gate bias, however, will lower this energy barrier near the substrate-dielectric interface and permit electrons to flow from the source to the positively biased drain. The MOSFET therefore operates on the principle that current flow between source and drain is enabled by the administration of sufficiently large gate bias.

In this interpretation, the MOSFET consists of a reverse-biased diode (drain) and a forward-biased diode (source). A positive gate bias induces a positive potential within the channel of the MOSFET. This positive surface potential appears as a forward bias to the source. The source behaves as an electron emitter while the drain is an electron collector. This treatment is more precise than the popular view that the gate bias generates an inversion layer under the MOS capacitor and that this inversion layer serves as a one-dimensional resistor bridging the source and drain.

A qualitative description of MOSFET operation generally focuses on the magnitude of source-drain current flow as a function of gate, drain, and bulk potentials relative to the source potential. As noted above, one must impose a sufficiently high gate bias in order to enable current flow from source to drain. The critical bias at which current flow commences is the

familiar threshold voltage appearing in our discussion of the MOS capacitor. Application of this threshold voltage to the gate induces a critical surface potential within the silicon that permits excessive electron injection from the source. Current flow with gate bias below threshold, which would ideally be nonexistent, is generally small but increases exponentially with increasing gate bias since the forward bias characteristic of a diode exhibits exponential voltage dependence. This subthreshold regime, in which the gate bias is less than the threshold voltage, is of great interest and concern since excessive subthreshold current is one barrier to continued miniaturization of MOSFETs.

With the gate bias above threshold, the (source) drain current increases algebraically with increasing gate bias. Simple analytical models for MOSFET behavior prescribe either a linear or quadratic relationship, depending on the drain bias, for the drain current dependence on gate bias minus threshold voltage. Regardless of the gate bias, there will be no current flow if there is no potential difference between source and drain. With gate bias above threshold, the drain current increases linearly with drain bias (relative to the source) for small drain bias and eventually saturates at high drain bias. It is not uncommon to specify a negative bulk potential. That is, maintaining the source at ground, one might operate the MOSFET with a negative bulk potential so that both the source and drain form reverse-biased diodes to the substrate bulk. The drain bias is still positive with respect to the source (ground), so that electrons will flow, if enabled by the gate, from source to drain. However, the negative bulk potential increases the electron energy barrier surrounding the source so that the gate bias at which source electron injection commences is increased. That is, there is a larger barrier for the gate to reduce. The nonzero bulk potential increases the MOSFET threshold voltage.

Simple, one-dimensional, circuit oriented models suffice to explain these qualitative observations on MOSFET behavior. One expects that a complete, quantitative description of, for example, the dependence of drain current on all terminal voltages will emerge only from a solution of the governing equations. Such a solution is exceedingly challenging. As we discussed with the one-dimensional junction diode, the semiconductor device equations are strongly coupled and nonlinear. Additionally, Fig. 2.4 shows that the MOSFET is inherently two-dimensional. There is no reasonable one-dimensional simplification. These equations for the two-dimensional problem are [7,8]

$$\nabla^2\psi = q/\epsilon_s(n(x, y) - p(x, y) - N_d(x, y) + N_a(x, y)), \tag{2.5}$$

$$\mathbf{J}_n = \mu_n(kT\,\nabla n - qn(x, y)\,\nabla\psi), \tag{2.6a}$$

$$\nabla \cdot \mathbf{J}_n = -g(x, y, n, p, \psi), \tag{2.6b}$$

$$\mathbf{J}_p = -\mu_p(kT\,\nabla p + qp(x, y)\,\nabla\psi), \tag{2.7a}$$

$$\nabla \cdot \mathbf{J}_p = +g(x, y, n, p, \psi). \tag{2.7b}$$

The notation here is similar to that encountered previously with the diode. We now see gradient operators instead of simple derivatives and the current densities $\mathbf{J}_n$ and $\mathbf{J}_p$ are now vector quantities. The scalar mobilities μ_n and μ_p are, in the most general case, functions of position, potential gradient, and electron and hole concentrations. Equations (2.5)–(2.7) describe current flow within the silicon. In the dielectric sandwiched between the silicon substrate and the gate, the electrostatic potential satisfies Laplace's equation and the electron and hole concentrations are negligible.

Solution of these semiconductor device equations for the MOSFET would be of inestimable value. The primary goal of continuing advances in fabrication technology is the physical size reduction of the MOSFET. Lower cost motivates this size reduction with the added benefit of improved performance. Even when the processing technology is sufficient to produce smaller MOSFETs, extraneous problems related to the small physical size erode the improved performance and, much worse, introduce reliability hazards and undesirable MOSFET characteristics. The best example from this latter category is the excessive subthreshold current mentioned previously. Analytical solution of the MOSFET equations would undoubtedly allow the engineer to see clearly the dependence of MOSFET performance, reliability, and characteristics on the relevant fabrication choices, such as substrate impurity atom density, channel (gate) length, gate dielectric thickness, and applied bias.

Unfortunately, such an analytical solution does not exist because of the mathematical complexity. Sophisticated treatments of simpler problems may be found in Refs. 12–14. Numerical solution of the relevant equations (see, for example, Refs. 7, 8, and 15) is a challenging problem and has contributed greatly to MOSFET design and physical understanding of the issues surrounding channel length reduction. Still, numerical methods are limited in terms of providing insight. Instead of inspecting an analytical solution, one must know which cases to "run" in the computer modeling. Several reports have studied analytically the subthreshold conduction problem [16–18]. The mathematics simplify considerably in this regime and permit the investigation of the related short-channel problems.

An important guiding principle in the study of the MOSFET equations is known as scaling [19,20]. While one cannot solve the relevant equations, one may study how the solution should change when one reduces the channel length. Note that this is precisely the important situation. Presum-

ably, the engineer can manufacture reliable MOSFETs with good characteristics and performance with some particular physical dimension. The next step is to reduce this physical dimension (i.e., channel length). The scaling concept stems from the observation that if we decrease the gate oxide thickness, source and drain junction depths, and applied voltages (gate, drain, and bulk relative to source) in the same ratio by which we reduce the channel length and we simultaneously increase the donor and acceptor impurity concentrations by this ratio, then the MOSFET equations [Eqs. (2.5)–(2.7)] remain almost unchanged. Thus, in some sense, we may expect behavior from the short-channel MOSFET similar to what we observe in its longer channel predecessor. Note that these prescribed alterations leave the electric field (negative potential gradient) unchanged so that one often refers to "constant field" scaling.

Application of these scaling concepts immediately provides ideas on how an existing process should be modified to accommodate reduced design rules. But scaling is not exact. For example, while we may certainly decrease the voltages we apply (i.e., the boundary conditions) in the hope that the electrostatic potential throughout the device will decrease, we cannot directly control the threshold voltage. In Eq. (2.4), the voltage, ψ_b, does not decrease as physical dimensions decrease and neither does the work function difference component of the flatband voltage. The quantity ψ_b, which is relevant for the surface potential at inversion, work function difference from gate to substrate, and junction built-in potential, actually increases in the scaling process since we also specify increasing impurity atom concentration. The failure of these silicon bandgap-related quantities to change appropriately with the other parameters of channel length, gate oxide thickness, junction depth, impurity concentration, and applied voltages implies that the engineer may need to deviate from these scaling principles at some point. At this point, we must rely on numerical modeling and experimentation.

Another type of scaling difficulty involves parameters that cannot be modified to accommodate scaling even though these parameters appear accessible. For example, it is difficult to reduce the depth of the n-type regions forming the source and drain due to the nonlinear, fast diffusion of the donor impurities in the high-concentration regime. Also, we would like to increase the donor concentration in the source and drain regions by the scaling factor and this is essentially impossible since present MOSFETs already contain the maximum soluble, electrically active impurity concentration possible. The inability to increase this impurity concentration implies that series resistance within the source and drain degrade the performance of short-channel MOSFETs.

A further complication is that the circuit and system designers would

prefer not to reduce the applied voltages as the channel length decreases. Retaining constant voltages ensures greater current flow, which gives faster circuits, but more importantly allows the next generation of semiconductor products to "plug in" to existing systems. This latter compatibility concern is legitimate and has driven the industry along the path of "constant voltage," as opposed to constant field, scaling. Clearly, the constant voltage scaling increases the magnitude of electric fields within the MOSFET. Channel hot electron instability becomes increasingly troublesome as design rules shrink because of this increased electric field. This subject is addressed in the Chapter 5 discussion of reliability. Increased field strength also speeds gate oxide wearout as discussed in a subsequent chapter. Additionally, the expected improvement in MOSFET performance does not fully materialize since the charge carrier mobilities decrease with increasing field strength.

Departure from constant field scaling, whether unavoidable or intentional, complicates advanced process design and implementation in several other ways. For example, both the nonscaling nature of the junction built-in potential and reluctance to scale down the applied voltages mandate increasing the substrate impurity concentration by a factor greater than the scale factor. As one increases this substrate doping level, however, the drain junction avalanche breakdown voltage decreases and eventually sets an upper bound on the applied voltage. The greater impurity concentration moves the threshold voltage in the "wrong" direction, increases the deleterious junction capacitance, and further reduces the mobility of charge carriers due to impurity scattering.

Subthreshold conduction ultimately limits continued feature size reduction in MOSFETs. The (logarithmic) slope of the drain current versus gate bias subthreshold characteristic is determined by the physics of the forward-biased diode. This slope does not scale with the parameters discussed previously. Thus, to ensure that the drain current flowing with zero gate bias (i.e., subthreshold regime) is less than some maximum tolerable value, one must specify a minimum voltage by which the threshold must exceed zero gate bias. This minimum voltage spacing between zero and the threshold voltage does not scale and is in the range of 0.5–0.8 V depending on the application (i.e., on the maximum tolerable current flow at zero gate bias). Thus, the target threshold voltage of a new, reduced design rule process will not decrease with the other physical parameters relative to an existing, larger design rule process. Furthermore, noise margin considerations in circuit design dictate that the maximum applied voltage be at least four times the threshold voltage. Thus, the threshold voltage constraint becomes a power supply constraint that serves to frustrate attempts to abide by ideal (constant field) scaling.

As a final comment on MOSFETs, we note that the physics and operational principles of a *p*-channel MOSFET are analogous to those of the *n*-channel MOSFET. We chose in this chapter to discuss *n*-channel MOSFETs for the sake of clarity and consistency. *P*-channel MOSFETs, which possess $p+$ surface regions within an *n*-type silicon substrate, form an integral part of CMOS design, as will be discussed later in this chapter. Hole flow comprises the device current in *p*-channel MOSFETs. *P*-channel MOSFETs tend to switch slower than their *n*-channel counterparts because of the inherently lower mobility of holes relative to electrons. However, *p*-channel devices are far less susceptible to high field problems, such as hot carrier injection into the gate dielectric as well as mobility reduction. Other *n*-channel/*p*-channel differences and trade-offs arise from inherently different properties of acceptor impurities and donor impurities.

IV. CIRCUIT CONSIDERATIONS

Circuit, product and system considerations motivate the choice of CMOS technology over the most common alternatives of NMOS and bipolar. As discussed in detail in this monograph, continued CMOS miniaturization poses many challenges to the device and process engineers. For example, latch-up is specific to CMOS and will always plague this technology. Simultaneous fabrication of *p*- and *n*-channel MOS devices is expensive. Given this "downside" of CMOS, it becomes incumbent on the circuit designer and systems engineer to justify the CMOS selection.

As one expects, there are indeed compelling reasons for implementing systems in CMOS. Several advantages accrue, directly and indirectly, from the inherently lower power dissipation of CMOS compared to NMOS and bipolar technologies. Reduced power consumption stems from the absence of standby current in CMOS logic elements (i.e., inverters and AND gates). Beyond easing the power supply requirements, the reduction in power dissipation leads to lower operating temperature for the system. Device performance and reliability are thus favorably affected as is the task of cooling the individual chips and system.

To put it bluntly, CMOS permits implementation of some circuits that are not feasible in NMOS. For example, a 256-K SRAM possesses two inverters, one in each state, in every cell. The total power dissipation arising from inverter standby current with an NMOS design presents unacceptable heating of the silicon chip. It is simply not possible to increase power dissipation per chip arbitrarily by increasing the SRAM cell count. In CMOS, however, we shall document the inherently lower

standby current and power dissipation of the inverter that allows contemplation, design, and fabrication of megabit SRAMs.

In this section, we discuss the basis of decreased power dissipation and its ramifications. Indirect benefits, such as system complexity and device reliability, are explored. We outline the increased design and process complexity as CMOS disadvantages.

A. Power Dissipation

With the reader's indulgence, we begin with some elementary observations. Power dissipation in this context denotes the (temporal) rate of energy deposition into the active circuit elements from the system (i.e., power supply or source of energy). Most generally, Joule power dissipation is represented by the dot product of the electric field and current density vectors. The electric field does work (i.e., expends energy) on the moving charge carriers and the lost energy must be replenished by the external energy source. From the standpoint of the circuit and system, one simply considers the product of the applied bias and the drawn current in lieu of the electric field/current density dot product. This current–bias product, then, represents a transfer of energy from the source to dissipation within the circuit in the form of heat. Such a transfer is, of course, highly undesirable.

For given bias, reduction of supply current lowers the dissipated power. Consider the inverter structures of Figs. 2.5(a) (NMOS) and 2.5(b) (CMOS). The inverter serves as an excellent vehicle for comparison of power dissipation because of its ubiquity in logic design. In the NMOS realization of Fig. 2.5(a), the input is the gate of an NMOS FET while the power supply V_{DD} feeds through a load resistor to the FET drain. One generally fabricates the resistor as a (depletion-mode) FET with gate tied to one of the source–drain terminals to mimic a two-terminal resistor. The inverter output is the node between the FET drain and the resistor. When the gate (input) is high, the FET is biased above threshold so that the drain bias is low (close to the grounded source) under the reasonable assumption that the MOSFET channel resistance is significantly less than the load resistance. Thus, the output is low while the input is high. There is considerable current in this state since there exists a direct current path through two resistors (load and FET) between V_{DD} and ground. With the input low, however, the FET is off and the drain is essentially disconnected from the grounded source. Thus, there is no current path and the output must rise to V_{DD}. Hence, the low input yields a high output.

Contrast this situation with that of the CMOS inverter of Fig. 2.5(b). The

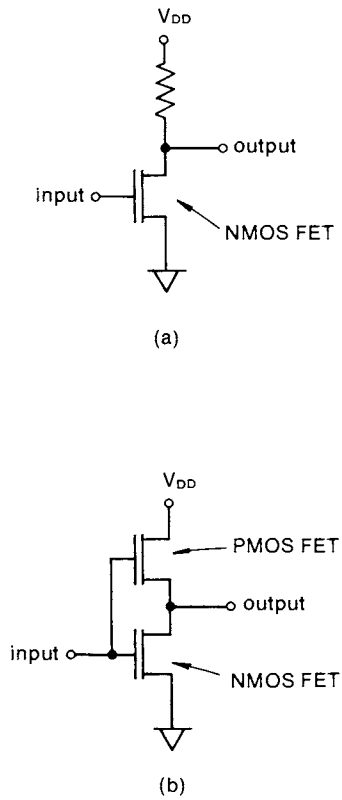


Fig. 2.5. We sketch the NMOS (a) and CMOS (b) inverter structures.

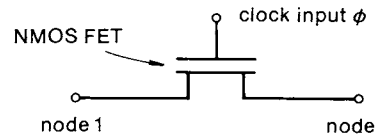
input now comprises the gates of an *n*-channel and a *p*-channel FET. The NMOS source is again grounded and the PMOS source is tied to V_{DD}. A low input level turns the PMOS device on while maintaining the NMOS device off. A high input level has the opposite effect of turning the NMOS FET on and leaving the other off. In the former case, the output is set by the PMOS FET as V_{DD} while in the latter case the output is ground. Hence, we have the inverter function. In neither (steady-state) case are both FETs on. There is no direct current path from V_{DD} to ground. In fact, the only appreciable current of a CMOS inverter arises during the switching transients so that there exists current for only a small fraction of the operation time.

The basis of reduced power dissipation in CMOS is thus revealed to be quite straightforward. Certainly, logic gates other than inverters do exist. But the comparison of NMOS to CMOS in terms of power dissipation is similar. Beyond logic gates, one finds that inverters and related elements

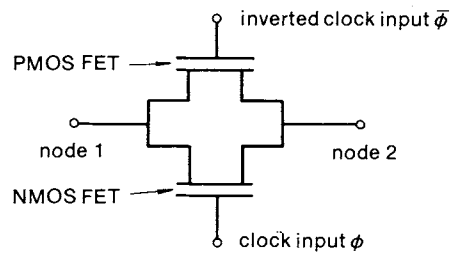
constitute many simple circuits. For example, a SRAM cell is essentially two inverters. Dynamic shift registers interpose inverters and transmission gates. CMOS simply represents a better idea for logic function implementation when steady-state current is undesirable.

Another advantage of CMOS is found in comparison of transmission gates in NMOS and CMOS technologies. The transmission gate acts as a switch between two nodes. When the switch, generally a MOSFET gate, is off, the two nodes are electrically isolated. With the switch turned on, the potential of the floating node rises or falls to the level of the other "controlled bias" node. Figures 2.6(a) and 2.6(b) show the NMOS and CMOS transmission gates, respectively. For concreteness, let us say that node 1 of Figs. 2.6(a) and (b) is the controlled bias node while node 2 is the floating node. The switch status is set by a clock input ϕ. Within a shift register, for example, a transmission gate permits propagation of an output signal of one stage to the input of the next stage when the clock level ϕ is high.

The NMOS gate of Fig. 2.6(a) is nothing but a single FET. The CMOS transmission gate realization is more involved and allows simultaneous charging of a node by NMOS and PMOS FETs in parallel. Figure 2.6(b) shows both the clock signal ϕ as well as its complement $\overline{\phi}$ tied to the gates



(a)



(b)

Fig. 2.6. We sketch the NMOS (a) and CMOS (b) transmission gates in which node 1 is the controlled bias node and node 2 is the floating node.

of the NMOS and PMOS FETs, respectively. Hence, one could place a neighboring inverter (to generate $\bar{\phi}$ from ϕ) in the sketch of the CMOS transmission gate. In practice, however, $\bar{\phi}$ is not created locally but is rather carried by bus lines throughout the chip as is ϕ. The NMOS transmission gate (i.e., FET) suffers from "threshold loss." That is, if the floating node is initially low and the controlled node initially high, turning the switch on will precipitate a MOSFET current to raise the potential of the floating node. When the floating node potential reaches the gate bias V_{DD} minus the FET threshold voltage, the MOSFET current declines abruptly and effectively ceases charging the floating node. Thus, the node is not charged to the full V_{DD} as desired. The CMOS transmission gate of Fig. 2.6(b) avoids this difficulty entirely. One of the two FETs is conducting with maximum gate drive in all cases so that no threshold loss will occur. The superiority of the CMOS implementation implies the need for fewer regenerative inverters in strings of transmission gates. This same property provides the rationale for faster charging of the CMOS transmission gate relative to the NMOS gate.

B. System Complexity

Having said very little thus far about the complexity of CMOS, we begin with system complexity at the package and board level. Heat dissipation, as discussed, is a primary factor at this level. The package design must succeed in maintaining a reasonable operating temperature. The placement of many integrated circuits on a single board in a computer, for example, relies on the ability to cool the board with ventilation systems. As one might expect, reduced power dissipation renders CMOS the most attractive alternative from this standpoint [21].

Furthermore, this advantage increases as device dimensions scale downward. The ideal, constant field scaling of Dennard *et al.* [19,20,22] implies a reduction in power dissipation per transistor equal to the square of the scaling factor κ. One obtains a factor of κ from both the power supply reduction and loss in transistor drive current. Since one expects to fabricate a factor of κ^2 more transistors within the same chip area, the total amount of power dissipated should remain constant. As discussed in the previous section on device considerations, however, designers are reluctant to remain faithful to constant field scaling. At the other extreme of constant voltage scaling, there is no reduction in the supply voltage. Furthermore, one may expect almost a factor of κ^2 increase in the drive current because of reduced transistor gate length and oxide thickness. The additional κ^2 factor due to the increased number of devices per area yields a horrendous

κ^4 increase in power dissipation upon dimensional scaling by a factor κ. Scaling decisions in industrial environments generally compromise between these two extremes. But, the lesson clearly shows that power dissipation is exacerbated as technology advances.

Common to all technology choices is the increased number of input/output pins per area as dimensions shrink. This issue in itself presents a challenge to the package engineering community. In this context, concurrent and drastic increases in the requirements of heat dissipation of NMOS and bipolar technologies are undesirable. There are situations, of course, where a special advantage of an alternative technology outweighs the heat dissipation problem. For example, high-speed computation is accomplished far more readily with (bipolar) emitter-coupled logic. This choice exacts a price in system complexity due to heat dissipation.

C. Reliability

Device and circuit reliability are of great importance in all semiconductor technologies. We devote an entire chapter to this subject later in this monograph and focus on those issues (hot carriers, oxide wear-out, and electromigration) that are most relevant to CMOS. It is worth noting here, however, the inherent differences in this topic in the presence of CMOS. First, the downside of CMOS is latch-up. Though this is sometimes viewed as a reliability concern, we preferred to incorporate an investigation of latch-up with the processing discussion of the elimination of this problem: adequate device isolation. Such a consideration is absent in NMOS and bipolar technologies.

As one suspects by now, a dominant theme in CMOS is the reduction of power dissipation due to the low standby current. Not only does one enjoy reduced requirements on heat removal, but the operating temperature of CMOS is generally quite close to room temperature. (Specifications for operating temperatures for commercial devices are typically $-40°-85°C$ [23].) As a consequence of the increased heat dissipation, NMOS and bipolar will tend to find a steady-state temperature at a higher value than that of CMOS. Two important reliability mechanisms (oxide wear-out and electromigration) become more troublesome at higher temperatures (see Chapter 5). Thus, a reasonable expectation of an operating temperature difference of 20°C between CMOS and other alternatives should result in a minor, though probably not negligible, reliability advantage for CMOS. Of course, we would note that high temperature actually decreases the hot carrier reliability problem and one could thus turn the argument around to suggest that CMOS has a relative disadvantage on this point. We would

counter that this gain in hot carrier reliability arises because of reduced channel mobility. That is, the improved hot carrier performance is a direct consequence of reduced carrier mobility. Thus, the small, yet perhaps nonnegligible, gain in hot carrier stability comes at the expense of an approximately equivalent loss of device and circuit performance.

The origin of reduced power dissipation, we recall, is reduced standby current in logic elements such as the inverter. This current reduction by itself is beneficial. Electromigration, for example, is quite simply macroscopic deformation and destruction of current conductors caused by electron flow. Great reduction in current, then, will reduce electromigration. A good estimate of this effect would be provided by simple comparison of on times for current. For example, the NMOS inverter will have full current about half the operating time if we assume that the inverter spends equal time in both logic states. The CMOS inverter, by comparison, allows current only during switching transients. Conservatively estimating an effective 5% on time for these transients, one might reasonably expect to gain a factor of ten increase in the lifetime because of electromigration for these logic elements. The effect on hot carrier degradation, though less drastic, is also present. While this mechanism also relies on the presence of current flow, one requires a high electric field, too. Such a situation exists in the switching transients of both NMOS and CMOS technologies. But, the large standby current of NMOS logic elements is not generally accompanied by the high field as well so that the comparison here is much smaller than that of electromigration.

D. Circuit Complexity and Considerations

One cannot deny the inherent complexity of CMOS device structures. Again, the simplest example is the inverter logic element. Both CMOS and NMOS implementations require two MOS transistors. In the NMOS case, of course, both transistors are *n* channel. Even when the load transistor requires a depletion implant, it is far more challenging to build the two FETs of the CMOS design since the *p*- and *n*-channel devices require individual optimization and mutual isolation. The bulk of this treatise is devoted to CMOS fabrication issues due to this fundamental concern of CMOS. The presence of both device polarities demands additional chip area to avoid deleterious latch-up and punch-through. CMOS is now the preferred technology in many applications simply because the advantages have outstripped this circuit complexity issue.

There do exist CMOS advantages in circuit construction, though. The peripheral circuitry in DRAMs is implemented far more efficiently in CMOS than in NMOS [24,25]. Having paid the process complexity price,

the presence of p and n channels gives added flexibility. In the case of the DRAM, the clocks and shift registers may be CMOS while the memory cell is purely n channel if this is deemed optimum. Others have argued that p-channel memory cells with CMOS peripheral circuitry are optimum because of enhanced reliability [25]. While this statement may be controversial because of other considerations, the point is that CMOS provides the flexibility for such an argument to be meaningful. One might certainly imagine applications in which the p-channel technology could be exploited. Though often maligned for lower carrier mobility and less precise source–drain junctions, p-channel devices are inherently more impervious to ionizing radiation [26], are able to withstand higher operating voltages [27], and exhibit lower $1/f$ noise [28] as compared to NMOS.

E. Summary

In terms of circuit design and system issues, CMOS provides a much different world than alternative technologies. The primary advantage of CMOS is the reduced power dissipation. Other advantages (heat dissipation, system complexity, and reliability) follow from this straightforward idea. Circuit and processing complexity pose the dominant drawbacks of CMOS. We tend to throw the latch-up issue in with the process complexity since the process must be designed and implemented to prevent latch-up under all operating modes. There do exist, and likely always will exist, specific circumstances for which CMOS is not the optimum technology choice. But CMOS does possess a large share of the present market. One expects this share to grow unabated because of increasing power dissipation with decreasing design size.

V. CMOS LATCH-UP

We have made many allusions to latch-up in CMOS. Even though Chapter 4 offers a concise description of this process, it behooves us to devote specific attention to this phenomenon here. To begin, we consider the thyristor [1,29] since parasitic thyristors in the CMOS structure play a leading role in latch-up. Just as one considers a diode as a $p-n$ junction, a bipolar transistor as a $p-n-p$ (or, alternatively, $n-p-n$) structure, the thyristor consists of a $p-n-p-n$ series of alternating p-type and n-type regions. The outer p and n regions are the anode and cathode, respectively. The inner p and n regions may be left floating (in which case the device is designated a Shockley diode) or controlled by independent leads. With one

of these inner regions contacted and one floating, the resulting device is a thyristor or semiconductor-controlled rectifier (SCR).

While description of the thyristor structure presents no conceptual difficulty, such is not the case with the electrical properties. Just as one cannot infer transistor action in a bipolar transistor by consideration of independent diodes in the transistor structure, one is also unable to understand thyristor operation by analogy with simpler devices. We restrict ourselves to the dependence of thyristor operation on the anode–cathode bias. That is, let us neglect for the moment current injection into one of the inner regions. A negative bias on the anode relative to the cathode will cause two of the three junctions in the $p-n-p-n$ structure to be reverse biased. The remaining center junction will acquire a small forward bias (just enough to supply the small leakage current of the thyristor). As the magnitude of this negative anode bias increases, the thyristor will conduct very little current since the anode and cathode are effectively (electrically) isolated by two reverse-biased junctions. The only point at which this high impedance state disappears is at the onset of avalanche breakdown in a reverse-biased junction or with punch-through of one of the inner (n or p) regions. Thus, one may construe negative anode bias as the thyristor reverse mode.

Consider now biasing the anode positive with respect to the cathode. The two junctions that were reverse biased in the negative anode case are now at a very small forward bias to accommodate a small leakage current. The middle junction is now reverse biased. We continue to claim, then, that the anode and cathode are electrically isolated in this forward mode (positive anode bias) as they were for the reverse mode. Only one reverse-biased junction serves for this isolation in the forward mode, though. But, this is not the whole story. If we measure thyristor current as a function of anode bias (with cathode grounded) by cautiously increasing this bias from ground, we will indeed find a very low current. This state is known as forward blocking. This condition has direct relevance for CMOS since we want all parasitic thyristors in the forward blocking or reverse modes.

Unfortunately, mathematics and physics conspire to destroy this state of affairs! Consider the mathematical formulation of the relevant semiconductor physics embodied in Eqs. (2.5)–(2.7). In principle, one may solve these equations with the appropriate geometry and boundary conditions of the thyristor. If one could indeed find an analytical solution, the result would show a finite range of positive anode bias in which *three* solutions exist simultaneously. This nonuniqueness of the solution is entirely consistent with the field of nonlinear differential equations. The upshot is that the forward blocking, low current behavior is only one of three possible states of the thyristor. We may discard one of the other two possible states since it is unstable. That is, the unstable state *does* represent a valid

solution of Eqs. (2.5)–(2.7). But, perturbations in any aspect of the system (temperature or bias, for example) precipitate a change in this solution. The remaining, stable solution dictates an exceedingly large anode–cathode current.

Deliberate exploitation of the existence of two solutions/states is the motivation for thyristor fabrication. The thyristor is essentially a switch that can handle high currents and finds application in industrial environments. In practice, one places a resistive load in series with the thyristor to place a current limit in the high current state. Switching to the high current mode is achieved in diverse methods such as injection of current into one of the inner (p or n) regions, voltage transients, or ionizing radiation. Subsequent demotion to the low current mode generally requires removal of the thyristor supply voltage. We recommend the reader consult Sze [1] and Ghandhi [29] for a more complete discussion of thyristor design and applications.

If a parasitic thyristor in the CMOS structure settles into its high current state, in which case we say it "latches," irreparable damage will follow. For obvious performance reasons, the circuit designer does not add series impedance to limit the latched current. This current is then orders of magnitude greater than that which the conductive lines were designed to carry. Latching is, therefore, unacceptable. The remainder of our discussion will focus on the forward blocking mode of the thyristor and prevention of switching to the latched state.

In our preceding discussion of the $p-n-p-n$ structure in the forward blocking mode (anode positive with respect to cathode), we noted that the middle $p-n$ junction is reverse biased while the outer two junctions exhibit a small forward bias. Since a forward-biased (emitter-base) and reverse-biased (collector-base) junction characterizes an active bipolar transistor, we may mentally decompose the forward blocking thyristor into two active bipolar transistors. See Figs. 2.7(a) and (b) in which we sketch this conceptual two-transistor model and its relationship to an n-well CMOS structure [30]. The forward-biased junctions of the thyristor correspond, naturally, to the emitter-base junctions of the conceptual bipolar transistors (one $n-p-n$ and one $p-n-p$) while the reverse-biased thyristor junction serves as the collector-base junction of both imagined transistors. These two transistors find themselves facing in opposite directions. The collector of each transistor is also the base of the other transistor.

Let us state clearly that derivation of thyristor characteristics requires solution of the relevant semiconductor equations in the specific geometry with proper boundary conditions. Furthermore, identification of multiple solutions and stability classifications will not follow from simple models. (We do not forget that experimentalists discovered all this first!) But, the

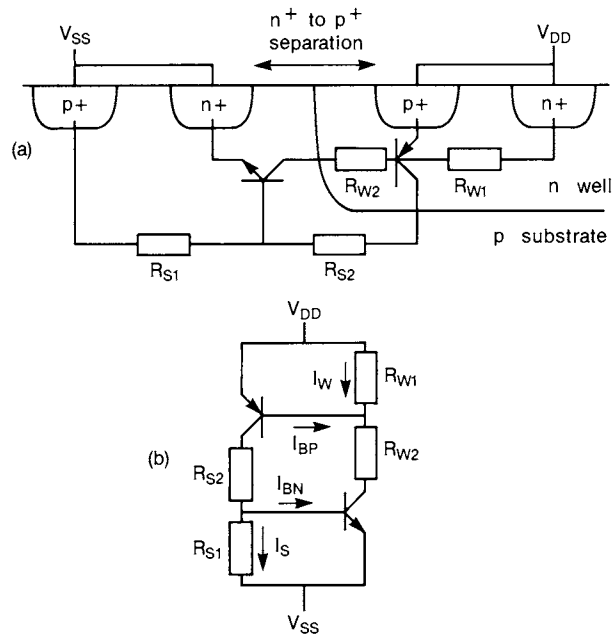


Fig. 2.7. We sketch a typical *n*-well CMOS structure (a) and a conceptual equivalent circuit diagram (b) for discussing latch-up. (From Lewis *et al.* [30].) ©1987 IEEE.

two-transistor model we have posed can convince us that thyristors do have unusual properties. If we inquire about the current flow in the $p-n-p$ transistor of our thyristor, it is natural to search for a base current (of electrons) into the n region in order to multiply by the appropriate current gain factor. With no additional contact to this region, this base current is equivalent to the collector current of the (other) $n-p-n$ transistor. Thus, the collector current of the $p-n-p$ transistor is the $n-p-n$ collector current multiplied by a ($p-n-p$) current gain. Conversely, the as yet unknown $n-p-n$ collector current arises from a hole base current, this time the $p-n-p$ collector current, multiplied by the $n-p-n$ transistor gain.

For details of this simple model we again refer the reader to Refs. 1 and 29. It is clear that there is a positive feedback in that an increase in one current component precipitates an increase in the other transistor for which the original current is the base current. Then the second current acts to increase the original current in the same manner. When the overall gain of this process is large enough, any "seed" current will be amplified to an arbitrarily large degree. This seed current is most often just the ubiquitous thermal leakage (i.e., "dark current"). The positive feedback process reminds one of avalanche multiplication in a reverse-biased diode [3]. The

conventional definition of "large enough" gain is that the sum of the $p-n-p$ and $n-p-n$ common emitter gains is unity. Increasing the anode-cathode bias increases both of these transistor gains by improving emitter efficiency (a current-dependent quantity) and reducing the effective base width. Thus, this simple model does predict an anode bias at which the thyristor current increases dramatically. Further investigation of thyristor turn-on with this model is difficult. We would note that the high current state of the thyristor forward biases all three junctions and drops the anode-cathode bias with the IR product at the high current level. Numerical solution of the semiconductor equations yields an important tool for studying latch-up since there is clearly no rigorously valid decomposition of the thyristor into simpler structures. One study demonstrated how a transient voltage pulse can induce latch-up [31]. Figure 2.8 plots the anode current of the parasitic thyristor of a CMOS structure for various temporal durations of a voltage pulse. The structure latches if the pulse width exceeds a critical value.

Since our goal is to avoid latch-up, one infers the need to minimize the current gains of these two conceptual bipolar transistors. Perhaps the simplest method to accomplish this task is to lengthen the two inner (base) regions of the thyristor by spacing the $p+$ (anode) and $n+$ (cathode) regions far apart. Large base widths give commensurately lower transistor gain. But, the penalty of this approach, which will always work, is reduced packing density of the devices in the CMOS circuit design. One therefore seeks other methods for the suppression of latch-up to implement in conjunction with a design rule for device spacing.

Discussion of these other solutions must account for the multidimensional nature of the true latch-up problem. One never finds our one-di-

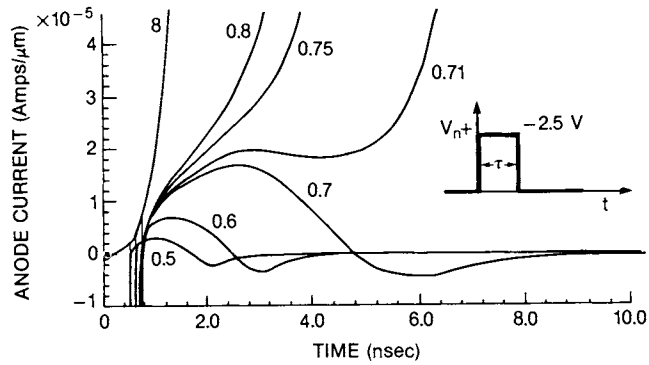


Fig. 2.8. We plot the simulated anode current of a parasitic CMOS thyristor subjected to transient voltage pulses of varying duration. (From Pinto and Dutton [31].) ©1985 IEEE.

mensional thyristor example in true CMOS implementations. Even the two-dimensional sketch of Fig. 2.7(a) does not recognize that the parasitic thyristors of CMOS involve devices with small widths and lengths so that the structures are three-dimensional. One may attempt to prohibit the establishment of the high current state in a three-dimensional parasitic thyristor by denying the lateral and vertical voltage drops necessary to sustain high current. For example, shorting neighboring n and p regions together would ideally prevent strong forward biasing of this junction. This technique is popular and useful. It does not eliminate latch-up since this junction will still develop a forward bias at a point removed from the surface shorting bar. Quite successful is the placement of a heavily doped substrate close to the surface device region (with a thin, lightly doped epitaxial layer). This structure, though expensive, greatly reduces lateral voltage drops necessary to sustain latch-up. Finally, "opening" the thyristor by interposing an insulator into the conduction path will, of course, completely prevent latch-up. Substantial but incomplete obstruction of the current path will also meet with success. We discuss these techniques in more detail in our discussion of isolation for CMOS in Chapter 4.

Let us be more specific about latch-up in CMOS [32,33]. Consider a cross section of a CMOS inverter fabricated by an n-well process in Fig. 2.9. To obtain worthwhile information from simulations and experiments, it is important to consider a truly representative CMOS structure. The prevalence of the inverter throughout digital design suggests consideration of latch-up in the inverter. The two different states of the inverter of Fig. 2.9 are susceptible to latch-up by different parasitic thyristors. With the input high and output low, for example, the $p+$ source (i.e., $p+$ region tied to V_{dd}) will act as the parasitic thyristor anode while the NMOS ($n+$) drain becomes the cathode. These assignments change when the input is low and the output high. One state will likely be more immune to latch-up than the

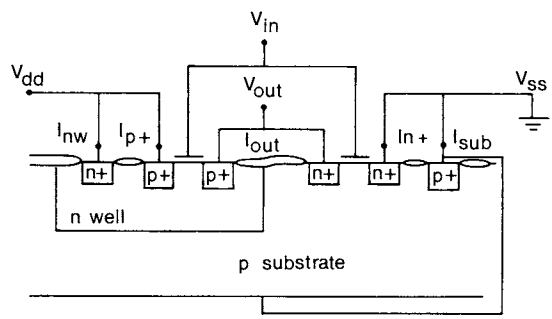


Fig. 2.9. We sketch a cross section of a CMOS inverter for discussion of latch-up. (From Lee [33].) ©1987 IEEE.

other. Furthermore, Lee has argued that the existence of "extraneous" nodes in a real CMOS structure, such as the inverter, will yield different latching behavior [33]. We expect that a combined study of soft error rates (i.e., single event upset) and latch-up in SRAM structures due to ionizing radiation and alpha particle bombardment would be both technically relevant and challenging.

Latch-up susceptibility is a direct and undesirable consequence of the CMOS technology. The price one pays for latch-up suppression is increased fabrication process complexity as discussed in Chapter 4. But CMOS is well worth the price. As we emphasized, CMOS permits levels of integration on a single chip that are impractical with NMOS and bipolar technologies because of cooling requirements. CMOS meets the required low power dissipation for continued technology advances in system integration and device miniaturization.

REFERENCES

1. S. M. Sze, "Physics of Semiconductor Devices," Wiley, New York, 1981.
2. A. S. Grove, "Physics and Technology of Semiconductor Devices," Wiley, New York, 1967.
3. S. K. Ghandhi, "The Theory and Practice of Microelectronics," Wiley, New York, 1968.
4. E. H. Nicollian and J. R. Brews, "MOS Physics and Technology," Wiley, New York, 1982.
5. R. N. Hall, Electron-hole recombination in germanium, *Phys. Rev.* **87**, 387, 1952.
6. W. Shockley and W. T. Read, Statistics of the recombination of holes and electrons, *Phys. Rev.* **87**(5), 835, 1952.
7. S. Selberherr, "Analysis and Simulation of Semiconductor Devices," Springer-Verlag, Berlin and New York, 1984.
8. For example, E. M. Buturla and P. E. Cottrell, Simulation of semiconductor transport using coupled and decoupled solution techniques, *Solid State Electron.* **23**, 331, 1980.
9. D. Redfield, Revised model of asymmetric p-n junctions, *Appl. Phys. Lett.* **35**, 182, 1979.
10. J. M. Pimbley, Depletion approximation for an exponentially graded semiconductor P-N junction, to be published in *IEEE Trans. Elec. Dev.*, 1988.
11. A. S. Grove, A. H. Snow, B. E. Deal, and C. T. Sah, Investigation of thermally oxidized silicon surfaces using metal-oxide-semiconductor structures, *Solid State Electron.* **8**, 145, 1965; B. E. Deal, Standardized terminology for oxide charges associated with thermally oxidized silicon, *IEEE Trans. Elec. Dev.* **ED-27**, 606, 1980.
12. C. P. Please, An analysis of semiconductor P-N junctions, *IMA J. Appl. Math.* **28**, 301, 1982.
13. P. A. Markowich, "The Stationary Semiconductor Device Equations," Springer-Verlag, Berlin and New York, 1986.
14. M. S. Mock, "Analysis of Mathematical Models of Semiconductor Devices," Boole, Dublin, 1983.

15. K. Board and D. R. J. Owen, eds. "Simulation of Semiconductor Devices and Processes," Pineridge, Swansea, Wales, 1984.
16. K. N. Ratnakumar and J. D. Meindl, Short-channel MOST threshold voltage model, *IEEE J. Sol. St. Circ.* **SC-17**(5), 937, 1982.
17. J. R. Pfiester, J. D. Shott, and J. D. Meindl, Performance limits of CMOS ULSI, *IEEE Trans. Elec. Dev.* **ED-32**(2), 333, 1985.
18. J. M. Pimbley and J. D. Meindl, MOSFET scaling limits, to be published in *IEEE Trans. Elec. Dev.*, 1988.
19. R. H. Dennard, F. H. Gaensslen, H. Yu, V. L. Rideout, E. Bassous, and A. R. LeBlanc, Design of ion implanted MOSFETs with very small physical dimensions, *IEEE J. Sol. St. Circ.* **SC-9,** 256, 1974.
20. G. Baccarani, M. R. Wordeman, and R. H. Dennard, Generalized scaling theory and its application to a $\frac{1}{4}$ micron MOSFET design, *IEEE Trans. Elec. Dev.* **ED-31**(4), 452, 1984.
21. W. J. Niewierski, 8- and 16-bit processors round out high-level CMOS architecture options, *Electronics,* 116, April 1984.
22. C. Hilsum, ed., "Handbook on Semiconductors," Vol. 4, Device Physics, North-Holland, Amsterdam, 1981.
23. R. E. Funk, Fast CMOS logic bids for TTL sockets in most systems, *Electronics,* 134, April 1984.
24. J. Fiebiger, CMOS—a designer's dream with the best yet to come, *Electronics,* 113, April 1984.
25. K. Yu, R. Chwang, and A. Mohsen, Dynamic RAM fabrication poised for continuing C-H-MOS invasion, *Electronics,* 121, April 1984.
26. K. G. Aubuchon, Radiation hardening of P-MOS devices by optimization of the thermal SiO_2 gate insulator, *IEEE Trans. Nucl. Sci.* **NS-18**(6), 117, 1971.
27. E. Takeda, Y. Nakagome, H. Kume, N. Suzuki, and S. Asai, Comparison of characteristics of *n*-channel and *p*-channel MOSFETs for VLSI, *IEEE Trans. Elec. Dev.* **ED-30,** 675, 1983.
28. K. H. Duh and A. van der Ziel, Hooge parameters for various FET structures, *IEEE Trans. Elec. Dev.* **ED-32,** 662, 1985.
29. S. K. Ghandhi, "Semiconductor Power Devices," Wiley, New York, 1977.
30. A. G. Lewis, R. A. Martin, T.-Y. Huang, J. Y. Chen, and M. Koyanagi, Latchup performance of retrograde and conventional n-well CMOS technologies, *IEEE Trans. Elec. Dev.* **ED-34,** 2156, 1987.
31. M. R. Pinto and R. W. Dutton, Accurate trigger condition analysis for CMOS latchup, *IEEE Elec. Dev. Lett.* **EDL-6,** 100, 1985.
32. See, for example, F.-S. J. Lai, L. K. Wang, Y. Taur, J. Y.-C. Sun, K. E. Petrillo, S. K. Chicotka, E. J. Petrillo, M. R. Polcari, T. J. Bucelot, and D. S. Zicherman, A highly latchup-immune 1μm CMOS technology fabricated with 1 MeV ion implantation and self-aligned $TiSi_2$, *IEEE Trans. Elec. Dev.* **ED-33,** 1308, 1986.
33. C.-T. Lee, Pseudocollector effect in a CMOS inverter, *IEEE Trans. Elec. Dev.* **ED-34,** 2212, 1987.

Chapter 3

Metallization

This chapter[1] will cover several metallization topics: gate electrodes, contacts, multilevel interconnections, and capacitors for analog function. In addition, this chapter will discuss several advanced interconnection and metallization concepts that will be useful for shrinking device and CMOS circuit size.

I. GATE ELECTRODES

Early MOS circuits used aluminum electrodes. The old PMOS aluminum gate technology produced high negative threshold PMOS FETs. These gates were not self-aligned with the source and drain diodes, and the design rules for them were quite coarse (5–10 μm). The advent of ion implantation allowed the device designer to tailor the thresholds and even produce depletion mode load devices. The introduction of refractory materials (e.g., Si, W, Mo) for gates in 1968 allowed for self-alignment [1,2], and hence, smaller dimensions since the gates could be used as diffusion or ion implantation masks. The work function of $n+$ polysilicon is about the same, as Al (4.1 V),[2] which produces an accumulated surface on n-type

[1] Portions of this chapter are reprinted with permission from *Proc. IEEE* **74**(12), 1986, pp. 1678–1702.

[2] The absolute values for the work functions of gate electrode materials vary somewhat in the literature. For instance, the electron affinity of Si as determined from subtracting the band gap from the photoelectric threshold is 4.1 ± 0.1 eV [3]. The work function of Al is about 4.1 eV. The work function of Mo as determined from flatband voltage measurements as a function of oxide thickness is 4.69 ± 0.03 eV [4]. The work functions of refractory metal silicides have been summarized [5]; however, there is variance in the literature. Reference 6 gives the work function of $n+$ polysilicon as 4.3 eV and that of $MoSi_2$ as 4.8 eV [6].

silicon, resulting in larger than desired negative PMOS thresholds and a depleted surface on p-type silicon, which produces low NMOS thresholds. The following contains a discussion of this problem.

In the absence of oxide space charge and fast interface state traps, the MOSFET threshold equation is

$$V_T = (\epsilon_G - 2\phi_B)/q - Q_s/C_o + \phi_{MS}, \tag{3.1}$$

where ϵ_G and ϕ_B are, respectively, the band gap and bulk Fermi level measured from the valence band edge in electron volts and Q_s, C_o, and ϕ_{MS} are, respectively, the semiconductor's space charge produced by the ionized impurities in the surface depletion region at the onset of minority carrier inversion, the gate oxide capacitance, and the metal to Si work function difference [7].

For purposes of illustration, consider a p-channel device in an n well. Typically, the p-channel device using an $n+$ polysilicon gate will have a threshold that is too negative, and a boron implant is required to increase it. This counterdoping produces a buried channel type device that has poor turnoff or subthreshold characteristics. A larger work function allows the device designer to reduce the boron implant dose, thus improving the subthreshold characteristics. Conversely, the threshold of an $n+$ polysilicon gate NMOS device is too low because a moderately doped p-type silicon surface is depleted by the low work function of the $n+$ polysilicon gate. The acceptor doping concentration must, therefore, be increased to raise the threshold.

Although in the past the work function has been compensated for by these channel implants, this becomes an even bigger problem for future devices using thinner gate oxides. This is because the channel space charge is multiplied by the gate oxide thickness coming from the C_o term in the threshold voltage equation and so even larger doping is required to offset with corresponding space charge an "improper" work function to obtain the desired threshold voltage. Since these types of problems are common to either the n- or p-channel device (or both) in downscaled CMOS circuits, it is advantageous to work with a gate electrode material with a larger work function.

The larger work functions of Mo (4.7 V), W, or refractory silicides produce low and nearly symmetrical thresholds for p- and n-channel devices on moderately doped substrates [4]. Figure 3.1 shows the reason for this. Figure 3.1 gives the surface potential and band bending diagrams for $n+$ polysilicon and Mo electrodes. To generate these diagrams, one must compute the surface electric field as a function of surface potential for a given substrate impurity concentration with the constraint that the potential drop in the silicon (band bending) plus that in the gate dielectric

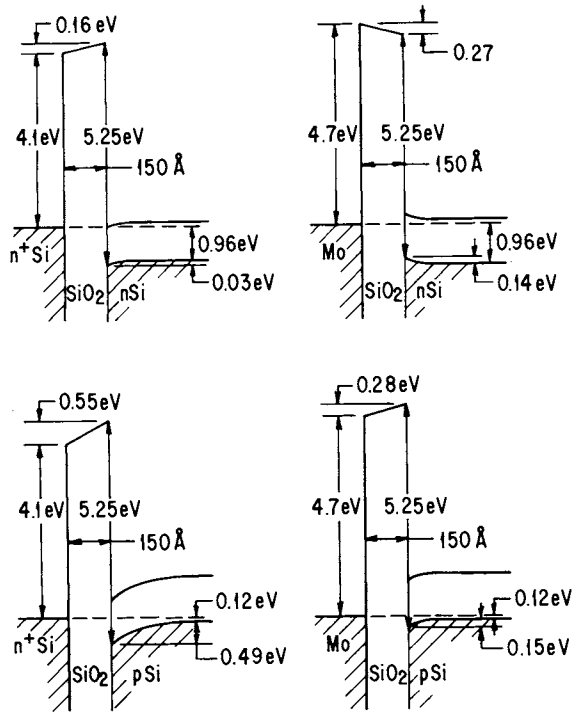


Fig. 3.1. Surface potential diagrams for $n+$ polysilicon and Mo gate devices on n- and p-type substrates with $10^{17}/cm^3$ impurities. The gate oxide is 150 Å thick.

balances the work function difference. Figure 3.1 shows a surface accumulation layer surface potential for the case of an $n+$ polysilicon electrode and an n-type silicon substrate. This means that the PMOS threshold implantation dose is required to reduce its negativeness. The surface of the p-type silicon is strongly depleted producing an NMOS threshold that is too low, and a boron implantation is required to increase it. This increase in doping increases the efficiency of hot electron injection into the gate oxide, which in turn degrades the hot electron reliability aspects of the device. Figure 3.1 also shows the surface potential diagrams for Mo or W gate NMOS and PMOS devices. These diagrams show that both of these substrate surfaces are slightly depleted by the same amount, which produces equal and opposite NMOS and PMOS thresholds without any additional doping steps. The diagrams are striking in that they graphically depict the reason for the strongly asymmetrical n- and p-type thresholds observed for $n+$ polysilicon gate devices and the more nearly symmetrical thresholds of refractory metal and refractory metal silicide gate devices. Obviously, the $n+$ polysilicon gate devices require more surface doping to produce sym-

metrical n- and p-channel enhancement mode devices since additional doping is needed to compensate for the asymmetry.

The possible advantage of using $p+$ polysilicon versus $n+$ polysilicon for gate electrodes is also worth discussing. In this case, the band bending diagrams are interchanged for p- and n-type substrates, which means that the PMOS subthreshold leakage problem is reduced and then the NMOS device becomes a "buried channel" type. This reduces the NMOS hot electron problem at the expense of subthreshold leakage. Some workers have even suggested that it would therefore be advantageous to make CMOS circuits where the electrodes of NMOS and PMOS devices are, respectively, $n+$ and $p+$ polysilicon to correct for the work function disparity [8]. This, of course, greatly complicates the normal CMOS process. In order to simplify the process, the choice of $p+$ polysilicon electrodes rather than $n+$ polysilicon would appear to be the best when viewed from a device physics point of view. However, the industry has a reluctance to do this because of another danger. Boron can diffuse rapidly from the $p+$ polysilicon into and through the gate oxide causing threshold instabilities and threshold voltage shifts. The diffusivity of boron is also greatly increased in a hydrogen ambient [9]. Introduction of a silicon nitride layer between the gate electrode and the gate oxide is a possible solution to this problem. But this in turn complicates the process and can produce hot electron instabilities in the NMOS device.

Another disadvantage of polysilicon gates is found in their low conductivity. In order to improve the gate conductivity, the so called "polycide"

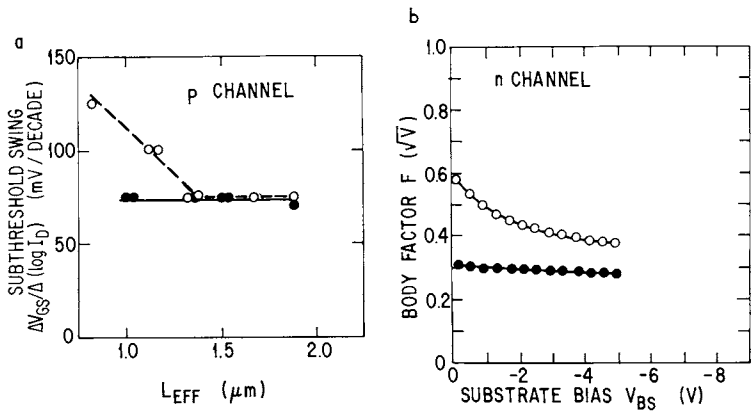


Fig. 3.2. a. Subthreshold characteristics versus effective channel length for polygate (open circle) and $TaSi_2$ gate (solid circle) PMOS devices; b. Backgate bias body effect versus substrate bias for polygate (open circle) and $TaSi_2$ (closed circle) gate NMOS devices; d_{GOX} = 22 nm, $L_{EFF} = 4.5$ μm [11].

gate has been widely adopted. The polysilicon, after being doped, is coated with a silicide, and then this gate material is patterned. Another approach is to selectively coat the gate after patterning the gates. This approach will be discussed later in this chapter. Another method of increasing the gate conductivity by an order of magnitude and also solving the work function problem is to eliminate the polysilicon entirely and utilize silicide directly on the gate oxide as the gate electrode. Another approach is to use refractory metal gates, thereby raising by two orders of magnitude the gate conductivity as well as solving the work function problem.

Recent work has shown that MOSFETs with refractory metal gates require about a factor of 6 times less channel doping than $n+$ polysilicon gate devices [10]. Reductions in subthreshold leakage have also been demonstrated when larger work function electrodes are used. For example, the use of $TaSi_2$ gates ($\phi_m \sim 4.5$ V) greatly reduces the subthreshold slope for channel lengths less than 1.3 μm as well as reducing the back-gate bias effect (Fig. 3.2) [11]. None of these advantages are obtained by using the so-called "polycide"[3] gate approach, because the electrode work function will still be that of the underlying $n+$ polysilicon. The fact that Mo or W have conductivities a hundred times higher than $n+$ polysilicon is also an advantage.

The difficulty with refractory metal gates in the past was in maintaining threshold voltage control because pure Mo was unavailable and because the columnar grain boundaries in Mo films trap impurities during the processing sequence. Past RMOS devices covered the Mo patterns with doped glasses, and the Mo was subsequently sealed and gettered using a high-temperature anneal [12]. Other methods include the use of Mo_2N to seal the top surface of the Mo [4] or the use of a Si_3N_4 coating [13].

Another difficulty has been the incompatibility of refractory metal gates with standard processing methods. Recent work has shown that refractory metal gates can be made more compatible with Si processing sequences by using a wet H_2 ambient so as to oxidize Si without oxidizing the metal gates [14,15]. For instance, this would be useful to form the screen oxide before the source and drain ion implantation step after forming gate sidewall oxide spacers used in lightly doped drain (LDD) devices. However, because of the columnar grain structure of these refractory metals, Si oxidation can

[3] A common approach for making polycide gates is to deposit WSi_2 on doped poly using silane and WF_6 in an LPCVD (low pressure chemical vapor deposition) reactor before patterning the gate level. Other approaches use cosputtered metal (e.g., Ta and Mo) and silicon. Since the sputter rate of Si is low compared to metals, the deposition rate is low and hard to control. The CVD method is therefore preferable. Other approaches coat the polysilicon with a silicide forming metal and then react the metal with the polysilicon to form the silicide either before or after forming the gate pattern.

also proceed under the gate electrode thereby thickening the gate oxide. A silicon nitride "cap" on top of the Mo has been utilized to prevent this.

In addition to all those advances, high-purity Mo and W sputtering targets are now available [16]. This work developed a chemical purification method to eliminate U and Th impurities before vacuum casting using high power e-beam melting. Uranium and thorium, being radioactive, can produce "soft" errors in large memories. CVD Mo was used for gates of refractory metal–oxide–semiconductor (RMOS) devices by reducing $MoCl_5$ in an H_2 atmosphere [17]. More recently W gates have been made using CVD W depositions by means of the H_2 reduction of WF_6.

Some of these problems are alleviated by the use of refractory silicide gates, which are more tolerant to oxidation. Also, if the refractory silicide materials are pure enough to begin with and their grain structure is more randomly orientated than the metals, the process contamination problems might be reduced. The $n+$ polysilicon gate or $n+$ polysilicon gate with a silicide coating are still the industry standard gate electrode materials. Time will tell if the advantages of refractory metal gates will be utilized in future submicron integrated circuits.

II. REDUCTION IN DEVICE PARASITICS

Above the device level, the areas that influence circuit performance, density, yield, and reliability are contacts, interlevel dielectrics, interconnections, and multilevels of metal. All these features have parasitic resistances and capacitances associated with them as well as reliability concerns.

These topics will be discussed: device parasitics, contacts, junction capacity, interconnections, and multilevel metallization technology. The reason for this topic order is that it is important for the reader to understand the various device parasitic factors and their relative importance in determining device and circuit performance and density, and in turn, understand how advances in metallization technology can lead to improvements in both. For instance, the use of the gate oxide sidewall spacer has made it convenient to strap both the polysilicon gate and the source and drain regions with high conductivity materials using either selective silicides [18,19] or W selectively deposited from WF_6 [20]. Many investigations have focused on the possible circuit speed enhancements achieved by doing this. What is discovered is that improvements occur only when the circuit speed is limited by transmission line signal delay times. In fact, in some cases, propagation gate delay as determined by using conventional ring oscillators shows minimal improvements with junction strapping or

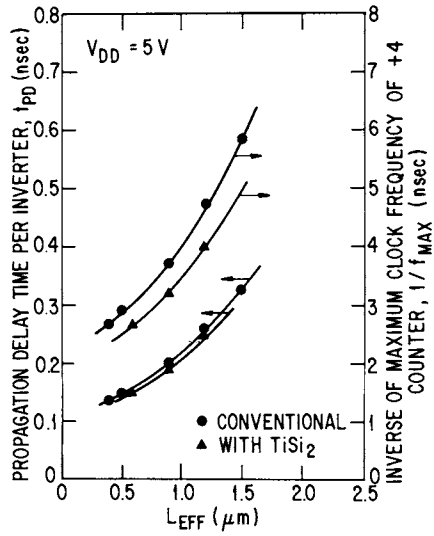


Fig. 3.3. Circuit speed versus effective channel length with and without $TiSi_2$ on S, D, and polygate [52].

even with gate and junction strapping. This is because gate delay in ring oscillators is primarily determined by channel transconductance, gate capacitance, and parasitic capacitances in the devices. At the device level, junction capacity is seen to be the dominant parasitic factor as observed by the curvature in the performance versus L_{eff} plots of Fig. 3.3. For instance, an analytical study has shown that junction capacitances account for up to 50% of the total capacitance in logic gates [21]. Therefore, reductions in device diode capacitances should produce corresponding decreases in gate delay. This conclusion is also supported by the data in Table 3.3 by comparing gate delays on insulating and noninsulating substrates.

As the following sections will show, the methods of reducing parasitic resistances are also means of reducing parasitic capacitances and in turn offer methods of increasing circuit density and speed. The following sections discuss these factors and the processing and layout options that can be used to reduce these parasitics.

A. Device Parasitic Resistances

Since current flow in the ohmic contacts to MOSFET source–drain regions is generally nonuniform [22] (Fig. 3.4), the proper scaling behavior of the contact impedance with decreasing device size is not immediately clear. In fact, experiments have shown [23,24] that the potential drop at

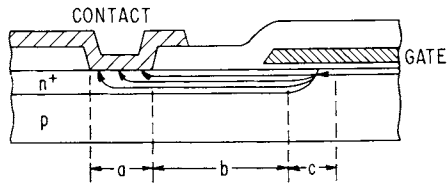


Fig. 3.4. The three regional components of device parasitic resistance: a, metal to silicon contact resistance; b, diode sheet resistance; and c, spreading resistance at channel end.

the device contact does not increase nearly as rapidly as the inverse square of the contact area because of the current crowding at the device edge of the contact.

The tenets of ideal scaling state that since all physical MOSFET dimensions (i.e., channel length and width, gate dielectric thickness, source–drain junction depth, contact hole diameter, and distance from gate edge to contact hole) decrease by some factor α, then the drive voltage also should decrease by α, while the dopant impurity concentration increases by α. As a consequence of these reductions, the total drive current decreases linearly with the scaling factor α, and since it is desired that all potential drops scale linearly as well, the series resistance should remain constant to accommodate the scaling rules. In order to maintain constant resistance for uniform current flow within the source and drain regions, the sheet resistance of these regions must be invariant to scaling. But this is difficult in practice, since decreasing the junction depth would require an increase of source–drain region conductivity. Unfortunately, the typical electrically activated impurity concentration for the source–drain is normally at the solid solubility limit. This also means that the specific contact resistivity for any given contacting material is already at the minimum attainable value since the specific contact resistivity decreases with the surface doping concentration.

Consideration of the potential drop across the contact is complicated by the observation that the current flow through the contact is not uniform because of the lateral current path to the FET channel under the gate. Therefore, this potential drop is not simply the product of contact area and specific contact resistivity (ρ_c). Analysis of the scaling behavior of the contact resistance is facilitated by the transmission line model (TLM) expression for the potential drop ΔV.

$$\Delta V = I(R_s \rho_c)^{1/2} \coth[L_c(R_s/\rho_c)^{1/2}]. \tag{3.2}$$

The TLM yields approximate values for the contact current density and electrostatic potential distribution. The ΔV of Eq. (3.2) is the potential

difference between the metal lead and the silicon source or drain region near the end of the contact closest to the MOSFET channel. Here, R_s and ρ_c are the source–drain sheet resistance and specific contact resistivity, respectively, while I is the transistor drive current per unit width and L_c is the contact length in the direction parallel to the current flow. Calculations using the above expression show that for small L_c ($L_c (R_s/\rho_c)^{1/2} \ll 1$), ΔV is proportional to ρ_c/L_c, which simply means that for small contacts the current is evenly distributed (no current crowding at the edge of the contact). However, when L_c is large ($L_c(R_s/\rho_c)^{1/2} \gg 1$), $\Delta V/I = (\rho_c R_s)^{1/2}$, and the voltage drop is independent of L_c (lots of current crowding). These calculations explain why measurements of V/I versus contact size, using 4-terminal Kelvin devices, increase less rapidly with decreases in contact size than simple dimensional scaling predicts [23,24].

These ideas are illustrated in Fig. 3.5. The dashed curve represents the contact voltage drop per current (per contact width) as a function of contact length for one set of contact parameters. When the contact length is large, the dashed curve (current crowding) shows that the voltage drop has little dependence on contact length. The solid curve (uniform current) is just the contact resistivity ρ_c divided by the contact length L_c. This ratio of contact resistivity to contact length is precisely the contact voltage drop in the simple, one-dimensional case of current flow *perpendicular* to the metal–semiconductor interface (e.g., the emitter contact in a vertical bipolar transistor). We add this solid curve for comparison of the two cases of vertical current flow (solid line) and planar flow (dashed curve) as in the MOSFET structure. The voltage drop for these two cases merge as contact

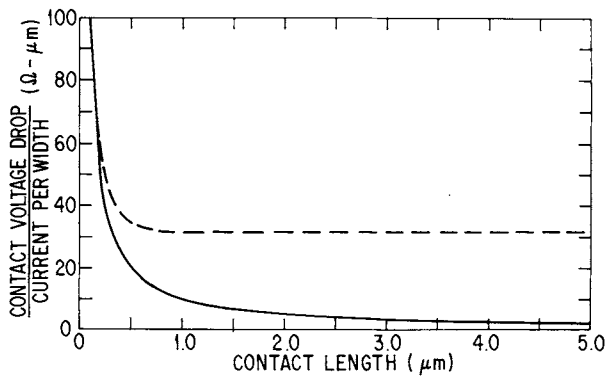


Fig. 3.5. Voltage drop normalized for current density and contact width versus contact length (voltage drop in units of V/IW or Ω-μm) assuming $R_s = 100\ \Omega/\square$ and $\rho_c = 10\ \Omega\text{-}\mu\text{m}^2$. Dashed curve: current crowding case. Solid curve: uniform current density in the contact.

length decreases. Figure 3.5 shows that if the contact length is reduced to $<0.5\ \mu$m, then the contact impedance rises dramatically because of the current crowding at the edge of the contact. In other words, the contact length must not be smaller than the current crowding region and ideally should be long enough to keep the contact impedance in the region of constant value or long contact limit ($L_c > 0.5\ \mu$m for the illustrated case).

The spreading or injection resistance at the point where the current enters the channel does not increase much with scaling. This resistance is

$$R_{inj} = (2\rho/\pi W) \cdot \log(\sqrt{8}\, d/\pi c), \tag{3.3}$$

where ρ ($=1/\sigma$) is the source resistivity and d, W, and c are the junction depth, width, and inversion layer thickness, respectively [25]. When all lengths (d, W, c) decrease by the same factor, the total parasitic resistance including source and drain sheet resistance increases, but the relative proportion of the total resistance due to source–channel current crowding remains constant.

Calculation of R_{inj} and the parameters used for Fig. 3.5 gives a value of 14 ohm-μ, which is close to the value previously calculated [26]. The contribution of the source–drain sheet resistance is critically dependent upon the spacing between the contact edge and the gate. For a 0.5-μm device, for instance, using the parameters in Fig. 3.5, the 50 ohm-μ resistance is slightly larger than the calculated 35 ohm-μ contact resistance for a $0.5 = \mu$m contact shown in this figure.

It is important to note that strapping the source and drain regions with self-aligned metal or silicide to improve the conductivity of these regions to about 1 $\Omega/\square$ is advantageous. This maximizes the silicon contact area, and interconnecting metal contacts can thereby be minimal in area since specific contact resistivities between metals is typically $\leq 10^{-9}\Omega$-cm^2. This is two orders of magnitude lower than nominal metal to silicon specific contact resistivities ($\sim 10^{-7}\ \Omega\ \text{cm}^2$) obtained at or near solid solubility silicon doping limits.

The above evaluation on the relative sizes of these parasitic device impedances allows for a quantitative analysis of the required contact technology for downscaling. The analysis below shows that Al contacts are sufficient for device sizes above 1 μm; refractory metal contacts would be advantageous for the 0.7–1 μm range and self-aligned metallization strapping of the source–drain is called for below 0.7 μm. Of course, other requirements such as better reliability and increased gate and substrate level conductivities for higher speed interconnections, may require these changes at larger design dimensions. Also, the requirements are more severe if a 5-V power supply is maintained.

1. Projected Limits of Parasitic Series Resistance in MOS Downscaling

In the past, when large design rules were used, the parasitic resistance was a minor component of the total MOS impedance of a turned-on device. Scaling changed this condition, because it decreased the channel resistance through a reduction of channel length and gate insulator thickness, while maintaining a constant voltage. Conversely, smaller contacts without a corresponding decrease of specific contact resistance had the opposite effect on the parasitic resistance, determining its increase with downscaling. Obviously, it is important to exploit the advantages of shorter channel length and inherent performance gains by controlling the rise of parasitic resistance. A suitable criterion could be to require that the parasitic series resistance itself could be no more than 10% of the channel impedance. Though the choice of 10% is somewhat arbitrary, it represents the ratio of these impedances in a typical 1.0-μm process. Therefore this is an acknowledgement of the present state of the art and can be a resolution to limit further inroads of the parasitic resistance.

The purpose of this section is double-fold. At first, the contact resistance requirements will be determined for a CMOS process with design rules ranging from 0.3 to 1.5 μm. The contact resistance has been singled out in this analysis since it can be more easily influenced by technology, with the choice of an improved contact metallurgy. Second, three different technologies will be compared in terms of their ratio of parasitic to channel resistance in the above scaling range. The intent is to forecast the application windows of these technologies and to determine at which resolution level there is a need for a change in technology.

As described, the MOSFET's channel impedance is in series with a parasitic resistance, which has three major components: (1) contact resistance, (2) source/drain sheet resistance, and (3) injection resistance at the edge of the channel. As a result, the observed MOS conductance is less than the channel conductance, being reduced by a factor containing the product of channel conductance and parasitic resistance [27]. Adopting conventional notation and indicating with lowercase the resistance terms normalized with unit width, the total MOS impedance, r_{tot}, is given by

$$r_{tot} = r_{ch} + r_{cont} + r_{sh} + r_{inj}, \tag{3.4}$$

where the right-hand expression lists sequentially the channel, contact, source/drain sheet and channel edge injection resistances. Their formulas are

$$r_{ch} = \frac{L_{eff} + V_{DS}/E_c}{\mu_o C_{ox} (V_{GS} - V_T - 0.5\ V_{DS})}, \quad \text{if} \quad V_{DS} < V_{DSAT}, \tag{3.5}$$

$$r_{cont} = 2(1 + W_{sp}/W_c)\,(R_s\rho_c)^{1/2} \coth\,[L_c(R_s/\rho_c)^{1/2}], \tag{3.6}$$

$$r_{sh} = 2R_sL_s, \tag{3.7}$$

$$r_{inj} = (4R_sd/\pi) \ln\,(8^{1/2}d/\pi c). \tag{3.8}$$

These equations need some explaining as to their origin and the meaning of the less conventional symbols.

Equation (3.5) is a modified form of the MOS channel resistance equation, which includes in the numerator the term V_{DS}/E_c. Its origin stems from the mobility dependence on the longitudinal electric field, which causes a more severe mobility reduction in a short channel MOS even for low V_{DS} [28,29]. This equation applies to the linear region of the MOS characteristics and was chosen for a worst case comparison with the parasitic series resistance. Indeed, the channel resistance is lowest if the gate voltage, V_{GS}, is highest and the drain voltage, V_{DS}, is lowest. In this analysis, V_{GS} is set equal to the power supply voltage, V_{DD}, and V_{DS} to 0.1 V. This condition is graphically represented by the inverse of the slope of the top curve of the family of characteristics near the origin. The symbol E_c represents the critical field at the onset of velocity saturation. However, instead of including E_c in a factor also containing L_{eff} and V_{DS}, as normally done [28], the term V_{DS}/E_c was added to L_{eff} to obtain a simpler formula. This leads to a mobility decrease as being equivalent to an increase of L_{eff}, for channel resistance or transconductance calculations.

Equation (3.6) is derived from Eq. (3.2), except that it is multiplied by $2(1 + W_{sp}/W_c)$. The reasons for this multiplication factor are (1) the current flows through both source and drain contacts; hence, the contact resistance is doubled; and (2) not all the channel width is used for contacts in 1-μm processes, since multiple standard contacts of minimum size, W_c, are employed, separated by a spacing, W_{sp}. If $W_c = W_{sp}$, the contact resistance is again doubled resulting in a fourfold total increase.

Equation (3.7) accounts for the sheet resistance contributions of the source and drain to the parasitic resistance; R_s is the sheet resistance and L_s is the spacing of the contact from the edge of the channel. Notice that lightly doped drain structures, commonly used to minimize hot electron effects, would result in additional sheet resistance, but their effects have been neglected here for simplicity.

Equation (3.8), representing the injection resistance, Eq. (3.3), has been multiplied by a factor of two, since in the linear region this resistance occurs at both source and drain. In addition, the source/drain resistivity appearing in the original formula has been replaced with the product of sheet resistance, R_s, and junction depth, d.

It is now possible to determine the contact resistance requirements in a

TABLE 3.1
Typical 1-μm CMOS Process Parameters

$V_{DD} = 5$ V	$\mu_{on} = 500$ cm^2/V sec
$V_{TN} = 0.8$ V	$\mu_{op} = 200$ cm^2/V sec
$V_{TP} = -0.8$ V	$E_{cn} = 1.2$ V/μm
$L_{eff} = 1.0\ \mu$m	$E_{cp} = 3.0$ V/μm
$L_s = 1.0\ \mu$m	$R_{sn} = 40\ \Omega/\square$
$L_c = 1.0\ \mu$m	$R_{sp} = 110\ \Omega/\square$
$W_c = 1.0\ \mu$m	$\rho_{cn} = 60\ \Omega(\mu\text{m})^2$
$t_{ox} = 250$ Å	$\rho_{cp} = 10\ \Omega(\mu\text{m})^2$
$C_{ox} = 138$ nF/cm^2	$d_n = 0.25\ \mu$m
$c = 100$ Å	$d_p = 0.40\ \mu$m

scaled-down CMOS process and its dependence on the minimum feature size of the design rules. For this analysis, some technological assumptions are required for assigning numerical values to the variables of the equations and for defining their changes with downscaling. Since CMOS is assumed, some symbols have been modified by adding n or p to distinguish between NMOS and PMOS properties.

A reference base must be established by listing the electrical and physical parameters for advanced state of the art CMOS, represented by a 1.0-μm process with aluminum alloy contact metallurgy. Table 3.1 lists typical values.

As downscaling takes place, many of the process parameters must change to optimize the devices at higher levels of resolution. Though rigorous scaling is not feasible because of technological constraints, it is nevertheless a useful guide. An educated guess of the evolution of the major process parameters is presented in Table 3.2. Only the primary process parameters are listed. Though L_{eff} is the only horizontal dimension in Table 3.2, its value represents the typical design rules feature size and is shared by L_s, L_c, and W_c.

TABLE 3.2
Predicted Evolution of Major CMOS Process Parameters in Actual Scaling

L_{eff} (μm)	V_{DD} (V)	t_{ox} (Å)	d_n (μm)	d_p (μm)	μ_{on} (cm^2/V sec)	μ_{op} (cm^2/V sec)	R_{sn} ($\Omega/\square$)	R_{sp} ($\Omega/\square$)
1.5	5	300	0.37	0.60	508	203	27	40
1.25	5	275	0.31	0.50	500	200	32	88
1.0	5	250	0.25	0.40	490	195	40	110
0.7	3	150	0.17	0.28	463	183	57	157
0.5	3	100	0.12	0.20	433	168	80	220
0.3	3	70	0.08	0.12	357	134	133	367

The channel resistances for both NMOS and PMOS versus L_{eff} are plotted in Figs. (3.6) and (3.7), respectively. After equating the total parasitic resistance to 10% of the channel resistance and subtracting the contributions from series and injection resistances, the remainder is the maximum contact resistance allowed. Examining these figures leads to several observations.

First, as expected, the minimum channel resistance becomes smaller and smaller with an increase of resolution. Second, because of the low field mobility difference between holes and electrons, the PMOS channel resistance is about 2.5 times larger than the NMOS. This implies that if the parasitic resistance is limited to 10% of the minimum channel impedance, then a larger value is allowed for PMOS than for NMOS.

Third, the source/drain sheet resistance, r_{sh}, and channel edge injection resistance, r_{inj}, are invariant with scaling and their sums are in ~3.5 ratio

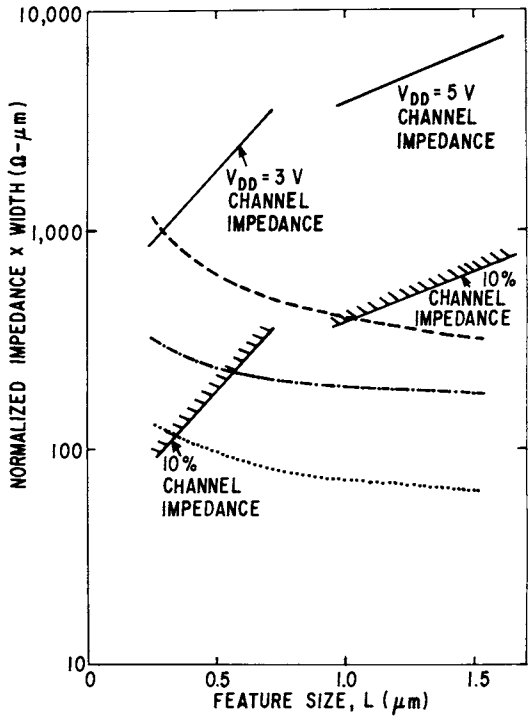


Fig. 3.6. Plot of minimum channel resistance and parasitic series resistance versus design rule feature size for NMOS. The impedance scale is normalized with channel width. The parasitic resistances are calculated for three different contact technologies: aluminum (------), titanium–tungsten alloy (-·-·-), and self-aligned tungsten strapping of source/drain (.....).

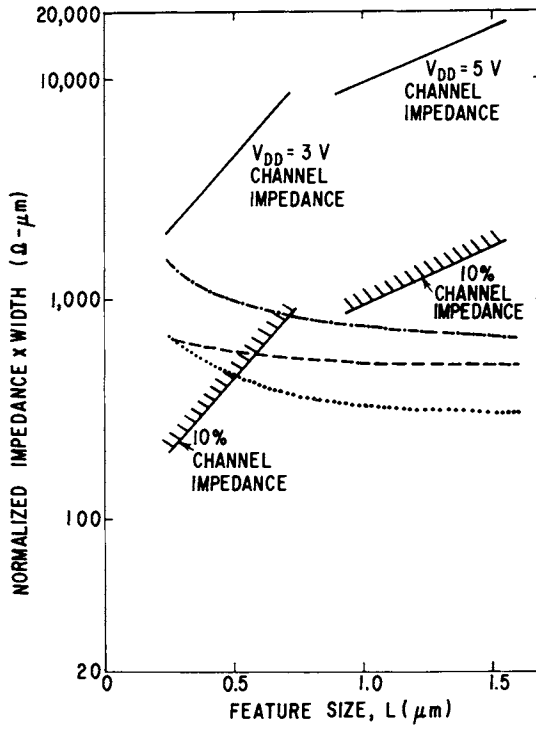


Fig. 3.7. Plot of minimum channel resistance and parasitic series resistance versus design rule feature size for PMOS. The impedance scale is normalized with channel width. The parasitic resistances are calculated for three different contact technologies: aluminum (------), titanium–tungsten alloy (-·-·-), and self-aligned tungsten strapping of source/drain (.....).

between PMOS and NMOS. The lack of dependence on scaling is mainly due to the increase of R_s, which is caused by using solid solubility doping concentration at the surface of heavily doped shallow junctions, and by the corresponding decrease of junction depth, as shown in Table 3.2. Since, in the formulas for r_{sh} and r_{inj}, the main term is the product of R_s and a linear dimension, L_s or d, the invariance of these resistances during scaling is justified.

Fourth, the contact resistance upper limits decrease very rapidly with an increase of resolution, becoming particularly severe for contact size below 0.7 μm. Notice that from 1.0 to 0.7 μm the contact resistance requirements stay nearly the same, if the power supply voltage is decreased from 5 to 3 V. This offsets the channel resistance decrease otherwise obtained from smaller channel length. Otherwise, if 5 V is maintained, the requirements will be more nearly equal those for devices below 0.7 μm.

However, the resistance quoted is for individual contacts and since the contact size is smaller during scaling, a lower specific contact resistance is still needed. This is process dependent and mostly affected by contact metallurgy and the doping level in the silicon at the contact interface.

Though aluminum and its alloys have been nearly universally used for first metal, the specific contact resistance to n+ is about six times larger than to p+, as shown in Table 3.1. This clashes with the device imposed conditions set forth in Table 3.2, whereby PMOS is allowed to have higher contact resistance than NMOS. It would be advantageous below 1.0 μm to use a contact metallurgy with a better match to the CMOS requirements. Titanium–tungsten alloys or tungsten itself are such metals, since their specific contact resistance values are reversed compared to aluminum in terms of silicon polarity [30].

Below 0.7 μm, the maximum contact resistance allowed is so small that it is beyond the reach of conventional MOS device structures, which are characterized by separate contact registration to the source/drain regions. The attendant sheet resistance from the contacts to the edges of the channel is nearly invariant with scaling. Hence, all the burden of keeping the parasitic resistance to the 10% fraction of channel resistance is placed on optimizing the contacts.

The obvious solution is then to eliminate this sheet resistance by means of self-aligned contacts, whereby the metal is strapped over the entire source/drain regions. A further advantage will be better utilization of the contact area in the channel width direction, since multiple contacts of standard size would no longer be required. The method of contacting the strapped metal to the interconnection metal is of little concern since the specific contact resistance between metals is usually very small.

This discussion leads to a comparison based on the previous scaling scenario between the present aluminum alloy contacts, the suggested refractory metals contacts (e.g., Ti, TiW, or silicides), and the self-aligned metallization strapping of source/drain using silicides or selective CVD W. For each of these technologies, the parasitic resistance is calculated versus design rule feature size. This is compared with the trend of minimum channel resistance, which is plotted in Figs. 3.6 and 3.7 together with the 10% channel impedance curve for a more direct comparison with the parasitic resistance.

Let us consider first the case of aluminum. As expected, there is a large discrepancy between the limiting feature sizes of this technology in NMOS and PMOS. These are easily derived from inspecting the graphs and recording the abscissas of the intersections between the 10% channel impedance curve and the "aluminum" parasitic resistance curve. The values are 1.0 μm for NMOS and 0.6 μm for PMOS, confirming the concern of using this metallurgy below 1.0 μm with optimal performance.

The situation is considerably more balanced with refractory metal (e.g., TiW) contacts. In this case, the limiting design rules dimensions are 0.6 μm for NMOS and 0.7 μm for PMOS. Therefore, this technology appears suitable for use in the high submicron range, down to 0.7 μm. This technology provides a convenient buffer between using aluminum and developing a self-aligned contact metallurgy with, for instance, tungsten strapping of the source/drain regions. This analysis, of course, ignores the possible beneficial reductions in parasitic junction capacity discussed later.

For still higher resolution, the graphs show the need for self-aligned contact metallurgy. The improvements are more pronounced in NMOS than in PMOS, for the smaller dimensions of 0.35 and 0.5 μm, respectively. At 0.5 μm, this is the optimal choice, because it is capable of retaining the 10% ratio of parasitic resistance to channel resistance.

Though these calculations have taken into account the most important factors foreseen in scaling, some elements have been neglected, because they were too difficult to quantify or their characteristics were still unsettled. The LDD structure falls in this category, though it is expected to play a major role in submicron devices because of the channel hot electron problem. However, it is difficult to predict its relative impact on parasitic series resistance, because its optimization is still underway and new implementations are continuously being developed. If the LDD structures were included in these calculations, they would have produced variations of the technology range of applicability, but they would not have altered their order.

In conclusion, the parasitic MOS series resistance and the minimum channel resistance have been predicted as a function of scaling for three different CMOS process technologies. With the constraint that the parasitic resistance be less than 10% of the channel resistance for both NMOS and PMOS, and omitting any potential reliability and other possible technical difficulties discussed in the text, the useful windows of these technologies were as follows:

Above 1 μm	Aluminum-type contact metallurgy
0.7 – 1.0 μm	TiW or refractory contact metallurgy
0.5 – 0.7 μm	Self-aligned tungsten or silicide strapping of source/drain

2. *Contacts*

Contacts to silicon for VLSI devices have received an enormous amount of attention. The various methods of determining the specific contact resistivity, ρ_c, have also been closely scrutinized [31,32,33,34]. Methods have been developed to correct for the current crowding effects normally present in Kelvin type contact resistance test structures [35].

Another subject that is equally important, but often overlooked, is the method of etching contact windows and the relationship between this process and the resultant contact resistance, including its reproducibility in manufacturing. Early VLSI work usually utilized a dry (RIE) + wet (BHF) etching sequence to obtain low contact resistance. However, for dimensional reproducibility and high yield, an all dry etching sequence is required. Early attempts at dry etching the contact vias were plagued with polymer formation at the bottom of the via. This problem was resolved by utilizing a RIE sequence that uses an appropriate oxide etching gas (e.g., $CHF_3 + Ar$) followed by resist removal and contact cleaning using O_2 plasma, O_2 RIE, or a small amount of Si RIE. This method produces excellent results [30].

Contact metallization is also beginning to change for reasons other than those required by the previously discussed scaling theory. For designs with junction depths greater than 0.3 μm, Al (1% Si) seems to be sufficient; however, to increase reliability and to avoid Al spiking failures, a TiW contacting/barrier layer has been utilized [36]. Such a layer has also been shown to reduce the incidence of open lines caused by Al electromigration failures, but it can produce interlayer shorts at high current densities [37]. This barrier layer has sometimes been combined with a silicide contact because most silicides are not good barriers against the diffusion of Al.[4]

A group of silicide forming materials (e.g., Pt, Ti, Pd) can be used to form a self-aligning contact by reacting the metal with Si and then removing the excess metal. The interconnecting metal is then deposited and patterned, with or without an additional barrier layer (e.g., Ti or TiW) between the silicide contact and the first-level metallization material. If the first-level metal is refractory (Mo, W), the barrier metal can be eliminated. A few studies have also concentrated on the use of deposited refractory silicide contacts ($MoSi_2$, WSi_2), which have very high thermal stability [42].

Obviously, the trend is in the direction of separating the contacting and interconnection functions. One exception to this trend might be in the use of nonselective CVD W [43], providing a contact/adhesion layer (e.g., Cr, Ti) is not required. Direct contacts using refractory metals (e.g., Mo) have also been tried [44], but the difficulty in using these metals is that they cannot reduce "native oxide" layers. This is also a problem with deposited refractory silicides. To overcome this difficulty, a number of active contacting materials that absorb or reduce native oxides have been investi-

[4] If Al is used as the first-level metal, a separate barrier material is usually used because silicides, which make good contacts to Si, are not good Al diffusion barriers [38,39]. TiN and CVD W (but not sputtered W) have been shown to be good Al barriers [40,41].

gated in conjunction with Mo interconnection metallization [30,45, 46,47]. Of these, the combination of a thin Ti layer in combination with a thick Mo (or W) layer holds great promise. The use of selective W [48,49] (or selective silicides) would seem to be a natural choice, since during the formation of these films, chemical reactions consume the native oxide. If selective W is formed at the bottom of the contact hole, the interconnection metal frame around the contact via can be eliminated because Mo and Al can be etched without etching W [50].

The minimal specific contact resistivities for these materials (W, PtSi) to heavily doped silicon is about 2×10^{-7} Ω-cm^2 for $p+$ contacts and 2×10^{-8} Ω-cm^2 for $n+$ contacts at surface concentration of about 2×10^{20}/cc [49].

The presence of gate side-wall oxide spacers could allow for the complete coverage of the junction area with selective metallization hopefully without producing gate-to-junction metallization shorts or excessive diode leakage. If a polysilicon gate is exposed to the reactants, the top of the gate line can be selectively coated, too. However, there are some unresolved problems with selective CVD W, for example, the parasitic replacement of Si with W at the contact interface. Unless properly controlled, this phenomenon can produce junction shorts through lateral encroachment of W at the junction perimeter. Another phenomenon that has been observed is the rapid etching of fine "wormholes" into the silicon at the junction edges. See, for instance, the description of these interfacial structures in Ref. 49. These fine holes could increase junction leakage. The use of silicides could also be a problem for shallow junctions, since their formation consumes silicon, and silicide encroachment into the junction can produce excessive junction leakage [51]. In addition, occasional losses of W CVD selectivity will coat the oxide spacer with W, which can produce gate-to-junction shorts if the top of the gate electrode is exposed. These problems, especially those associated with diode leakage, are currently receiving considerable attention. Currently, the use of selective $TiSi_2$ seems to be favored. Recent reports indicate that selective W using SiH_4 reduction of WF_6 rather than H_2 reduction holds promise. This process, if feasible, would be ideal because of its simplicity. The potential device and circuit improvements that could be gained by using these advanced processing techniques are described in the following section.

B. Device Junction Capacitance

A detailed discussion of junction capacitance is considered next. Figure 3.3 shows how performance begins to saturate for $L_{\text{eff}} \leq 0.7\ \mu$m. This saturation is attributed to parasitic junction capacitance [52].

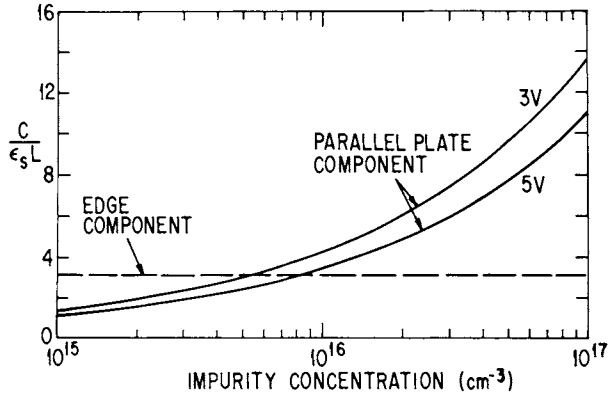


Fig. 3.8. Components of junction capacity versus impurity concentration, W = 3 μm.

Junction capacitance is composed of the parallel plate component and the edge component given by this equation [53]:

$$C/L \text{ (per unit length)} = \epsilon_s W \left[\frac{qN_a}{2\,\epsilon_s(V+\phi_b)}\right]^{1/2} + \epsilon_s \pi. \qquad (3.9)$$

The edge component, $\epsilon_s \pi$, does not decrease with scaling. As the doping increases to maintain threshold and punch-through control for smaller L_{eff}, the relative importance of the parallel plate term increases, especially if the operating voltage is decreased. The relative importance of the two terms can be evaluated in Fig. 3.8, where the two terms are compared as a function of doping concentration. The parallel plate component for fixed diode area becomes increasingly important for high doping concentrations required for submicron channel lengths. Indeed concentrations higher than about 10^{17}/cc are necessary to eliminate source-to-drain punch-through, consequently increasing junction capacitance.

Methods of Reducing Junction Capacitance

The edge component could be reduced nearly 50% if one junction edge could abut a vertical oxide isolation sidewall by using a deep trench isolation well. The silicon on insulator (SOI) method is a way of greatly reducing junction capacitance, especially the parallel plate term since MOSFET source and drain diodes can be extended downward to abut the insulator's surface. In fact, the SOI data in Table 3.3 supports the hypothesis that reducing junction capacity will greatly enhance speed. Notice that for the same or even slightly larger effective channel lengths, SOI gate delay is about half that of Si devices built on noninsulating Si substrates. Gate delays for MESFET and MOSFET GaAs digital switching circuits with submicron dimensions are also listed in Table 3.3.

TABLE 3.3
Gate Delay Performance versus Technology[a]

L_G (μm)	L_{eff} (μm)	τ_d (psec)	Technology
2.0	1.4	270	NMOS
1.0	0.85	150	CMOS
1.3	0.7	80	NMOS
	0.4	31	NMOS
	0.3	50	NMOS
	2.0	115	SOI
	1.0	95	SOI
	0.8	80	CMOS/SOS
0.8		77	GaAs MESFET
1.2		72	GaAs MOSFET
1.2	0.4	66	GaAs MESFET
0.5		34	GaAs
0.3		17	GaAs MESFET

[a] For reference listing see Reference 53.

The junction area is determined by layout rules, which in turn are controlled by the contacting processing methodology. The new methods being developed to make unframed and self-aligned contacts should lead to methods of making smaller device junctions. One method could use selective etching for patterning the top metal without etching the underlying contact metal [50]. This method etches Mo without etching the contact material (e.g., selective W or $TiSi_2$) in the bottom of the contact hole. Lift-off patterning of the interconnection metal (e.g., Al) would also be

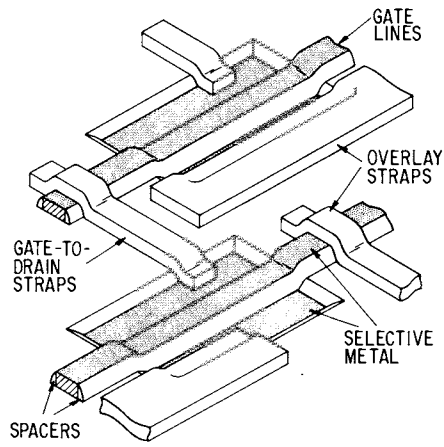


Fig. 3.9. Interconnections between gate and drains using selective metallization of S, D, and gates and unframed contacts whereby interconnections are formed before the first-level interlevel dielectric is deposited.

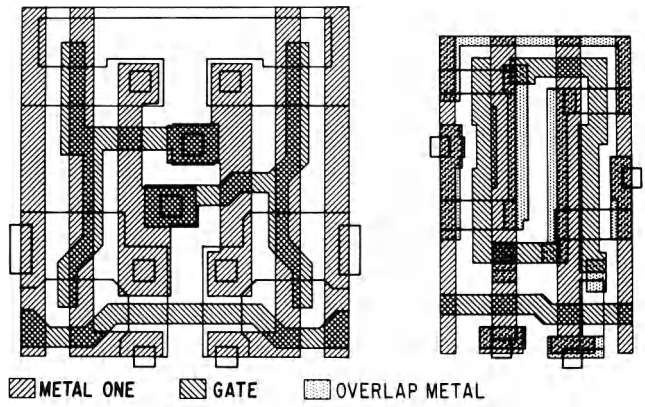


Fig. 3.10. Illustration of the size reduction obtained using unframed contacting overlay metal to locally interconnect gate and drain in a six-transistor memory cell.

suitable. However, it is doubtful that "lift off" is a viable method of producing submicron metallization patterns.

Figure 3.9 shows how, after coating the source, drain, and gate with selective metal, an unframed contacting overlay interconnection metal links gates to diodes of adjacent devices. The local interconnections are formed at the gate level before the interlevel dielectric between gate and first-level metal is deposited. The use of unframed contacts and short metallization runs to locally interconnect gates and drains in CMOS circuits, as described previously, should increase packing density dramatically. Another method of doing this has been described recently. It uses the TiN layer normally formed during the selective or self-aligned $TiSi_2$ contact metallization process as a local interconnection [54]. An example of how packing density is increased is shown for a six-transistor CMOS memory cell in Fig. 3.10. This kind of interconnection at the gate level is analogous to the old so-called buried contact used in high-density NMOS circuits in which the $n+$ polysilicon gate material performs the interconnection and contact function (Fig. 3.11) of adjacent gates and drains. This interconnect is not a full level of interconnection because gate crossovers are not allowed unless the top of the gate is covered with a sufficiently thick dielectric. This would then have to be selectively removed wherever a

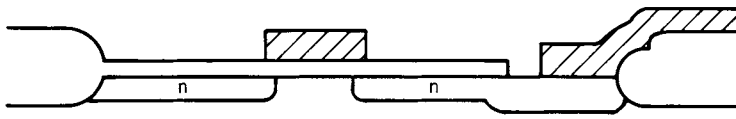


Fig. 3.11. Buried contact used in NMOS circuits to locally interconnect adjacent gates and drains. Contact hole to drain is cut in gate oxide before deposition and doping of $n+$ polysilicon. Out diffusion of phosphorus from $n+$ polysilicon links contact to drain.

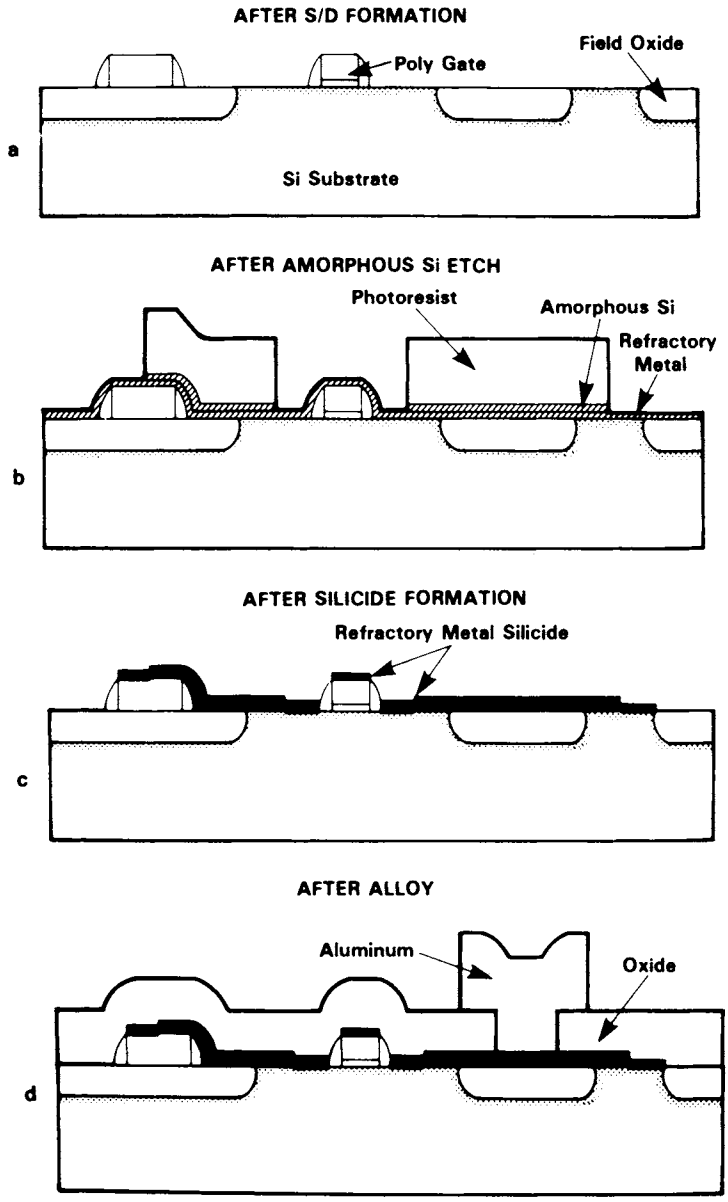


Fig. 3.12. Method of reducing junction area and device capacitance by contact extensions. a. Etched back isoplanar field oxide and polygate with oxide spacers; b. Following deposition of refractory metal (e.g., Ti) and amorphous silicon, the amorphous silicon is patterned without removing the underlying metal; c. After removing the resist, metal silicide is formed by reaction of the refractory metal with the single crystal silicon in the diode regions, top of the polysilicon gate, and patterned amorphous silicon (the unreacted metal is then selectively removed); d. Contact window etching forms the diode contact to the metal extension over the field oxide. The patterns formed in b can be used as a gate-to-diode local interconnect, thereby reducing layout area as in Fig. 3.9. (From Chen *et al.* [55].)

gate-to-overlay metal contact is needed. Another method of reducing the source and drain areas by extending the diode contact areas out over field oxide has been described [55]. This processing concept is shown in Fig. 3.12. The metallization extensions formed by performing a selective metallization sequence on a patterned amorphous silicon layer is subsequently used as a contact landing pad over the field oxide. Similar advanced device structures can be formed by using Mo as the local interconnection material since Mo can be etched without etching, for instance, selective W contacting and "strapping" the source/drain and gates.

The practice of using selective metallization to "strap" entire source and drains and thereby reduce diode areas requires that all the problems associated with excessive diode leakage produced by selective "strapping" of the entire surface of the diode are solved. In addition, the best material and process for the local interconnection pattern must be selected.

III. INTERCONNECTIONS

The previous discussion on the advantages and some of the possible methods of locally interconnecting gates and drains leads to the discussion of multilevel metal. Interestingly, the original 1970 Mo gate "RMOS" circuits were really two-level metal circuits because of the Mo gate level and the upper Al metallization lines [56]. Polygate circuits with two levels of metal are now becoming prevalent, and the addition of silicided or W-coated polygates produces three levels of high conductivity interconnections. If one were to add the gate level "local" interconnections described previously, a fourth layer would be available and useful for short, high-density, local "cell" interconnections and unframed contacts.

The reason multilevel interconnections are so important is that since most of the chip is covered with interconnections, multiple metal layers can reduce chip size, thereby decreasing propagation delay. In fact, it has been shown that the chip area dependence on multilevel wiring density is inversely proportional to the number of wiring levels [57].

A usual practice is to use closely packed lower levels of metal for local interconnections and thicker higher conductivity, but wider pitched metal patterns at upper levels for power supply buses and longer interconnections. It has been suggested that the interconnection cross sections of each level can be optimized to reduce parasitic capacitances between lines [58]. An exception to these spacing rules is found in gate arrays, where the metal pitch for all levels is determined by the coarser upper layer. The reason is that upper layer masks must be able to contact any of the underlying

features using a single router grid. Therefore, for gate array applications, there is a desire to have equal metal pitch for first- and second-level metal.

A. Trends in Methods and Materials

The traditional contact and interconnection material has been Al and various alloys of Al.[5] Aluminum has a number of ideal properties: (1) low resistivity ($\rho_{BULK} = 2.8\ \mu\Omega$ cm); (2) excellent adhesion to SiO_2; and (3) excellent wire bonding properties. However, because of its very low melting point (660°C), electromigration occurs at relatively low temperatures and low current densities. (This is in strong contrast to refractory metals, e.g., Mo and W, whose melting points are 2620°C and 3370°C, respectively.) Electromigration of Al atoms occurs at the grain boundaries within the metallization line. The electron stream creates a flow of these atoms because they are less tightly bound than those within grains where the atoms are bound in lattice positions.[6] Because the atom flow occurs along the grain boundary in the direction of electron flow, a grain boundary that extends completely across the metallization pattern will not erode as rapidly. A line with this type of grain structure is called "bamboo." Of course, this type of structure is not practical because it cannot be produced with any certainty and therefore can only be examined on an experimental basis. The practical method of reducing grain electromigration is to introduce impurities that "pin" or "stuff" the boundaries. In the past, these impurities have been primarily Si and Cu.

Until now, the additions of high percentages of Cu were required to significantly reduce electromigration. These alloys are difficult to etch and are prone to corrosion problems; therefore, they are normally patterned using a photoresist lift-off technique. Because of these complications, large numbers of circuits are still being made with Al (1% Si) or Al (1% Si) with a small percentage of Cu ($\leq 0.5\%$). One of the difficulties that has been occasionally experienced is the failure of Al lines caused by Si "nodule" formation. Nodule formation is a precipitation of silicon within the line that can occur during cool down or during the operation of the device [59,60]. As the nodule grows in size, the current density in the Al around the nodule increases and the line can eventually crack because of the stress

[5] Traditionally, this has been Al (1% Si) in order to prevent junction shorting or "spiking" failure that occurs when Al alloys with Si. Additions of Si inhibit the alloying and penetration of the Al–Si contact into the junction during contact "sintering" and device operation.

[6] At higher operating temperatures, lattice electromigration can also occur. Another electromigration failure mode has recently been identified. This is the electromigration of Al along a metal–SiO_2 interface. Contact electromigration can also be a problem.

around the growing nodule. Nodule growth can also cause interlevel metal shorts or time-dependent breakdowns [61]. Another similar phenomenon has been observed: the precipitation of Si at the Al–Si contact interface, which increases the contact resistance. Another traditional problem with Al metallization has been hillock formation. It is speculated that hillocks form by solid-state diffusion of Al to relieve the film stress during thermal cycling at temperatures below those where plastic flow can occur. Hillocks are also formed by electromigration [62]. Notice that there is a threshold current density that is inversely proportional to strip length [63]. If the strip length is short enough, no electromigration will occur because Al back pressure in the line is too high to allow void formation at the cathode. This is why pinhole-free passivation layers are important in retarding electromigration because they keep the line "pressurized." Annealing of the films before passivation can also be important [64]. The addition of Cu to the film reduces the rate of grain boundary diffusion that normally occurs in Al films to relieve the stress.

Recently, Ti has been substituted for Cu, and it has been shown that electromigration is reduced by small (0.2–0.5 wt. %) additions of Ti to an Al (1% Si) alloy [65]. In order to completely eliminate hillock formation and subsequent interlevel shorts, a rather high percentage of Ti (~3 at. %) is required in homogeneous films [66]. These additions of Ti increase the resistivity to between 4 and 6 $\mu\Omega$-cm. Another study has shown that thin multilayers of Ti and Al (1% Si) should also be effective [67,68]. The layered structures are promising because they maintain a lower resistivity through standard annealing temperatures than the homogeneous films (see Table 3.4). Also, if the top layer is Ti, the reflectivity is reduced, which makes photolithography easier. Another advantage of these systems is that the addition of Ti does not increase the difficulty of etching. Electromigration data comparing Al (1% Si) films to multilayered Al(Si)–Ti films show a 10–100 times improvement in the mean time to failure for the layered films [69,70]. It must be noted that it is especially important to use a good barrier between the contact to silicon if an Al–Si–Ti metallization is used because the solubility of Si in the Al_3Ti intermetallic compound can be as high as 15% [69].

A disadvantage of Al alloy systems is the high thermal mismatch to Si. Because the thermal expansion coefficient of Al (23.5×10^{-6}/°C) is about 8 times higher than that of Si (3.3×10^{-6}/°C), metal extrusion failure and cracking have been observed after thermal cycling [71]. Refractory metal thermal expansion coefficients (4.5×10^{-6}/°C) are a closer match and have been traditional packaging and heat sinking material in power devices.

Figure 3.13 shows the measured mean time to failure of Al alloy systems as a function of linewidth [72]. Obviously, this factor presents a problem

TABLE 3.4
Resistivity of the Metal Systems before and after Annealing[a] [68]

Metal systems	Type of system[b]	Resistivity before ($\mu\Omega$-cm)	Resistivity after ($\mu\Omega$-cm)
Al/1% Si	H	3.9	3.6
Al/1.2% Cu/1% Si	H	5.2	3.9
Pure Al	H	2.9	2.9
Al/1% Si/0.4% Ti[c]	H	4.8	4.3
Al/1% Si/1.4% Ti[c]	H	8.3	4.8
Al/1% Si/4.0% Ti[c]	H	7.8	6.0
Al/1% Si with Ti top (500 Å)	SL	3.6	3.5
Al/1% Si with W top (500 Å)	SL	3.8	4.0
3 layers of Al/1% Si and Ti (160 Å)	ML	3.7	4.1
3 layers of Al/1% Si and W (160 Å)	ML	3.7	4.1
Al/1.0% Ti	H	6.9	6.6
Al/0.4% Ti	H	4.6	4.7
3 layers of Al (pure) and Ti (200 Å)	ML	3.6	5.5

[a]From D. S. Gardner *et al.* [68].
[b]H, homogeneous film; SL, single layer of refractory metal used in film; ML, multiple layers of refractory metal used in film.
[c]Measured by electron microprobe. All values are atomic percent.

with continued shrinkage. There is another practical aspect to this problem. Current densities in metal lines often exceed design rule specifications because many times there is not a carefully considered "reflection rule" [73]. Also, the manufacturing engineer, in an effort to increase yield by eliminating intraline shorts in those portions of the circuit most densely

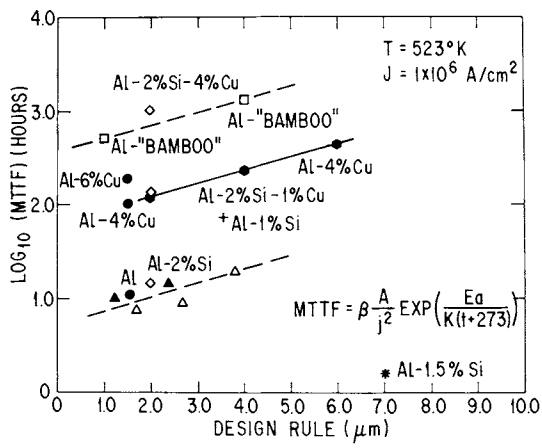


Fig. 3.13. Mean time to failure versus Al design rule. (From K. E. Gsteiger *et al.* [72].)

packed with metal, intentionally overexposes the positive photoresist pattern, thereby narrowing the lines. In addition, metal step coverage may not always be as high as expected. Any combination of these factors can cause unexpected circuit failures even for moderate current densities between 1 and 5×10^5 A/cm^2. These reliability problems are so serious that the viability of Al lines ≤ 1 μm wide is being questioned [61].

The reasons for the reliability concern are as follows. The size of the Si precipitates in Al (1% Si) alloys is about 0.4 μm, which is about the grain size of Al in fine-grained metal lines. If the films are cooled slowly, the size of the precipitates can grow to be as large as 1.5 μm. The same problem can occur with Cu additions, which limit the size of the silicon precipitates; the Cu–Al intermetallic grain size can grow to be as large as 1.5 μm. In any case, the line is extremely inhomogeneous, and the occurrence of one grain structure triple point or a large, 1-μm Si precipitate in a 1 or 0.5-μm line will produce a current flux divergence and rapid electromigration at this site [74]. These problems are so severe that Au is being considered for 0.5-μm lines [75].

An obvious solution to this problem is to use refractory metal interconnections (Mo, W). These interconnections have been utilized in high-density NMOS circuits where the power dissipation is very high [76]. In this instance, W metallization was used both at the first and second metallization levels with Al being confined to the bonding pads. These materials have other advantages in that they are easier to dry etch than Al alloys because their chlorides are volatile and, therefore, no corrosive by-products are produced during RIE. Also, their reflectivity is less than Al, which makes photolithographic linewidth control easier.[7] Although Mo and W interconnections eliminate electromigration and hillock formation, they do have some disadvantages. One disadvantage is that their adhesion characteristics are not as good as Al. Nevertheless, CVD and magnetron sputtered Mo is adherent to SiO_2. However, W seems to require an adhesion layer to promote adhesion to SiO_2. An excellent adhesion layer is a thin film of sputtered Mo [77], Ti, or TiW. The major disadvantage of these materials is the higher resistivity ($\rho_B = 5.4-5.7$ $\mu\Omega$-cm). However,

[7] The use of various materials in integrated circuit fabrication requires a knowledge of their optical properties when photolithography is used. Reflectivity and diffuse reflectivity influence the degree of linewidth and alignment control the process can maintain. For instance, both poly Si and Mo have low reflectivity and diffuse reflectivity, which is advantageous. CVD W has a very rough surface, and although its reflectivity is lower than Al, its diffuse reflectivity is very high, which can affect photoresist exposure and alignment accuracy. Double-level metal capacitor yield might also be decreased (Fig. 3.28) when thin dielectrics are used. The lithographic problems at submicron dimensions can require the use of thin antireflection layers on these metal surfaces.

TABLE 3.5
Mo at M1

No hillocks
Smooth surface
Caps for analog
Stable
No electromigration
No structural changes
Low reflectivity
Accurate patterns
Easy to RIE
No nonvolatile chlorides
Differential etching
Compatible with Al M2
Two deposition methods
Magnetron sputtering
CVD
Clean
Simple contact system
No barrier required
Thin sintering layer (CVD W, Ti, TiW)
No spiking failure
Compatible with CVD W plugs
Temp and no fluorides
Resistivity ~ 8 $\mu\Omega$ cm

the designer might consider the advantages of using higher current drive to charge capacitive loads within a cell when using these materials. The advantages of using refractory metals (e.g., Mo) at the lower levels of interconnection are given in Table 3.5.

When one considers the use of metals with higher resistivity than Al in a circuit, the implications of this to circuit performance must be considered. In the case of specific, 1.2-μm CMOS high-speed logic circuits, the use Mo at M1 (metal 1) and Al at M2 (metal 2) has been demonstrated to have the same performance as two levels of Al. The reasons for this are as follows. First, all power busing must be done using the lower resistivity or lowest sheet resistance material at the upper level to avoid voltage drops along the power bus lines that supply drain voltage to the underlying devices. Of course, for the long interconnections (e.g., intermacrocell or long memory read and write lines), the lowest resistivity should also be used. The signal paths can then either be carried at the lower level or, if need be, by the upper conductor. The reason that signal propagation delays do not suffer in 1.2-μm CMOS logic circuits when either Mo or Al are used for all the signal paths is explained by the analysis in Ref. 78. Figure 3.14 gives the

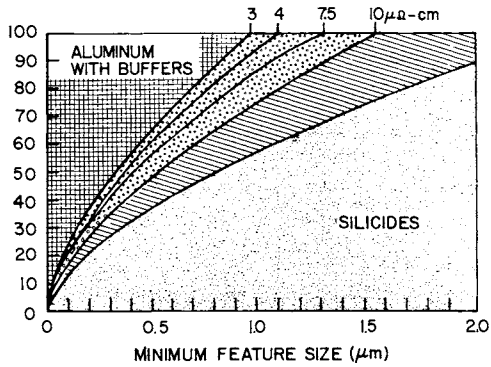


Fig. 3.14. Breakdown of materials for interconnections based on resistivity. The resistivities of Al, Mo, W, and silicide are 3–4, 7.5, 10, and 30 $\mu\Omega$ cm, respectively. The metal linewidth used is 1.5 times the minimum feature size, and the value tolerated performance loss (solid curves limit) is 15%. (From D. S. Gardner *et al.* [78].)

fractional percentage of the total number of signal paths versus scaling size of lines that can be used by materials of different resistivity before either a lower resistivity is required or the longer lines must be broken and connected with repeaters (buffer amplifiers). For instance, Al at 3 $\mu\Omega$-cm is good for design rules down to 0.9 μm. Below that buffers are required. A 10 $\mu\Omega$-cm (e.g., W) material is all right for 1.4-μm feature sizes. Also, Mo at about 7.5 $\mu\Omega$-cm is 100% good for a 1.2-μm CMOS VLSI process. Therefore, refractory material resistivities are suitable for logic circuits and actually should not reduce speed even when utilized for all the signal paths at these feature sizes.

It is also interesting to carry this discussion to scaling below 1.2 μm. At 0.5 μm, one finds (Fig. 3.14) that the comparative and practical differences in the use of either a refractory metal (M1) at 7.5 $\mu\Omega$-cm or Al or Au at 4 or even 3 $\mu\Omega$-cm are very small. All those materials will not permit more than about 65% of the interconnections to be made without inserting repeaters in the longer lines. The same kind of argument can be made at 0.8 μm but here 80–85% of the lines are good without repeaters. These conclusions are also supported by recent work which shows that for typical circuits the delay is dominated by RC delays associated with the source resistance as long as Al interconnection lengths are shorter than 1 cm at 300K and 2.5 cm at 77K. Therefore, arguments have been made that the use of superconductors will not enhance 77K circuit performance [79], but other insights might disagree with this view because pulse dispersion is virtually nonexistent [80].

As to the reliability aspects of these refractory metal systems, the Mo/TiW system shows no failures at 3.6×10^6 A/cm^2. Also, Mo/TiW or

Mo/Ti contacts are reliable to 600–650°C with no observed "spiking" or electrical degradation of the junctions [30,47].

Because of the preceding discussion, the thin film resistivities of refractory metal films need to be considered in comparison with Al alloys. Although Al maintains its resistivity in thin films because it deposits with large grain size, refractory metal films (e.g., Mo) deposit with finer grains and result in resistivities between 6 and 8 $\mu\Omega$-cm. Impure films, if contaminated with oxygen, will have higher resistivities. Comparison of these film resistivities with the new Ti/Si Al alloys (Table 3.4) shows that these refractory metal films will have to be thicker to produce the same sheet resistance. However, at the higher concentrations of Ti required to completely eliminate hillock formation, the resistivity of the homogeneous or layered films with resistivities of about 6 $\mu\Omega$-cm are only slightly lower than elemental Mo films ($\rho \sim 7.5$ $\mu\Omega$-cm). One advantage of these two refractory metals is that they can be deposited by CVD using either fluoride or chloride sources. Apparatus for CVD Al is not yet available, whereas CVD W equipment of both the selective and nonselective types is available. Some of the advantages of CVD metal will be discussed in the following section.

B. Multilevel Metal Processing

The trend toward modular VLSI design with computer aided routing of interconnections is pushing the technology toward multilayer metal structures. Besides easing the routing problem, this results in smaller chip size with attendant cost reduction because of the larger number of chips per wafer. In fact, it has been shown that the chip area dependence on multilevel wiring density is given by

$$A^{1/2} = (PG^{2/3})/n,$$

where A is the chip area, P the metallization pitch, G the number of gates, and n the number of wiring levels. This formula is an extension of Rent's rule. Analytical work [81] has shown that for high-density designs, the $G^{2/3}$ power is more likely to be $G^{0.2}$. Performance also benefits from multilayer metal designs since the interconnection length is correspondingly reduced.

The desire to maintain or even to reduce the interlevel capacitance of the interconnections to the underlying and overlying metallization lines requires that the interlevel dielectric thicknesses be maintained or made thicker. Unfortunately, this is done in spite of feature and via size shrinkage, which therefore presents the processing engineer with an interesting set of difficulties.

Metal pitch is determined by the minimum line and space dimensions

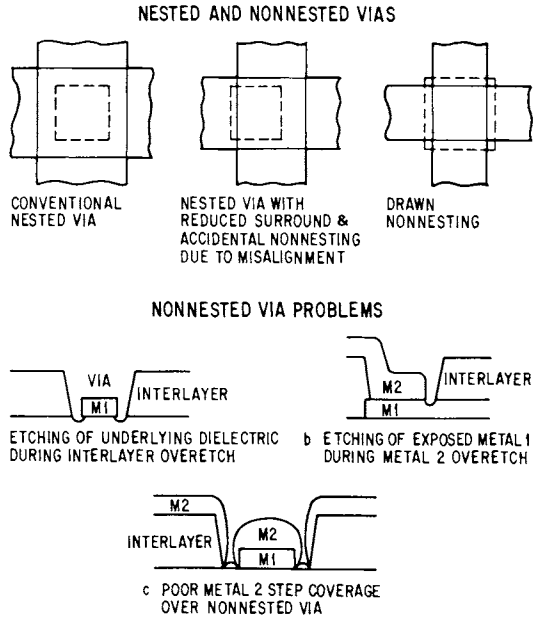


Fig. 3.15. Illustrations showing framed or nested vias (framing or surround size can be required for either or both levels of metal). Problems with unframed vias are caused by undercutting M1 during via etching (a) or when etching M2 pattern with M1 exposed to etch (b); the problem shown by (c) is even more severe when (a) occurs.

made possible by recent advances in lithographic techniques. Metal pitch is also determined by the via size and underlying metal pad size. When the underlying and interlevel dielectric is the same (e.g., SiO_2), the metal pad must be bigger than the via and must frame the via so as to allow for misalignment and possible metal undersizing (nested or framed via, Fig. 3.15). Unframed interlevel vias are therefore very desirable, but as described in Fig. 3.15, they have been plagued with many problems.

The use of two different metals (e.g., Al over Mo or Mo over W) to make unframed contacts or vias has been demonstrated when the underlying metal is different from the top interconnection metal, providing that the latter can be etched without removing the exposed portions of the metal underneath. This approach solves the problem shown in Fig. 3.15(b), but does not solve the oxide overetch problem [Fig. 3.15(a)]. For these reasons, different types of "etch stop" dielectrics are being investigated so that via etching will not undercut the underlying metal pattern [82].

Still another problem is that of filling or providing good metal coverage in small, deep interlevel via holes. Figure 3.16 shows that when conventional sputtering is utilized the metal thickness at the bottom edge of the hole can become very thin, much less than high reliability requires [83].

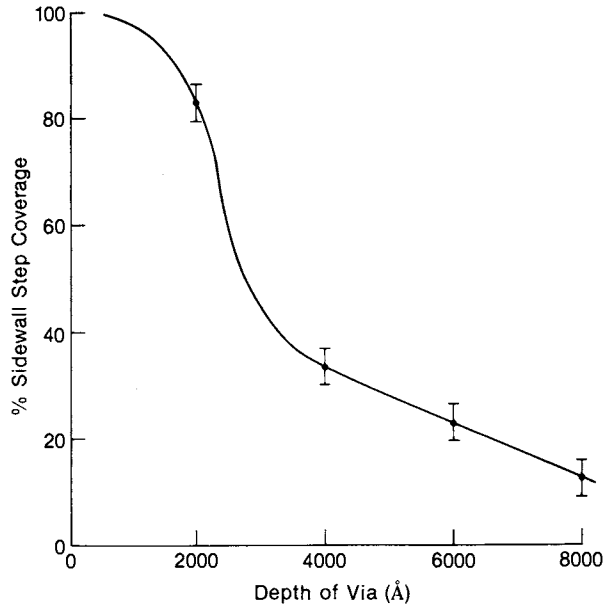


Fig. 3.16. The percent sidewall step coverage of a standard dc sputtered aluminum film (8000 Å thick) into 1.2-μm diameter vias of different depths. (From Saia, *et al.* [83].)

This problem can become very severe when planarization of the interlevel dielectric is used because of the variable via depth to different underlying features. Planarization is advantageous for patterning fine lines with good step coverage over the sharp edges of underlying RIE metal features. The disadvantage is that the via depth varies greatly, and certain types of "stacked" vias (e.g., M2 to M1 diode) can be very deep (Fig. 3.17).

Planarization can be achieved by spinning photoresist on the wafer and baking at elevated temperatures to cause the resist to flow and thereby level

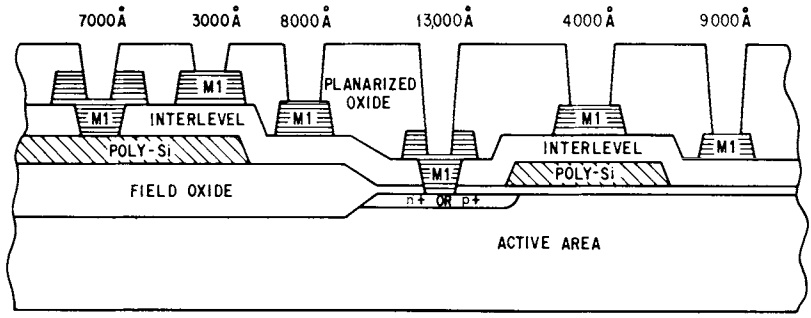


Fig. 3.17. Various via depths that can occur after planarization of the second-level interlevel dielectrics.

the surface [84,85]. Subsequently, plasma or reactive ion etching simultaneously etches resist and oxide. Best results are obtained when the etch rates of the interlevel dielectric (SiO_2) and photoresist are identical. Identical etch rates can be achieved by adding oxygen to a fluorine-based plasma, such as CF_4 or CHF_3, or by diluting NF_3 with Ar [86]. Another method of planarization uses RF bias sputtering of SiO_2 [87] or plasma-enhanced biased LPCVD. Other methods utilize "spin on glass," which deposits small glass particles suspended in a liquid on the wafer. Subsequent baking of the wafers drives off the liquid. Subsequent etch back of this layer followed by deposition of SiO_2 results in a "semiplanar" surface.

The use of doped "flow glasses" can also achieve a partial planarization of underlying features if the glass can be deposited thick enough and the flow temperature is not too high to degrade the underlying devices. Examples of "flow glass" are $P_2O_5-SiO_2$ containing 5–15% phosphorous, and $P_2O_5-B_2O_3-SiO_2$. Unfortunately, high temperatures are required to achieve any appreciable flow. Because of this requirement, several efforts have been undertaken to increase the high temperature tolerance of the underlying contact and metallization systems. Whereas refractory metals can withstand these temperatures when they are over silicon dioxide, specialized Si contacting layers must be utilized to limit any chemical interaction with silicon. Reference 42 describes how this was done using WSi_2. Nevertheless, these glass layers have excellent passivation properties and are traditionally used between the gate level and first metallization layer to block the passage of "dirt," such as alkali metals (Na, K), from entering the device regions. The use of the TiW barrier layers requires these passivation layers under the TiW because sintered TiW sputtering targets have been traditionally very dirty. High-purity TiW targets have been recently made available, but doped glass passivation layers are still required to eliminate the random dirt that occurs during manufacturing.

There are a number of basic approaches being used or studied for high-density, multilevel metal. These are now summarized.

1. No planarization of interlevel dielectrics:
   a. In order to solve the step coverage problem, the process could use a conformable coating CVD metal or place oxide "spacers" [88] on the underlying metal edges. If the via holes are small, deep, and untapered, one of the via filling methods mentioned in 2.b. will be required.
   b. An advantage is that via depth is the same for all features independent of their topographical height above the silicon surface; this reduces the via dimensional control problem (especially if tapered vias are utilized to attain good metal coverage in the via).

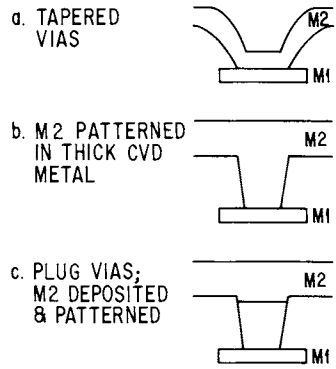


Fig. 3.18. Methods of filling small vias. a. Tapering the via during via etching by RIE erosion of the photoresist will give good metal step coverage but the dimensions are difficult to control especially when the vias have different depths; b. M2 is a CVD refractory metal that produces excellent step coverage; however, since M2 must be thick, M2 is more difficult to pattern; an alternate approach is to use bias sputtered Al or Al alloys; c. Vias are filled using a CVD metal process and the excess metal in field is etched back (without masking); M2 is then deposited and patterned. In c, all vias should be small enough to be completely "plugged," otherwise the maskless etch-back step is likely to etch M1. An alternate method is to use selective CVD W plugs.

c. However, patterning of overlying metal is more difficult (especially if tight pitch is required).

2. Planarization of interlevel dielectrics:
   a. Via depths are variable (good via etch stop is therefore required).
   b. Via hole filling requires either a plug-filling process (Fig. 3.18), using either selective W [89] (deposits selectively on conductors but not insulators) or nonselective W, or bias sputtering [90,91] of

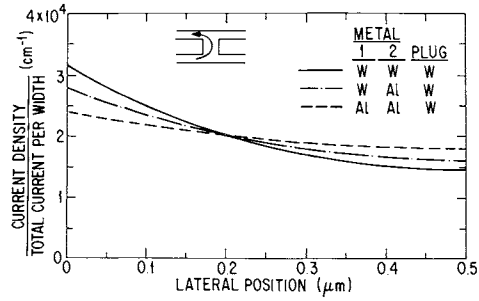


Fig. 3.19. Current density versus position within the via plug. Plug dimensions: 0.5 μm wide and 0.5 μm deep. Solid line covers the case when all levels (and plug) use the same metal (W or Al), and the other two cases are as designated for M1, M2, and plug material. (From Pimbley and Brown [101].)

metal.[8] A variation on this is to deposit a thin film of nonselective CVD W and cover it with Al and then pattern both layers [93]. An alternate method (Fig. 3.20) of forming interlevel metal connections using via "pillars" has been described [94], but its practicality for small feature sizes and variations in underlying topographical feature height has not been demonstrated. One of its advantages is that no underlying metal frame is required because the interlevel dielectric is deposited after forming the via metal pillars.

An example of recent processing advances for filling advanced vias is shown in Fig. 3.21. A W plug in a 1.3-μm-diam via hole over 2 μm deep is shown. This via plug was formed using a new cold wall, very high rate ($T \sim 600°C$), highly selective CVD W process. The underlying first-level metal is Mo/TiW; the second-level metal would subsequently be placed on the planar surface and patterned. Molybdenum is an ideal metal for the underlying interconnection material because of its high thermal stability and because WF_6 being more stable than MoF_6 eliminates Mo fluoride formation. This makes Mo compatible with the WF_6 used to form the selective plug. If Al were used as the underlying metal, a barrier material, such as Mo, W, or $MoSi_2$ or WSi_2, might be required to passivate the Al against the formation of nonvolatile AlF_3, which if formed greatly increases the contact resistance between the plug and the underlying Al metallization [95]. Recently, however, it has been shown that the contact resistance between selective CVD W plugs and Al can be reduced by using deposition temperatures above 400°C [96]. The use of different types of metals within the same structure requires the understanding of the interdiffusion and electrical properties of these couples [97] as well as the electromigration reliability [98]. The advantage of using the selective deposition method is that it fills the via from the bottom whereas nonselective CVD W coats all surfaces conformably. The use of nonselective CVD W presents the possibility, especially for high aspect ratio vias, of closing off the top of the via early in the process, thereby leaving a void. Another advantage is that the selective process described here is tolerant to variations in thickness of the deposited metal. This is in contrast to the use of nonselective CVD W and metal planarization etch back because variations in thickness produced by a CVD "blanket" coating make it difficult to remove the thicker W in the field without reducing the height of the W via plugs in those areas of the wafer where the W in the field is thinner.

[8] An interesting note is that the current crowding that occurs in the upper conductor is reduced when the plug material has higher resistivity than the upper level metal (Fig. 3.19). This observation is similar to that which shows decreases in contact resistivity will decrease the electromigration mean time to failure as described by Ref. 92.

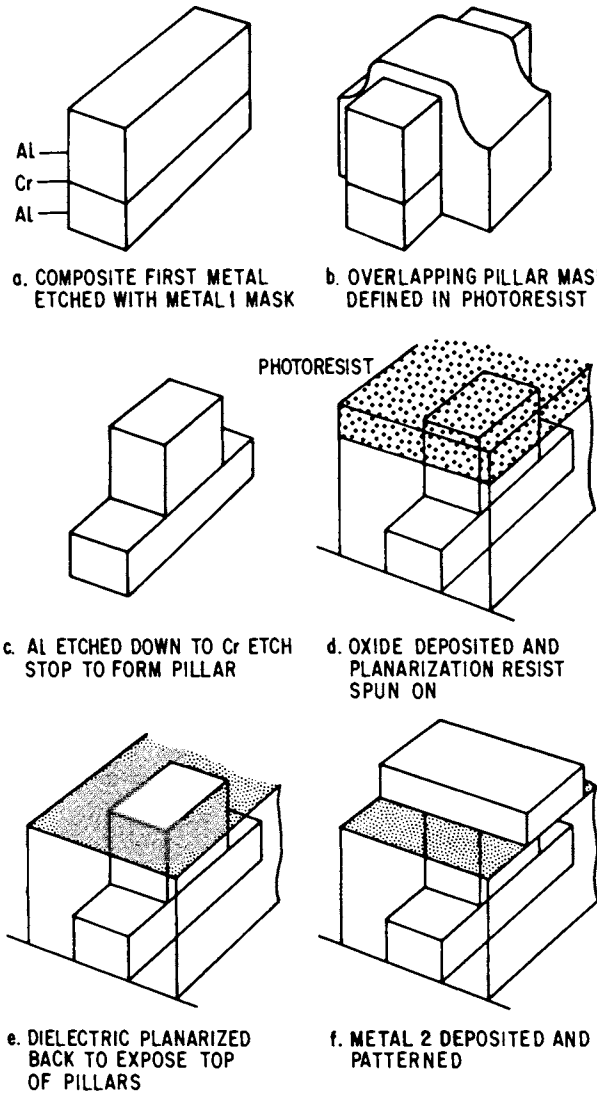


Fig. 3.20. Process sequence for the pillar process. a. M1 and the pillar metal are deposited and etched in M1 pattern; b. Photoresist overlap pattern is defined for pillar formation; c. Pillar is etched; d. and e. Interlevel dielectric is deposited and planarized; and f. M2 pattern is formed. (From R. E. Oakley *et al.* [94].)

(Microloading occurs in these areas.) In addition, a nonselective conformable coating produces a problem because of the thicker W that occurs over any unplanarized oxide step produced by the topography of underlying features. In this case, the metal planarization process will tend to leave a W "stringer" on these steps. The presence of metal stringers can produce

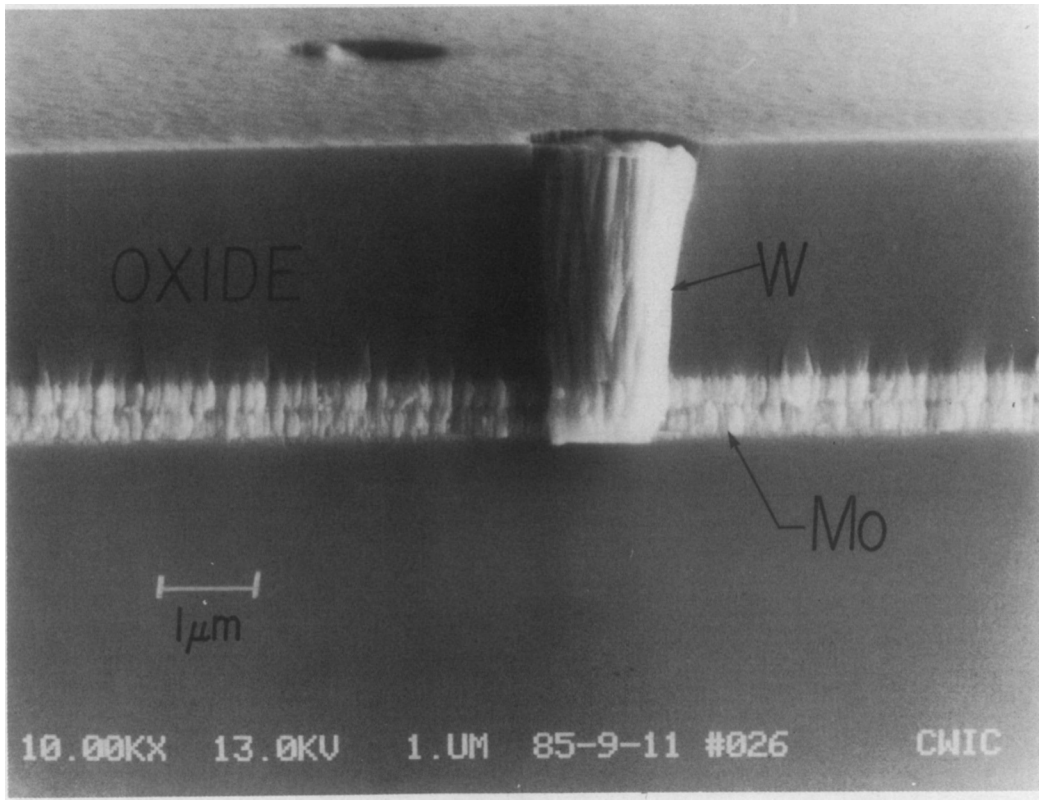
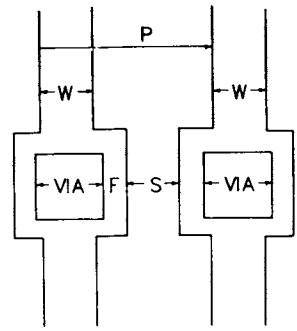


Fig. 3.21. A W via plug 1–1.4 μm in diameter formed by filling a 2.2-μm-deep via hole using selective CVD W. The first-level metal layer is 0.6 μm of Mo. The sample has been cracked through the middle of the via and through the Mo layer and interlevel dielectric. The picture, taken at an angle, shows how well the W plug conformably coats the via. The entire planarized structure is ideal for depositing and patterning the next metal interconnection layer. (From R. H. Wilson *et al.* [89].)

intralevel shorts. Removal of these stringers would require a large degree of overetch, which would reduce the height of the W in the via holes.

The selective W via filling process has been shown to be compatible with planarization of the interlevel metal dielectric which makes tight pitch patterning of M2 easier. Furthermore, this process results in nearly 100% feature coverage by M2 because M2 is not required to fill vias and there are only a few minor steps in the planarized dielectric and at the via openings [99]. The process is also thermally and chemically compatible with Mo as a first-level metal. Electrical TEG measurements on processed wafers exhibit very high yield and very low M2/M1 contact resistance ($\leq 10^{-9}$ Ω-cm^2) [99]. Furthermore, this methodology provides a means of substantially reducing interconnecting metal pitch. This is because the via holes can be small and straight and because M2 frames around the via can be eliminated since Al etch does not etch W. The advantages of reducing the via size and eliminating the M2 frame around vias is shown in Fig. 3.22. For instance, for purposes of illustration, the M1 and M2 pitches can be made the same by reducing the via hole to 1.3 μm and eliminating the M2 frame but without requiring more critical alignment of via to M1. This approach makes unrestricted vias (Fig. 3.17) feasible at tight metal pitches [99]. A more detailed description of the selective W plug process and its advantages is given at the end of this section.

After planarization of the interlevel dielectric, very small, straight vias are etched through the dielectric to the underlying metal landing pads. The



UNSTAGGERED VIA METAL PITCH

WIDTH	SPACE	FRAME	VIA	M2 VIA PITCH
2	2	0.6	1.8	5.0
2	2	0.6	1.3	4.5
2	2	0.0	1.3	4.0
				M1 VIA PITCH
1.5	1.5	0.6	1.8	4.5
1.5	1.5	0.6	1.3	4.0

Fig. 3.22. Metal pitch at unstaggered interlevel vias.

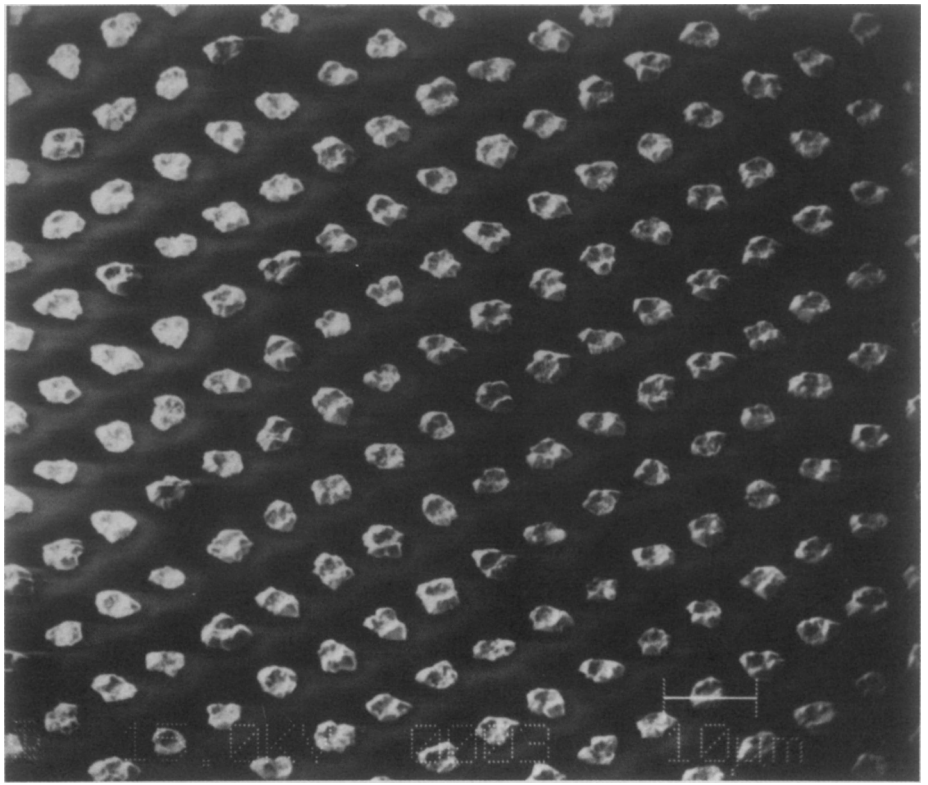


Fig. 3.23. Nailheads that are produced when via holes are overfilled with selective W. Note that the absence of W in the field illustrates the high selectivity of the process.

small via diameter allows the landing pad size to be reduced without violation of the via framing rule. In addition, deep and shallow vias are allowable because the via is straight. Small tapered vias of different depth would be difficult to fabricate, and the etching time required to open the deeper vias would enlarge the bottom of the shallower vias causing them to become "unframed." After etching the straight hole in the interlevel dielectric, high-rate, selective CVD W is then utilized to fill the via holes.

The time required to fill the deep vias using the selective process overfills the shallower vias resulting in "nailheads" above the interlevel dielectric's top surface (Fig. 3.23). These nailheads are selectively removed by a planarization etch-back technique that planarizes the protruding via plugs so that the tops of the via plugs are even with the surrounding dielectric. This metal planarization technique, which can etch photoresist and refractory metals at the same rate, is used to remove the nailhead and any growth of tungsten on the oxide field that could occur because of any localized nucleation site. Metal 2 (e.g., Al or Al alloys) interconnections can now be formed using conventional sputtering and patterning. The interconnection metallization layer that is subsequently formed, therefore, has 100% step coverage over via plugs and underlying topology. Furthermore, high-resolution patterning of this layer is easy because of the planarity of the surface. Since Al (or Mo) can be etched without etching W, the top metal layer need not frame or completely cover the via. This greatly increases top level metal pitch. Figure 3.24 shows a stacked contact between M2 and an active area diode (see Fig. 3.17).

The high-rate, highly selective CVD W process has been studied in detail and has been shown to produce low contact resistances to both refractory metallization and Al metallization layers. Furthermore, the reliability aspects of these metallization systems has been studied and initially defined to be highly reliable provided a highly reliable Al alloy is utilized [98].

IV. AN APPLICATION OF ADVANCED METALLIZATION TECHNOLOGY

Present design rules are usually quite restrictive on via types allowing only one or at most two types out of the many types possible. Allowance for vias of many different via types, including stacked vias and variable depths, is useful for all design methodologies but becomes especially useful for gate arrays. This is because gate arrays utilize an array of gates previously fabricated before metallization starts. Circuit fabrication is obtained by interconnecting the gates with metallization patterns. It is desirable, therefore, to contact all the underlying features with M1 and M2

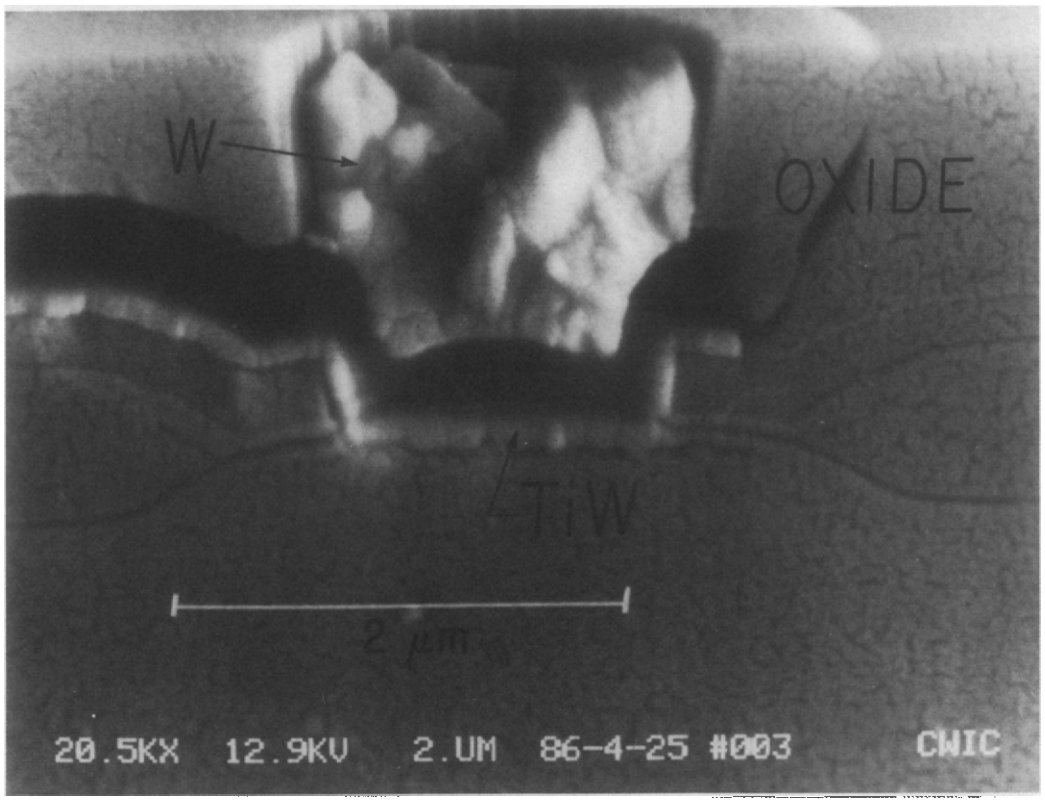


Fig.3.24. Stacked via contact between M2 and contact with active area diode region. The Al and M1 Mo metallization has been removed by selective etching leaving the TiW contact metal and W plug intact.

patterns. As mentioned, this produces vias of greatly differing depth if the interlevel dielectrics are planarized for improved step coverage and high-resolution lithography.

The layout and density of the array of gates is primarily determined by the gate array routing and metallization grid. The metallization grid spacing is usually determined by the metallization design rules. Reducing the metal pitch, therefore, has a direct relationship to gate array chip size. In fact, one relative merit of gate array design rules can be obtained by multiplying M1 and M2 pitches. Use of the W via plugs for minimizing the via size and eliminating some of the metal frame rules by the methods described previously can aid this reduction.

Another application of this new refractory metal technology is described here. Decreases in size of devices and circuits can also be accomplished by using these same methods and materials at the gate and contact levels. In this instance, the unframed metal contact landing pads (Figs. 3.9,3.25, and 3.26) enable the device and active area diode areas to be decreased in size.

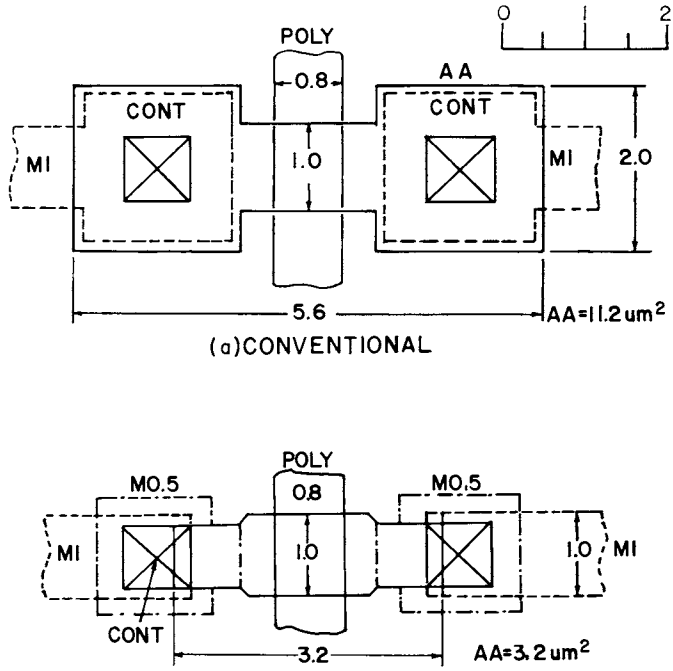


Fig. 3.25. Two types of device layouts: a. Conventional with bottom of contact framed within diode or device active area region and M1 framed around the top of contact hole; b. Use of M0.5 to contact selectively strapped diode areas (notice that diode area is smaller and does not frame M0.5) and M1 is utilized as an unframed contact to M0.5.

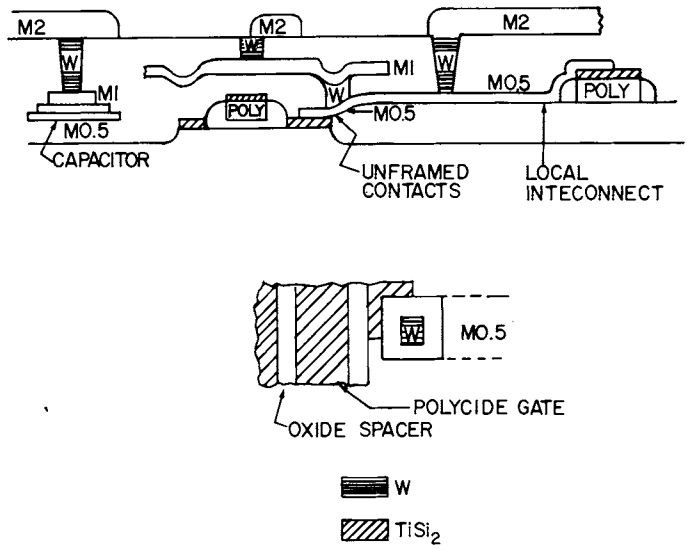


Fig. 3.26. Possible future multilevel metal configurations, with W via plugs between all levels. M0.5, the local interconnection metal, is an unframed contact to selectively metalized gates, sources, and drains, as shown. M1 and M2 are crossover interconnection levels. Capacitors for mixed analog/digital are also feasible, as shown at the left.

Obviously, the unframed contacting metal (Mo), labeled M0.5, must be etched selectively over SiO_2 and the active area region (usually strapped, for instance, with metal silicide, e.g., $TiSi_2$).[9] Figure 3.25 shows that the device area can be reduced with the diode areas being decreased almost fourfold without requiring a fundamental change in basic design rules (e.g., alignment tolerance). The use of this concept enables very small contact windows to be filled with selective CVD W since this process works favorably when selective CVD W depositions are made on another metal. Gate crossovers would be made utilizing M1 and M2 patterns, as shown in Fig. 3.26. Figure 3.26 also shows how M2 could make direct contact to M0.5, thus bypassing M1. This might be useful for new power busing layout. For instance, M2 to M0.5 contacting diffusion might be useful. This architecture is made possible by the W via plug method.

In a standard cell design approach, M0.5 could also be utilized as a local interconnect within the standard cell. Both of these new concepts (unframed contacts and local interconnections) result in reductions in cell size. For logic circuits that require a large number of M1/poly crossovers, the local interconnection is only occasionally used but the unframed contact reduces circuit size. An example is shown in Fig. 3.27.

[9] Table 3.6 identifies by number and function the different metallization layers. See also Fig. 3.26.

TABLE 3.6
Metal Terminology

M0.0: Contacting metal, contact gates, and silicon, may or may not be in the same pattern as the interconnection metal at the silicon contact level.
M0.5: Local interconnection done at the contact gate level before depositing interlevel dielectric (interconnects adjacent gate and diode, similar to buried contacts in NMOS). It might also be used as a capacitor base plate.
M1: Interconnects underlying device features and connects to M2 through interlevel dielectric vias (also can serve as base plate for M1/M1.5 capacitors or top plate if M0.5 is used as base plate).
M1.5: Top plate of M1/M1.5 capacitor; its size determines the value of the capacitor (used in 1.2-μm GE CMOS analog process).
M2: Interconnects devices at power bus level, interconnects M1.5 capacitor plates and makes contact to M1 through interlevel via. An Al alloy is typically used.
M0.0–M1.5: Can be refractory materials (e.g., $TiSi_2$, TiW, W, and Mo alone or in combination).

In any case (for gate arrays, "sea of gates" approach for gate arrays, custom design, or standard cells), the selective W contact plug between M0.5 and M1 enables the designer to unframe M1 over contact. This reduces M1 pitch since the framing rules at M1 are no longer required. Of course, this can be combined with the previously discussed use of unframed M2 over M1/M2 vias to also decrease M2 pitch. These methods can reduce chip size considerably.

Another area where multilevel metal techniques have been shown to be advantageous is in analog VLSI circuits. In this application, Mo at M1 is used as a high-precision, stable, comparator resistor ladder and, in addition, the base plate of very low voltage coefficient precision capacitors, wherein the top electrode also utilizes a thin Mo film [100] (Fig. 3.28). Another advantage of these metal-to-metal capacitors is their small loss angle or high Q. This is in contrast to, for instance, double polysilicon capacitors.

V. SUMMARY

- Gate electrodes (Mo, W, silicides) that have a higher work function than that of $n+$ polysilicon can reduce subthreshold leakage and alleviate other degredation problems associated with the doping levels required to make good n- and p-channel devices in CMOS circuits.
- Parasitic device resistance and capacitances were examined in detail. The injection resistance does not scale with size, and the contact length

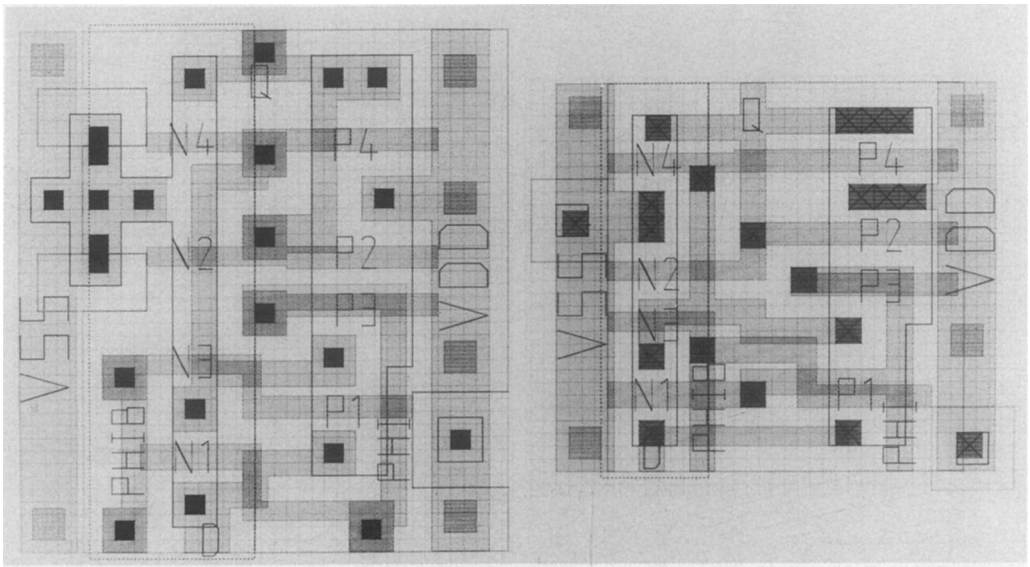


Fig. 3.27. Reduction in circuit size obtained by using M0.5 and unframed contacts (8-transistor CMOS latch circuit).

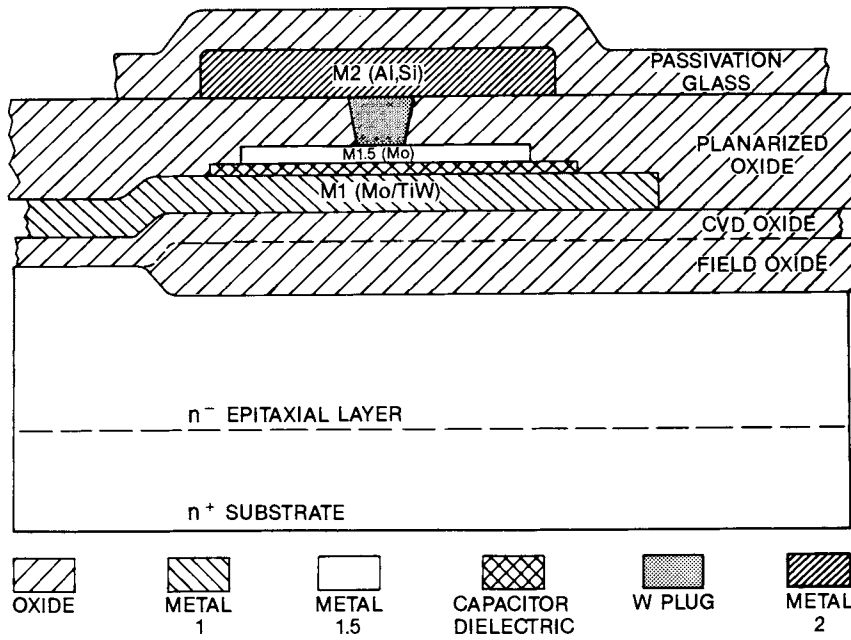


Fig. 3.28. Double-level metal capacitor utilizing M1 (Mo/TiW) as the base plate and a thinner refractory metal (Mo) as the top plate. The top plate's size determines the capacitance value with M2 connecting to and interconnecting capacitors. Such a structure has advantageous electrical properties. (From Brown *et al.* [100].)

must be long enough to avoid the increase in resistance that occurs when the contact is shorter than the current crowding region at the edge of the contact. An analysis indicated that changes in contact metallization are advantageous for reducing contact parasitic resistances. Refractory metal contacts and selective source/drain strapping metallization have advantages below 1 μm. Junction capacity is seen as the major parasitic component that limits the speed of logic gates. Methods of reducing junction capacity using new self-aligning contact metallization techniques and isolation methods were discussed. The present difficulties with selective metallization strapping of the entire source and drain areas were discussed in this context.

- Some of the changes in contact metallurgy being studied and utilized should produce higher reliability to shallow junctions by eliminating spiking and electromigration failure.
- Multilevel metallization is an effective way of decreasing chip size and enhancing circuit performance without shrinking active device dimensions that are currently limited by economical lithography methods

and hot electron effects. The new materials and methods being utilized for interconnections and multilevel metallization were discussed. Material properties were compared with conventional Al metallization systems. Planarization and other techniques can be used to give good step coverage over sharply defined fine metal lines; and the methods that are being explored to fill small interlevel metal vias were discussed.

Advances in "downsizing" and higher performances utilizing advanced metallization concepts will be more dependent upon innovative process and device research as contrasted to the previous two decades when scaling laws were a dependable means of predicting and accomplishing these gains. In the future, the interplay between process, device, and circuit designers must be closely interactive when designing a new and smaller integrated circuit product line.

ACKNOWLEDGMENTS

The authors are thankful to P. Frank, T. Nakagawa, and J. Richotte for their layouts of Figs. 3.10, 3.25, and 3.27. The authors would like to acknowledge the fine e-beam microphotography of J. Norton. The authors are indebted to Elizabeth Harris for her careful manuscript work. Finally, a thanks to all the workers in the field who wrote so many fine articles exhibiting their excellent work, much of which, but not all, was used in this article.

REFERENCES

1. D. M. Brown, W. E. Engeler, M. Garfinkel, and P. V. Gray, Self-registered molybdenum-gate MOSFET, *J. Electrochem. Soc.* **115,** 874–876 (1968). Also, Electrochem. Soc. Meeting, Chicago, Oct. 15–19, 1967, Abstract RNP-7.
2. D. M. Brown, The self-registered MOSFET, *Solid State Tech.* **15,** 33 (1972).
3. D. M. Brown, W. R. Cady, J. W. Sprague, and P. J. Salvagni, The *p*-channel refractory metal self-registered MOSFET, *IEEE Trans. Electron Devices* **ED-18,** 931–940 (1971).
4. M. Kim and D. M. Brown, Mo_2N/Mo gate MOSFETs, *IEEE Trans. Elec. Devices* **ED-30,** 598–599 (1983).
5. T. P. Chow and A. J. Steckl, Refractory metal silicides: thin-film properties and processing technology, *IEEE Trans. Electron Devices* **ED-30,** 1480–1497 (1983).
6. T. Mochizuke, T. Tsujimaru, M. Kashiwagi, and Y. Nishi, Film properties of $MoSi_2$ and their application to self-aligned $MoSi_2$ gate MOSFET, *IEEE Trans. Electron Devices* **ED-27,** 1431–1435 (1980).
7. D. M. Brown and P. V. Gray, MOS science and technology, I. Analytical model for the semiconductor surface channel and the MOS varactor and field effect transistor, General Electric Class 1 Report No. 67-C-026, March 1967. Also, see the appendix of Fast interface state measurements, *JECS* **115,** 760–766 (1968). These reports contain

an exact analytical solution for the MOS surface channel device which uses Fermi Dirac statistics instead of the commonly used Maxwell Boltzmann approximation. The use of Fermi Dirac statistics is especially important for the heavily doped semiconductor. These equations are also easy to understand.

8. L. C. Parrillo, Process and device considerations for micron and submicron CMOS technology, *IEDM Tech. Digest,* 398–402 (1985). See also K. M. Chan, D. W. Wendcur, J. Lin, C. K. Lau, and H.-S. Fu, Submicrometer thin gate oxide *p*-channel transistors with *p*+ polysilicon gates for VLSI applications, *IEEE Elec. Dev. Lett.* **EDL-7,** 49–50 (1986).
9. D. M. Brown, H_2-induced diffusion in MOS devices, *J. Electrochem. Soc.* **123,** 412–415 (1976).
10. E. Takeda, G. A. C. Jones, and H. Ahmed, Constraints on the application of 0.5 μm MOSFETs to VLSI systems, *IEEE Trans. Electron Devices* **ED-32,** 322–327 (1985).
11. V. Schwabe, F. Neppl, and E. P. Jacobs, $TaSi_2$ gate for VLSI CMOS circuits, *IEEE Trans. Electron Devices* **ED-31,** 988–992 (1984).
12. D. M. Brown and R. J. Connery, Mo gate tetrode, *IEEE Trans. Electron Devices* **ED-25,** 1302–1307 (1978).
13. T. P. Chow, M. Ghezzo, A. J. Steckl, and D. Brown, "Silicon Nitride Passivation for Short-Channel Molybdenum-Gate Devices," *ECS Meeting, Minneapolis, Minnesota, May 10–15, 1981, Abstract 296.*
14. N. Yamamoto, S. Iwata, N. Kobayashi, K. Yagi, and Y. Wada, Sub-micron tungsten gate process compatible with silicon gate process, *2nd International Symposium on VLSI Science and Technology, Cincinnati, Ohio, May 6–11, 1984,* pp. 1–13.
15. R. F. Kwasnick, E. B. Kaminsky, P. A. Frank, G. A. Franz, K. J. Polasko, R. J. Saia, and T. B. Gorczyca, An investigation of molybdenum gate for submicron CMOS *IEEE Trans. Electron Devices* **ED-35,** 1432–1438 (1988).
16. H. Oikawa and T. Amazawa, Highly reliable Mo gate and interconnection technology, *Proc. 3rd Int. VLSI Symp. ECS Meeting, Toronto, Ontario, May 12–17, 1985,* pp. 131–145.
17. D. M. Brown, W. R. Cady, J. W. Sprague, and P. J. Salvagni, The *p*-channel refractory metal self-registered MOSFET, *IEEE Trans. Electron Devices* **ED-18,** 931–940 (1971).
18. T. Shibata, K. Hieda, M. Sato, M. Konaka, R. Dange, and H. Iizuka, An optimally designed process for submicrometer MOSFETs, *IEEE Trans. Electron Devices* **ED-29,** 531 (1982).
19. M. E. Alperin, T. C. Hollaway, R. A. Haken, C. D. Gosmeyer, R. V. Karnaugh, and W. D. Parmantie, Development of the self-aligned titanium silicide process for VLSI applications, *IEEE Trans. Electron Devices* **ED-32,** 141–149 (1985).
20. W. A. Metz, N. J. Szluk, G. W. Miller, and H. O. Hayworth, Effect of selective tungsten as a polysilicon shunt on CMOS ring-oscillator performance, *IEEE Elec. Dev. Lett.* **EDL-6,** 372–374 (1985).
21. R. J. Bayruns, R. L. Johnson, D. L. Fraser, and S. C. Fang, Delay analysis of Si NMOS GBIT/S logic circuits, *IEEE J. Solid State Circuits* **SC-19,** 755–764 (1984).
22. H. Murrman and D. Widmann, Current crowding on metal contacts to planar devices, *IEEE Trans. Electron Devices* **ED-16,** 1022 (1969).
23. S. S. Cohen, G. Gildenblat, and D. M. Brown, Size effects on contact resistance and device scaling, *J. Electrochem. Soc.* **130,** 978-908 (1983).
24. J. M. Ford, Al/Si contact resistance for submicrometer design rules, *IEEE Trans. Electron Devices* **ED-32,** 840–842 (1985).

25. J. M. Pimbley, Two-dimensional current flow in the MOSFET source-drain, *IEEE Trans. Electron Devices* **ED-33,** 986-998 (1986).
26. K. K. Ng, R. J. Bayruns, and S. C. Fang, The spreading resistance of MOSFETs, *IEEE Elec. Dev. Lett.* **EDL-6,** 195-198 (1985).
27. A. S. Grove, "Physics and Technology of Semiconductor Devices," p. 329. Wiley, New York, 1967.
28. Y. El-Mansy, MOS device and technology constraints in VLSI, *IEEE J. Solid State Circuits,* **SC-17,** 197-203 (1982).
29. B. Hoeneisen and C. A. Mead, *IEEE Trans. Electron Devices* **ED-19,** 382-383 (1972).
30. M. J. Kim, D. M. Brown, S. S. Cohen, P. Piacente, and B. Gorowitz, Mo/TiW contact for VLSI applications, *Trans. IEEE Electron Devices* **ED-32,** 1328-1333 (1985).
31. J. G. J. Chern and W. G. Oldham, Determining specific contact resistivity from contact end resistance measurements, *IEEE Electron Device Letters* **EDL-5,** 178-180 (1984).
32. W. M. Loh, K. Saraswat, and R. W. Dutton, Analysis and scaling of kelvin resistors for extraction of specific contact resistivity, *IEEE Electron Device Letters* **EDL-6,** 105-108 (1985).
33. W. M. Loh, S. E. Swirhun, E. Crabbe, K. Saraswat, and R. M. Swanson, An accurate method to extract specific contact resistivity using cross-bridge kelvin resistors, *IEEE Electron Device Letters* **EDL-6,** 441-443 (1985).
34. J. Chern and W. G. Oldham, Reply to comments on determining specific contact resistivity from contact end resistive measurements, *IEEE Electron Device Letters* **EDL-5,** 349 (1984).
35. W. M. Loh, S. E. Swirhun, T. A. Schreger, R. M. Swanson, and K. C. Saraswat, 2-D simulations for accurate extraction of the specific contact resistivity from contact resistance data, *IEDM Technical Digest International Electron Devices Meeting,* pp. 586-589. Washington, D.C., Dec. 1-4, 1985.
36. S. S. Cohen, M. J. Kim, B. Gorowitz, R. Saia, T. F. McNelly, and G. Todd, Direct W-Ti contacts to silicon, *Appl. Phys. Lett.* **45,** 414 (1984).
37. J. M. Towner, Electromigration-induced short circuit failure, *23rd Proceedings Reliability Physics,* pp. 81-86 (1985).
38. H. Grinolds and G. Y. Robinson, Study of Al/Pd_2Si contacts on Si, *J. Vac. Sci. Technol.* **14,** 75-78 (1977).
39. D. C. Chen, P. Merchant, and J. Amano, Thermal stability of Al-Si/$TiSi_2$/Si Schottky diodes, *J. Vac. Sci. Technol.* **A3,** 709-713 (1985). Also, P. Merchant and J. Amano, Thermal stability of diffusion barriers for aluminum alloy/platinum silicide contacts, *J. Vac. Sci. Technol.* **A1,** 459-462 (1983).
40. H. Kaneko, M. Koyanagi, S. Shimizu, Y. Kubota, and S. Kishino, Novel submicron MOS devices by self-aligned nitridation of silicides, *Proc. IEDM,* pp. 208-211. Washington, D.C., December 1-4, 1985. Also, T. Maeda, S. Shima, T. Nakayama, M. Kakuma, K. Mori, S. Iwabuchi, R. Aoki, and J. Matsunaga, Highly reliable one-micron-rule interconnection utilizing TiN barrier metal, *IEDM Technical Digest International Electron Devices Meeting,* pp. 610-613. Washington, D.C., Dec. 1-4, 1985.
41. T. Hara, S. Enomoto, N. Ohtsuka, and S. Shima, Barrier effects of tungsten inter-layer for aluminum diffusion in aluminum/silicon ohmic-contact system, *Jpn. J. Appl. Phys.* **24**(7), 828-831 (1985).
42. G. B. Bronner and J. D. Plummer, Characterization of transient process phenomena using a temperature-tolerant metallurgy, *IEEE Elec. Dev. Lett.* **EDL-5,** 75-77 (1984).
43. J. P. Roland, N. E. Hendrickson, D. D. Kessler, D. E. Nory, and D. W. Quint, Two layer refractory metal IC process, *HP Journal,* August 1983.

44. S. S. Cohen, M. J. Kim, and D. M. Brown, Direct molybdenum contacts to silicon, *Appl. Phys. Lett.* **46**(7), 659 (1985).
45. S. S. Cohen, P. A. Piacente, and D. M. Brown, Thermal stability of platinum silicide contacts to silicon with molybdenum as final metallization, *Appl. Phys. Lett.* **41,** 976–978 (1982).
46. R. N. Singh, D. W. Skelly, and D. M. Brown, Palladium silicide ohmic contacts to shallow junctions in silicon, JECS **133,** 2390–2393 (1986).
47. M. J. Kim, D. W. Skelly, D. M. Brown, and J. F. Norton, Mo/Ti double layer contact for VLSI, *ECS Meeting, Las Vegas, Nevada, October 13–18, 1985, Abstract 289.* Also, M. J. Kim, D. W. Skelly, R. Saia, G. Smith, and D. M. Brown, *J. Electron Chem. Soc.* **134,** 2603–2606 (1987).
48. J. M. Shaw and J. A. Amick, Vapor-deposited tungsten as metallization and interconnection materials for silicon devices, *RCA Rev.* **31,** 306–316 (1970); also, *ECS, Detroit, Michigan, RNP231,* October 1969.
49. K. C. Saraswat, S. Swirhun, and J. P. McVittie, Selective CVD of tungsten for VLSI technology, Electrochem. Soc. Meeting, Cincinnati, Ohio, May 1984. Also, S. Swirhun, K. C. Saraswat, and R. M. Swanson, Contact resistance of LPCVD W/Al and PtSi/W/Al metallization, *IEEE Elec. Dev. Lett.* **EDL-5**(6), 209–211 (1984). Also, W. T. Stacy, E. K. Broadbent, and M. H. Norcott, Interfacial structure of tungsten layers formed by selective low pressure chemical vapor deposition, *J. Electrochem. Soc.* **132,** 444–448 (1985).
50. D. M. Brown, B. Gorowitz, R. Wilson, and R. Saia, Unframed contacts using refractory metals, *IEEE Elec. Dev. Lett.* **EDL-6,** 408–409 (1985).
51. D. C. Chen, T. R. Cass, J. E. Turner, P. Merchant, and K. Y. Chiu, The impact of $TiSi_2$ on shallow junctions, *IEDM Technical Digest International Electron Devices Meeting,* pp. 411–414. Washington, D.C., Dec. 1–4, 1985.
52. T. Yamaguchi, S. Morimoto, G. H. Kawamoto, and J. C. DeLacy, Process and device performance of 1 μ-channel n-well CMOS technology, *IEEE Trans. Electron Devices* **ED-31,** 205–214 (1984). Also, Y. Yamaguchi, S. Morimoto, H. K. Park, and G. C. Eiden, Process and device performance of submicrometer-channel CMOS devices using deep trench isolation and self-aligned $TiSi_2$ technologies, *IEEE Trans. Electron Devices* **ED-32,** 184–193 (1985).
53. D. M. Brown, M. Ghezzo, and J. M. Pimbley, Trends in advanced process technology —submicrometer CMOS device design and process requirements, *Proc. IEEE* **74**(12), 1678–1702 (1986).
54. T. Tang, C.-C. Wei, R. Haken, T. Holloway, C.-F. Wan, and M. Douglas, VLSI local interconnect level using titanium nitride, *IEEE IEDM Technical Digest, International Electron Devices Meeting,* pp. 590–593. Washington, D.C., Dec. 1–4, 1985.
55. D. C. Chen, S. S. Wong, P. V. Voorde, P. Merchant, T. R. Cass, J. Amano, and K-Y. Chiu, A new device interconnection scheme for sub-micron VLSI, *IEDM Tech. Digest,* pp. 118–121. San Francisco, California, December 9–12, 1984.
56. W. E. Engeler and D. M. Brown, Performance of refractory metal multilevel interconnection system, *IEEE Trans.* **ED-19,** 54–61 (1972).
57. S. Asai, Device and material requirements in very large scale integrated circuits, *Proc. 32nd Annual Symp. American Vac. Soc., Nov. 19–22,* 1985.
58. H. B. Bakoglu and J. D. Meindl, Optimal interconnection circuits for VLSI, *IEEE Trans. Electron Devices* **ED-32,** 903–909 (1985).
59. J. Curry, G. Fitzgibbon, Y. Guan, R. Muollo, G. Nelson, and A. Thomas, New failure mechanisms in sputtered aluminum-silicon films, *Int. Rel. Phys. Sym.,* pp. 6–8 (1984).

60. S. S. O'Donnel, J. W. Barling, and G. Hill, Silicon inclusions in aluminum interconnects, *Int. Rel. Phys. Sym.*, pp. 9–16 (1984).
61. R. W. Thomas, private communication.
62. R. W. Thomas and D. W. Calbrese, Phenomenological observation on electromigration, *1983 21st Annual Proc. Int. Rel. Physics Symposium,* p. 1. Phoenix, Arizona, April 5–7, 1983.
63. I. A. Blech, Electromigration in thin aluminum films on titanium nitride, *J. Appl. Phys.* **47,** 1203 (1976).
64. H. H. Hoang, Effects of annealing temperature on electromigration performance of multilayer metallization systems, *Int. Rel. Phys. Symp.*, pp. 173–178 (1988).
65. F. Fischer and F. Neppl, Sputtered Ti doped Al-Si for enhanced interconnect reliability, *22nd Annual Proceedings Reliability Physics,* pp. 190–198. Las Vegas, Nevada, April 3–5, 1984.
66. D. S. Gardner, T. L. Michalka, P. A. Flinn, T. W. Barbee, K. C. Saraswat, and J. D. Meindl, Homogeneous and layered films of aluminum/silicon with titanium for multilevel interconnects, *1985 Proc. 2nd International IEEE VLSI Multilevel Interconnection Conf.*, pp. 102–113. Santa Clara, California, June, 1985.
67. J. K. Howard, J. F. White, and P. S. Ho, Intermetallic compounds of Al and transition metals: effect of electromigration in 1–2 μm wide lines, *J. Appl. Phys.* **49,** 4083 (1978).
68. D. S. Gardner, T. L. Michalka, K. C. Saraswat, T. W. Barbee, J. P. McVittie, and J. D. Meindl, Layered and homogeneous films of aluminum and aluminum/silicon with titanium and tungsten for multilevel interconnects, *IEEE Trans. Electron Devices* **ED-32,** 174–183 (1985).
69. B. W. Shen, T. Bonifield, and J. McPherson, An evaluation of titanium Interlayered aluminum films for VLSI metallization, *1985 Proc. 2nd International IEEE VLSI Multilevel Interconnection Conf.* pp. 114–120. Santa Clara, California, June, 1985.
70. R. E. Jones and L. D. Smith, Contact spiking and electromigration passivation cracking observed for titanium layered aluminum metallization, *Proc. 2nd Int. IEEE VLSI Multilevel Interconnection Conf.*, pp. 194–200. Santa Clara, California, June, 1985.
71. E. Philofsky, K. Ravi, E. Hall, and J. Black, *9th International Reliability Physics Symposium,* p. 120. Las Vegas, Nevada, March 31–April 2, 1971.
72. K. E. Gsteiger, Current status and problems relating to dielectrics and conducting films on VLSI Surfaces, *SRC Topical Res. Conf.* (VLSI Interface Engineering), Chicago, Illinois, February 16–17, 1984.
73. M. T. Yin, Layout Verification of VLSI Designs, *VLSI Design,* pp. 30–38, July 1985.
74. R. W. Thomas, D. Calabrese, B. Vastag, and D. Roberts, Analysis of VLSI metallization microstructure by high resolution mechanical cross section and auger analysis, *Proc. Int. Sym. for Testing and Failure Analysis,* p. 98 (1985).
75. P. B. Ghate, *Int. Rel. Phys. Sym. Tutorial,* Orlando, Florida (1985).
76. J. M. Mikkelson, L. A. Hall, A. K. Mahotra, S. P. Seccombe, and M. S. Wilson, An NMOS VLSI process for fabrication of a 32-bit CPU chip, *IEEE Journal of Solid State Circuits* **SC-16,** 542–547 (1981).
77. D. W. Woodruff, R. H. Wilson, and R. A. Sanchez-Martinez, Adhesion of nonselective CVD tungsten to silicon dioxide, Tungsten and Other Refractory Metals for VLSI Applications. Albuquerque, New Mexico, October 7–9, 1985. Materials Research Society, Pittsburgh, Pennsylvania, 1986, p. 182.
78. D. S. Gardner, J. D. Meindl, and K. C. Saraswat, Interconnection and electromigration scaling theory, *IEEE Trans. Electron Devices* **ED-34,** 633–643 (1987).

79. J. T. Watt and J. D. Plummer, Effect of interconnection delay on liquid nitrogen temperature CMOS circuit performance, *Int. Elec. Dev. Meeting,* Washington, D.C., *Tech. Digest,* pp. 393–396 (1987).
80. O. K. Kwon, B. W. Langley, R. F. W. Pease, and M. R. Beasley, Superconductors as very high system level interconnects, *IEEE Elec. Dev. Lett.* **EDL-8,** 582–585 (1987).
81. D. K. Ferry, Interconnection length and VLSI, *IEEE Circuits and Device Magazine,* July 1985, pp. 39–42.
82. D. Barton and C. Maze, A two level metal CMOS process for VLSI circuits, *Semiconductor International, January 1985,* pp. 98–102. Also, *IEEE VLSI Multilevel Interconnections Conference Proceeding,* p. 268 (1984).
83. R. Saia, B. Gorowitz, D. Woodruff, and D. Brown, Plasma etching methods for the formation of tungsten plugs used in multilevel VLSI, *J. Electrochem. Soc.* **135**(4), 936–940 (1985).
84. A. C. Adams and C. D. Capio, Planarization of phosphorus-doped silicon dioxide, *J. Electrochem. Soc.* **126,** 423–429 (1979).
85. R. H. Wilson and P. A. Piacente, Effect of planarization on VLSI processing, *Semiconductor International,* April 1986, pp. 116–121. Also, R. H. Wilson and P. A. Piacente, Effect of circuit structure on planarization resist thickness, *J. Electrochem. Soc.* **133**(5), 981–984 (1986).
86. N. G. Einspruch and D. M. Brown, eds., "VLSI Electronics," Vol. 8, p. 326. Academic Press, San Diego, (1984).
87. M. Morimoto, T. Mogami, H. Okabayashi, and E. Nagasawa, SiO_2 planarization by two step RF bias sputtering, *1983 Symposium VLSI Tech.,* pp. 100–101.
88. L. Koyama and M. Thomas, Metal step coverage improvement in double level metal process using oxide spacers, *Second International VLSI Multilevel Interconnection Conference, Santa Clara, California, June 25–26, 1985.*
89. R. H. Wilson, R. W. Stoll, and M. A. Calacone, Highly selective, high rate W metal process deposition, *Second International VLSI Multilevel Interconnection Conference, Santa Clara, California, June 25–26, 1985.*
90. T. Mogami, H. Okabayashi, E. Nagasawa, and M. Morimoto, Planarized molybdenum interconnections using via hole filling by bias sputtering, *Second International VLSI Multilevel Interconnection Conference, Santa Clara, California, June 25–26, 1985.*
91. Y. Homma and S. Tsunekawa, Planar deposition of aluminum by RF/DC sputtering with RF bias, *JECS* **132,** 1466–1472 (1985).
92. J. G. Ehern, W. G. Oldham, and N. Cheung, Contact electromigration induced leakage failure in aluminum-silicon to silicon contacts, *IEEE Trans.* **ED-32,** 1341 (1985).
93. T. D. Bonifield and S. M. McDavid, Development trends in VLSI interconnects and metallization, *SRC Topical Research Conference Interconnections and Contacts, Madison, Wisconsin,* October 30, 1984.
94. R. E. Oakley, S. J. Rhodes, E. Armstrong, and A. Marsh, Pillars—the way to two micron pitch multilevel metallization, *Proceedings Second International IEEE VLSI Multilevel Interconnection Conference. New Orleans, Louisiana, June 21–22, 1984,* pp. 23–29.
95. T. Moriya, S. Shima, Y. Hazuki, M. Chiba, and M. Kashawagi, A planar metallization process—its application to tri-level aluminum interconnection, *IEDM Tech. Digest,* pp. 550–553, Washington, D.C., Dec. 5–7, 1983.
96. R. H. Wilson, B. Gorowitz, H. G. Williams, R. Chow, and S. Kang, Achieving low contact resistance to aluminum with selective tungsten deposition, *J. Electrochem. Soc.* **134,** 1867 (1987).

97. R. N. Singh, D. M. Brown, M. J. Kim, and G. A. Smith, Study of molybdenum-aluminum interdiffusion kinetics and contact resistance for VLSI applications, *J. Appl. Phys.* **58**(12), 4598–4604 (1985).
98. R. N. Hall, D. M. Brown, R. H. Wilson, and D. W. Skelly, Electromigration reliability studies of intermetal contacts having CVD tungsten via plugs, "Tungsten and Other Refractory Metals for VLSI Applications," Vol. III, (V. A. Wells, ed.), Materials Res. Soc., Pittsburgh, Pennsylvania, 1988. pp. 231–237.
99. D. M. Brown, B. Gorowitz, P. A. Piacente, R. Saia, R. H. Wilson, and D. Woodruff, Selective CVD tungsten via plugs for multilevel metallization, *Elec. Dev. Lett.* **EDL-8**, 55–57 (1987). Also, IEEE International Electron Device Meeting (IEDM), Los Angeles, California, December 1986, Technical Digest, pp. 66–69.
100. D. Brown, S. Chu, M. Kim, B. Gorowitz, M. Milkovic, T. Nakagawa, and T. Vogelsong, Advanced analog CMOS technology, *IEDM Tech. Digest,* pp. 260–263. Washington, D.C., Dec. 1–4, 1985.
101. J. M. Pimbley and D. M. Brown, "Current crowding in high density VLSI metallization structures, *IEEE Trans. Electron Devices* **ED-33**(9), 1399–1401 (1986).

Chapter 4

Isolation Techniques

I. INTRODUCTION

The origin of integrated circuits is rooted in the desire of fabricating on an active substrate as many devices as possible while electrically connecting them only according to circuit requirements. Therefore, though the devices share the same silicon chip and are physically located near each other, they must operate independently, as if they were discrete components. For this purpose, special fabrication processes, which are known as isolation techniques, have been developed and are judged by their ability to minimize the spacing required for electrical isolation of adjacent devices. Within the constraints of the design rules, smaller spacings result in reduced chip area, and consequently higher die yield and increased number of chips per wafer, thereby reducing the cost per circuit function. Thus, there is a great economical incentive toward improving the isolation techniques in conjunction with scaling, which is further enhanced by the related circuit performance improvement due to shorter interconnections and reduced gate propagation delay.

In the past, the design rules and feature sizes for a given technology were mainly determined by lithographic resolution, overlay accuracy, and pattern transfer capabilities. On the other hand, device isolation was not an issue, since field oxide encroachment, lateral diffusion, and the extent of the space charge regions were small compared to the distance separating the individual devices. However, as the minimum feature size fell below 2 μm, these factors became important because of their inability to scale with the other dimensions. This development initiated a search for new isolation processes, which is still proceeding unabated.

In CMOS technology, the isolation system is more complicated than in either its NMOS or PMOS counterparts, because the well requires a separate isolation in addition to the active area. Although there are cases where similar isolation techniques can be used for both requirements, for example, deep trenches, normally the techniques are different. For instance, relative to the active area, the preferred approach might be some form of dielectric isolation, while for the well, the modern solution could be a twin-tub approach with retrograde wells separated by deep trenches.

We will consider the active area first. In modern processes, LOCOS (local oxidation of silicon) has been extensively used since the early 1970s [1]. Its major advantages are simple fabrication, partially recessed field oxide, self-aligned field implant, and accurate active area definition. However, LOCOS is afflicted by significant field-oxide encroachment along the borders of the active regions, which is commonly known as "bird's beak," because of its characteristic profile [2]. Since the bird's beak width usually ranges from 0.5 to 1.0 μm per side, the underlying area represents a significant "overhead" in wafer surface utilization and a stumbling block toward achieving higher packing density. The situation will worsen with increased downscaling, since the active area will be more finely subdivided and the perimeter will grow accordingly. New isolation techniques are thus required for submicron CMOS technologies to reduce or eliminate the bird's beak completely, even at the cost of a radical departure from LOCOS.

As suggested by Holton and Cavin [3], the new active area isolation techniques can be classified into four major groups: (1) advanced LOCOS, (2) trench isolation, (3) selective epitaxial growth, and (4) full dielectric isolation. Though we defer a detailed discussion to the following sections, we point out that these techniques are listed in ascending order of fabrication complexity, and consequently, of decreasing popularity for use in advanced CMOS processes. Nevertheless, there are cases where the selection of the isolation technology is dictated by the application, setting aside any consideration of cost and fabrication complexity to a secondary issue. For example, a well-known case is provided by radiation-hardened circuits, where stringent system specifications can be met only by full dielectric isolation [4,5].

A few years ago integrated circuit (IC) process development groups were very optimistic about their ability to completely eliminate the bird's beak with advanced LOCOS or trench isolation methodologies [6]. However, despite great efforts and impressive evidence demonstrating a lack of encroachment, the new techniques were plagued by a vast array of problems, such as a high junction leakage current in silicon and a high density of gate oxide defects. These conditions were traced back to an increase of stresses

in silicon at the active area edges during field oxidation and their subsequent relief through the generation of lattice defects [7]. As a result, LOCOS is still widely used for 1.25-μm CMOS processes, and there are strong indications that with some modifications it will be also used for 0.8-μm technologies.

In dielectric isolation, there are many opportunities for synergistic action between various technological developments. For instance, the development of trenches for DRAM capacitors may lead to an early utilization of trenches for isolation in generic CMOS/BULK processes. BULK process refers to fabrication of an integrated circuit on silicon wafers as distinguished from silicon on insulator (SOI), where a thin silicon layer is mechanically supported by an insulating substrate. Similarly, radiation-hardened field oxides, which are often deposited [8], may offer an alternative to thermal growth of the field oxide, hopefully avoiding edge defects and lateral encroachment.

Another important aspect of device isolation is the surface doping under the field oxide and at the active area edges, which has a strong influence on both the field threshold voltage and the punch-through voltage between adjacent diffusions. Though the importance of the field threshold voltage is well recognized, only lately the punch-through voltage has become a concern, because of the close proximity of active devices in advanced CMOS processes, particularly in the submicron regime [9]. Fortunately, both field threshold and punch-through voltages can be adjusted with ion implantation by increasing the surface doping near the field-oxide–silicon interface. If high-energy MeV implantation is available, this adjustment may be accomplished after field oxidation to avoid dopant segregation effects and related changes of the doping profile.

As stated earlier, CMOS also has the unique problem of well isolation. In the past, deep wells were formed at high temperature by diffusion, causing large lateral spreads, which had to be accounted for in the design rules. Moreover, the parasitic thyristor action across the well, commonly called "latch-up," is triggered more easily if the well edge is near active devices, the well or the substrate are lightly doped, or the ground plane is remote from the surface [10]. These observations have spurred the use in advanced CMOS of epitaxial wafers, "retrograde" wells, multiple wells, and trench isolation along the well edges. Often, only one or two of these techniques are used in a process, because using more would be redundant and excessively expensive.

Within the realm of isolation we need to include a discussion of SOI technology [4,5]. Although a special case of SOI, represented by SOS (silicon on sapphire), has existed for over a decade, renewed interest in SOI has been revived by the emergence of 3-D IC technology [11]. Potential

benefits of using SiO_2 rather than sapphire as the insulating layer are a lower cost of starting material, elimination of back-channel leakage, and the ability to stack multiple active devices at various levels. However, this type of SOI technology has yet to reach the production floor, mainly because of low yield. On the other hand, SOS CMOS is commercially available and is used for special applications, such as radiation-hardened circuits.

Using an insulating substrate many isolation problems vanish, especially in CMOS. Foremost, latch-up and field inversion leakage are intrinsically eliminated. However, source to drain leakage at the channel edges may still be a problem and must be suppressed with adequate doping. The worst problem in SOI is the quality of the crystalline silicon layer, which lags considerably with respect to bulk silicon. This results in mobility degradation and thin oxide defects, thereby decreasing performance and yield, respectively. Nevertheless, three-dimensional prototype ICs have been fabricated with this technology, which appears promising for future fabrication of very compact chips [12].

The following sections will first review the progress in dielectric isolation between active areas, including a discussion of methods for bird's beak reduction and trench formation. The well isolation will be considered next, elucidating its dependence from the method of well formation, such as twin-tub, retrograde, or trench-isolated approaches. Latch-up immunity and leakage suppression will be discussed in this context. Finally, new developments in SOI technology will be described, showing current achievements and potential uses for the future.

II. DEVICE ISOLATION

A. Features, Issues, and Applications

The trend toward increased packing density and higher resolution, which is driven by cost reduction and performance improvement, has spawned a variety of new isolation techniques. For making an appropriate choice in terms of specific applications, it is important to review first the issues of CMOS device isolation by analyzing the requirements and examining the available solutions. Because of the important role played by the choice of substrate, that is, bulk or insulator, these two cases will be discussed separately, starting with bulk.

A cross section of a typical CMOS/BULK circuit is shown in Fig. 4.1 to illustrate the isolation requirements [13]. Since active devices share either

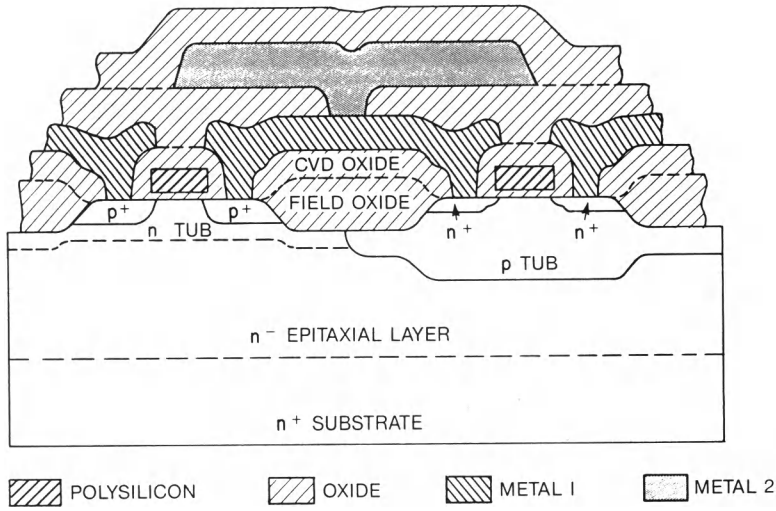


Fig. 4.1. Cross section of CMOS inverter. (After Brown *et al.* [13]. Copyright 1986 by the IEEE.)

the substrate or the well, the field-oxide isolation needs to be supplemented by junction isolation through the presence of space charge regions. As spacings between devices are reduced, the small back-to-back diode leakage may increase because of lateral punch-through [9]. In addition, inter-device leakage may be caused by powered-up interconnection lines crossing the isolation region, which may invert the silicon surface underneath if the field oxide is too thin or the silicon is too lightly doped. Both effects can be jointly modeled with a short-channel MOS field transistor, whose channel current, representing the isolation leakage, depends on the 2-D characteristics of the isolation structure [9]. A major reason for this complex interaction is the adoption of constant voltage scaling, which results in higher electric field over reduced size structures.

A simple way to offset the negative impact of scaling on electrical isolation is to increase the doping level of the channel-stop, which surrounds the active area, in order to suppress leakage currents due to either punch-through or field-induced surface conduction. This solution is a natural extension of the definition of channel-stop, since its higher doping is supposed to form a barrier against leakage. A self-aligned channel-stop is required to avoid extra masking steps and to ensure perfect registration with the active area edges. Despite this precaution, lateral diffusion takes place from the channel-stop into the active area causing the following device degradation effects. First, the channel width is reduced by lateral diffusion encroachment, resulting in smaller transconductance and other

narrow-width effects, such as an increase of threshold voltage. Second, the capacitance of source/drain regions and diffused interconnections is increased because of a narrower space charge region at the perimeter, which also leads to lower junction breakdown voltage. This effect is particularly prominent in submicron devices operating at 5-V supply, because the space charge region width is comparable to the minimum feature size, henceforth increasing the ratio of fringe capacitance to total capacitance. To avoid these drawbacks, the logical solution is to form the channel-stop and the active area on different planes, in order to minimize their mutual interference. Structures using this approach include either a recessed field oxide or deep trenches.

In conventional CMOS with diffused wells, the channel-stops are formed after well drive-in and active area patterning to be able to mask the active area with an oxide/nitride stack during the channel-stop implants, in good agreement with the LOCOS sequence [1]. This method achieves self-alignment in regard to the active area and also ensures effective masking due to the low energy of these implants. However, during subsequent field oxidation, the original doping profile may change substantially because of dopant segregation at the $Si-SiO_2$ interface and thermal redistribution. This problem is particularly acute for boron, because this element segregates preferentially into SiO_2, leaving a depleted region near the interface. This explains why many CMOS technologies exhibit a bare minimum specification for the field threshold voltage of NMOS devices. A similar effect is caused by thermal redistribution, but in this case, the responsibility for lowering the dopant peak near the interface rests with dopant diffusion into silicon. The result is nevertheless the same and the effectiveness of the channel-stop is reduced.

One way to overcome these problems is to form the isolation oxide by high-pressure oxidation or by low-temperature deposition. High-pressure oxidation reduces the thermal cycle and the associated dopant redistribution, with additional benefits of lower defect density and reduced wafer warpage [14]. However, dopant segregation still persists. On the other hand, using deposited oxides, neither segregation nor diffusion can occur at low temperature. But, unlike oxidation, deposition is not selective, and therefore it must be followed by planarization or other techniques to remove the isolation oxide from the active area. Recent progress in planarization techniques have made this method more appealing, as demonstrated for instance by oxide refill of silicon trenches [15].

Another way to retain the integrity of the original channel-stop profile is to form it after field oxidation with a high-energy (MeV) implant, or, even better, to use the retrograde well approach, because it provides an automatic channel-stop during well formation without additional masking

[16]. This occurs because the isolation region is covered by field oxide, so that, by adjusting the range of the high-energy implant to match the oxide thickness, the impurity concentration peak is formed near the $Si-SiO_2$ interface, as highly desirable for the channel-stop. On the other hand, since in a retrograde process the active area is bare or covered by a thin screen oxide, the high-energy implant penetrates deeply into silicon over the MOS channel with only a negligible change of the surface concentration, leaving its final adjustment to the following threshold control implants.

The suppression of interdevice isolation leakage is just one aspect of the isolation requirements. Equally important is the capacitive decoupling between the interconnection lines and the silicon substrate. This is critical for DRAM operation because of the need of maximizing the ratio of storage capacitance to parasitic capacitance, but it is also beneficial for ICs with long interconnections to reduce the propagation delay. For this reason, technologists have been reluctant to scale the field-oxide thickness proportionally to the linear dimensions. Eventually, this started to affect the design rules, because, with thicker field oxide, the thick–thin transition region along the isolation borders was also wider than normal. Hence, new isolation techniques were required, featuring steeper silicon sidewalls to reduce the width of the transition region without limiting the field-oxide thickness. These arguments apply in particular to the widely used LOCOS isolation [1], where the problems of scaling the bird's beak have prevented the use of this technology in the submicron regime. Instead, modified LOCOS techniques have been developed to reduce the bird's beak and steepen the silicon sidewalls, as discussed later.

The changes made to the original LOCOS process have created another class of problems represented by silicon defects along the active area perimeter. The origin of these defects is attributed to stresses during field oxidation, which are relieved by forming edge-type and screw-type dislocations [17]. When contaminated by heavy metals, these defects increase the diode reverse current by many orders of magnitude, leading to parametric test failure. On the other hand, in LOCOS, these defects are not formed because of the presence of a buffer oxide layer between the nitride mask and the active area silicon. However, this buffer oxide is also the conduit of oxygen under the mask edges, causing lateral oxidation and the bird's beak. For this reason, modified LOCOS processes either seal the edges of the buffer oxide with nitride (SWAMI [18,19], see Section II.B.2) or block the transport of O_2 to the interface with a thin nitride layer (SILO [20], see Section II.B.3). Consequently, the bird's beak is reduced, but the field oxide is butted against steep silicon sidewalls, which are subject to heavy strain during oxidation because of volumetric expansion, since the ratio between produced oxide and consumed silicon is about 2.2 by volume.

Hence, the challenge is to develop methods with carefully adjusted parameters, particularly thickness of various layers, in order to obtain the desired bird's beak reduction without causing edge defects. Empirically, it was found from a correlation study of sidewall steepness and edge-defect density that the oxidation angle should not exceed 75° [21].

Isolation can produce defects in oxide in addition to silicon. Those at the active area perimeter are more dangerous than those under the field, because this region is covered by thin oxide. The most common manifestation of these defects is through low-voltage gate breakdown of transistors and MOS capacitors with a border overlap. A well-known example is the Kooi effect [22], which occurs in LOCOS because of silicon nitridation under the active area nitride mask and consequent partial inhibition of oxide growth. The corresponding thermal oxide is therefore thinner and more susceptible to early breakdown. Similarly, there is a strong correlation between silicon edge defects and oxide defects, which is understandable since stacking faults and dislocations locally undermine the perfection of the oxide layer.

Another important requirement is surface planarity, with the maximum step height allowed being dependent on the minimum feature size of the process. One reason for this need is that optical exposure systems normally increase resolution at the expense of depth of field through the use of a lens with larger numerical aperture. Therefore, in submicron processes, the depth of field is reduced to a few thousand angstroms, which is the upper limit for the height difference between active area and field regions to accurately pattern the overlaying gate conductor. Otherwise, this pattern would not be simultaneously focused on both regions, resulting in linewidth variations and yield losses. In addition, large step heights can be responsible for another type of failure, which is represented by the formation of conductive "filaments" on the oxide sidewalls. These result from the combination of conformal deposition of a conductor film, usually polysilicon for the gate electrode, and its subsequent patterning with RIE (reactive ion etching), which is well-known for its anisotropy. Consequently, while the film covers uniformly the underlying pattern, including the sidewalls, it etches faster over the flat regions than over the sidewalls, where the etching path is longer because of the oblique intersection of this film with the vertical etching direction. Henceforth, the thicker the film and the higher the sidewall step, the wider will be the filament residue. Presently, a short isotropic etch is used after RIE to eliminate this filament, but the attendant lateral undercut of the final pattern can impair the gate electrode linewidth control with disastrous consequences for minimum size MOS, such as increased source/drain leakage and reduced hot electron lifetime. Thus, for further scaling, the only viable solution is surface planarity at the end of isolation.

The trends toward surface planarity, a relatively thick field insulator, steep sidewalls, and insulator deposition instead of growth have steered advanced isolation efforts in the direction of an inlay-type structure. This could be implemented either by reactively ion etching (RIE) trenches in silicon and refilling them with oxide, for example, BOX [15] (see Section II.B.5), or depositing the thick oxide first, reactively ion etching it, and refilling epitaxially the opening with silicon, for example, SEG [23] (see Section II.B.7). Variants of these methods are probably going to be used in CMOS/BULK technology at resolution below 0.5 μm.

Although trench isolation appears to be the way of the future, it has been used until now more frequently for well isolation [24] than for active area isolation [6], because of sidewall leakage problems with the NMOS device. These are partially caused by the difficulty of forming a channel-stop on the trench sidewall surface with ion implantation because of (1) the small angle of incidence required for penetrating into the trench, (2) the surface doping reduction due to the cosine law, and (3) the shadowing effect of the tilt of the implanter's beam. Without this indepth channel-stop, the fixed charge in the oxide may induce surface inversion in NMOS devices, since the polarity of this charge is always positive. Moreover, on $\langle 100 \rangle$ wafers, which are normally used for MOS ICs, the orientation of the trench sidewalls is close to $\langle 111 \rangle$ and hence exhibits a large built-in surface charge [25].

In addition, subthreshold leakage can be caused by nonuniform oxide etch back during planarization, which may leave exposed some of the trench upper edges and create a 90° bent in the MOS channel surface. The attendant crowding of the equipotential lines and higher electric field will lower the local NMOS threshold voltage and induce subthreshold leakage [26].

The expanding application of CMOS ICs in space and military systems has toughened the environmental specifications, particularly radiation hardness and temperature range. We will consider the effects on device isolation of radiation exposure and review means of improving it for enhanced radiation hardness. Of the various effects of radiation on CMOS ICs, the one that affects active area isolation most is the generation of positive charge in the field oxide. Gamma rays and other high-energy beams are responsible for the generation of these charges by creating hole–electron pairs in the oxide. While electrons leak through the oxide to adjoining electrodes, holes remain trapped in the oxide and form positive charges. The thicker the oxide, the larger is the total charge per unit area, because the probability of hole–electron pair formation is volume related. It also follows that the charge is proportional to the total dose of radiation absorbed per unit area, so that this mechanism is known as the *total dose effect*. The electrical properties most affected by the oxide charge are the

threshold voltage and the subthreshold characteristics of both the gate and field transistors [27,28]. It turns out that the field threshold shift is much larger than the gate threshold shift, not only in absolute numbers but also as a percentage of initial value. The reason is that the threshold shift increases as a function of oxide thickness at a faster rate than the threshold itself, because its dependence from thickness is a square (or cubic) law in contrast to a direct proportionality for the threshold voltage.

Since the oxide charge is positive, the threshold shift is negative, decreasing the NMOS threshold and increasing the PMOS threshold (in absolute value). Thus, the total dose effect does not endanger the PMOS isolation but may lead to failure of the NMOS isolation if the field threshold becomes so low that surface inversion is possible. Eventually, this will occur at some value of the total dose, but the challenge lies in raising this level higher and higher.

Accordingly, any hardness improvement for isolation must include either a reduction of the oxide charge generation rate or of the effect of the charge on the silicon underneath. The option of increasing the initial field threshold by using thicker field oxide is not advised because of the rapid increase of threshold shift with oxide thickness, as explained previously. Instead, an increase of channel-stop doping is suggested, because the corresponding increase of field threshold does not affect the threshold shift while providing extra margin in anticipation of this shift. This is why a CMOS retrograde *p*-well technology is suitable for radiation hardness, because the channel-stop for NMOS isolation is done after field oxidation, and thus retains a high boron concentration near the $Si-SiO_2$ interface.

For further improvements, the formation of the field oxide must be optimized to reduce the generation rate of oxide charge during radiation. Simple changes of the field oxidation cycle are helpful in this respect, but for high total-dose hardness, good results have been reported with a deposited PSG (phospho-silicate glass) layer over a thin, thermally grown SiO_2 film [8]. The current understanding is that in PSG the hole–electron recombination rate is higher than in SiO_2, partially compensating the radiation induced generation rate and therefore improving hardness [8]. Fortunately, deposited oxides are compatible with trench isolation, allowing to forecast higher total-dose tolerance for submicron CMOS isolation without major deviations from the baseline process.

B. Field Isolation Methods

1. LOCOS

LOCOS was invented in 1970 [1] and soon after became widely used in advanced integrated circuits. The gentle slope between thin and thick oxide

improved the step coverage of the gate electrode increasing yield and reliability. Moreover, at parity of field-oxide thickness, this step was only half of the amount found in isolation methods, which are based on field-oxide etch cut. Patterning was also more accurate, because only a thin nitride film had to be etched to define the active area. This contrasts with the deep field-oxide etch previously required, where edge control was difficult because of oxide undercutting.

As expected, LOCOS presented its own problems, but the major one, nicknamed bird's beak, started to hurt only at feature sizes below 2 μm. In an effort to extend the use of LOCOS to 1-μm technology, modified LOCOS techniques were invented with the objective to scale the bird's beak proportionally to the other dimensions. The degree of success of these attempts and their trade-offs and limitations will become apparent through the following description of the basic LOCOS process and its derivatives.

Figure 4.2 shows the LOCOS process sequence [22]. A stack, consisting of a thin layer of thermal SiO_2 and a thicker layer of deposited nitride, is patterned with the active area mask and etched off in the field region. At this point, in a single polarity process, such as NMOS or PMOS, the field

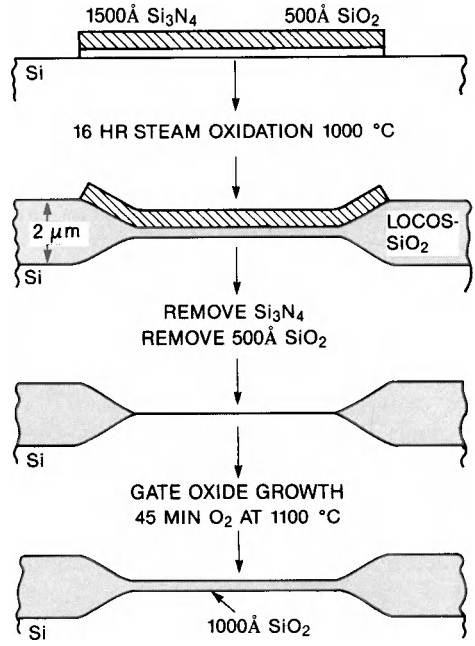


Fig. 4.2. LOCOS fabrication sequence. (After Kooi *et al.* [22]. Reprinted by permission of the publisher, The Electrochemical Society, Inc.)

channel-stop is implanted, while the active area is still protected by photoresist. However, in CMOS this is difficult to do, because a select mask is required after active area patterning to confine the field implant to the appropriate well. Nevertheless, since this is a low-energy implant, the oxide–nitride stack is sufficient to block it over the active area. Then, exploiting the property of nitride to inhibit thermal oxidation, a thick oxide is grown selectively over the field forming a partially recessed field insulator structure. Finally, after stripping the stack, the gate oxide is grown and the CMOS process is completed.

Observing the cross section after field oxidation, one can see that lateral oxidation occurs under the nitride edge, resulting in the so-called bird's beak due to its shape. Consequently, the actual active area edge is shifted inward by an amount roughly proportional to the field-oxide thickness. To reduce the bird's beak, various combinations of thicknesses have been tried for the nitride and oxide layers of the stack. It was found that the bird's beak is reduced by a thicker nitride and a thinner oxide [2]. However, the choice of thicknesses cannot be carried to the extreme, because dislocations are formed. Dislocation-free conditions have been found for less than 2000-Å Si_3N_4 and more than 78-Å SiO_2 films in conjunction with 1000°C field oxidation in wet O_2 [17]. In addition, the ratio of Si_3N_4 to SiO_2 should be less than 3:1 [29].

To improve planarity, a fully recessed LOCOS process was proposed [2]. Instead of stopping at silicon during etching of the LOCOS stack, a silicon layer is also removed to form a receptacle for the expanded SiO_2 produced by Si oxidation. The resulting change of cross section at the transition region is illustrated in Fig. 4.3. Notice that the bird's beak develops a well-defined crest, which could cause step coverage problems. For this reason and the difficulty of uniformly etching silicon without an etch-stop, this process has been used less than the standard LOCOS.

When the 1-μm MOSFET VLSI technology was introduced, the typical bird's beak size of $\approx$ 7500 Å was deemed excessive in comparison to the minimum feature dimension [30]. However, since the slope of the bird's beak is variable and very small near the active area, it appeared possible to significantly reduce the bird's beak size with a short, unmasked, oxide etch back. This was confirmed experimentally, by reducing the bird's beak from 7500 Å to 3500 Å with an etch back of only 1100 Å in a diluted HF solution [30]. Notice that in anticipation of the etch back, the field-oxide thickness was increased by an equal amount to compensate for the loss due to etch back. In addition, this technique has these advantages: (1) the surface becomes more planar, because the top field-oxide layer is removed, decreasing the height difference between this surface and the active area surface; and (2) the edge of the active area becomes better defined, because

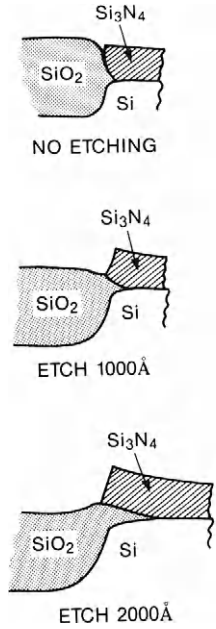


Fig. 4.3. Cross sections of recessed oxide structures showing the change of bird's beak and surface planarity. The LOCOS stack consists of 100 Å pad oxide and 2000 Å silicon nitride. The recessed field oxide thickness is 4500 Å. (After Bassous *et al.* [2]. Reprinted by permission of the publisher, The Electrochemical Society, Inc.)

the field oxide joins the active area at a steeper angle. Beyond these simple changes, which extend LOCOS to the border with submicron isolation, more complicated methods are needed for further scaling. These will be discussed in the following sections.

2. SWAMI

SWAMI (sidewall mask isolation) [18,19] is one of the best known LOCOS-based isolation techniques and was developed with the objective of retaining the advantages of LOCOS while drastically reducing the bird's beak. Since this is due to oxygen lateral diffusion through thin oxide from the active area edges, an obvious solution is to block it with a nitride barrier. This is the approach taken by SWAMI and illustrated in Fig. 4.4 with a schematic of the process flow [31].

Similar to LOCOS, the SWAMI process starts with the formation of the nitride–oxide stack. After patterning with the active area mask, the stack is etched down to silicon. A recess is formed in the field region with an anisotropic, orientation-dependent Si etch. Then, the channel-stop is im-

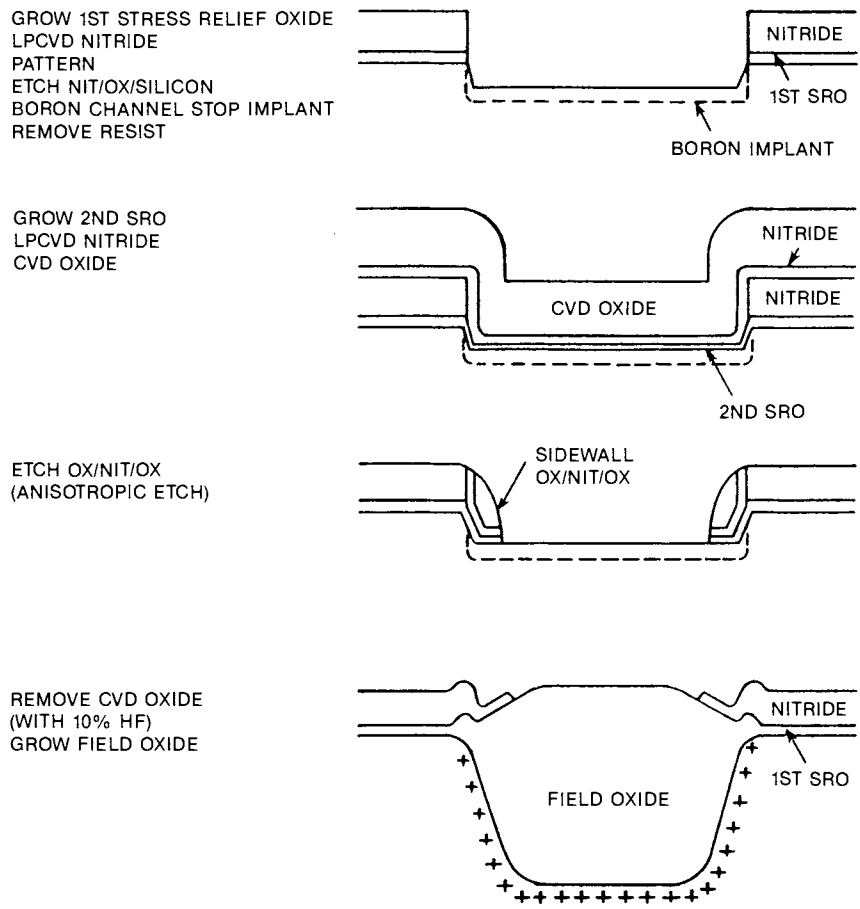


Fig. 4.4. Process flow for sidewall masked isolation (SWAMI) technique (After Teng *et al.* [31]. Copyright 1985 by the IEEE.)

planted while the active area is still protected by resist. At this point, SWAMI deviates from LOCOS by requiring additional steps to seal the sidewall edge, as described later. A second, thin, thermal oxide is grown over the field for stress relief purposes, followed by deposition of a sidewall nitride barrier using low-pressure CVD for conformal step coverage. Then, a third SiO_2 layer is deposited by CVD to increase the total thickness of the sidewall barrier seal. Taking advantage of the combination of conformal deposition and anisotropic etching, a "filament" or "spacer" is formed by RIE at the edges of the active area stack and Si recess. The width of this spacer is determined mostly by the total thickness of the sidewall barrier. After removal of the CVD oxide, still present in part of the spacer, the field

oxide is grown in wet O_2 and the nitride–oxide stack is etched off in preparation for subsequent device processing.

SWAMI has demonstrated a large reduction of bird's beak with scanning electron micrographs (SEMs) of device structures and electrical measurements of MOS effective channel width, which were obtained by extrapolating to zero the source/drain current versus the nominal width. While in standard LOCOS the channel width loss amounted to 1.3–1.5 μm, in SWAMI it was less than 0.1 μm, resulting in a large transconductance gain for narrow MOSFETs of the same drawn dimensions [18].

As natural in any new development, the earlier version of SWAMI presented fabrication weaknesses, susceptibility to Si defects, and higher source/drain leakage than LOCOS because of a "kink" in the subthreshold curve [26]. Later versions eliminated some of these problems, as will be discussed.

One troublesome yield loss was due to the tendency of the oxidation mask to fail at the joint between the first nitride and the second sidewall nitride because of overetching in the formation of the sidewall spacer. This failure was eliminated by intentionally undercutting the first thermal oxide at the edges of the stack and replacing the void with the second nitride [31].

The high susceptibility of SWAMI to generation of Si defects was brought under control by finding a relationship between the defect density and the recessed Si depth. The latter is important, because it controls the vertical length of the sidewall nitride barrier, which causes stresses in Si. By keeping the etch depth below a threshold of about 1000 Å, the structure appeared defect free [31].

However, in doing so, the attainment of surface planarity was sacrificed. To restore this feature, another Si etch was introduced after forming the sidewall spacer to create a second recess, further away from the active area edges [31]. Since the depth of this recess is not related to defect generation, it can be adjusted to optimize the surface planarity according to the field-oxide thickness. In conjunction with this recess, a second field implant can be done to increase the field threshold voltage of the channel-stop without increasing the electrical encroachment at the edges of the MOS channel. The version of SWAMI, which incorporates all these modifications, is called modified fully framed fully recessed (MF^3R) isolation and is schematically illustrated in Fig. 4.5.

Another important innovation introduced in SWAMI consists of stacking another CVD oxide layer over the original nitride–oxide stack. The process is then called STOMI (stacked oxide masked isolation) [32] and presents these advantages, as illustrated in Fig. 4.6: (1) protection of the first nitride film with the overlaid oxide during RIE of the sidewall spacer; (2) better control of the spacer width due to increased vertical support for

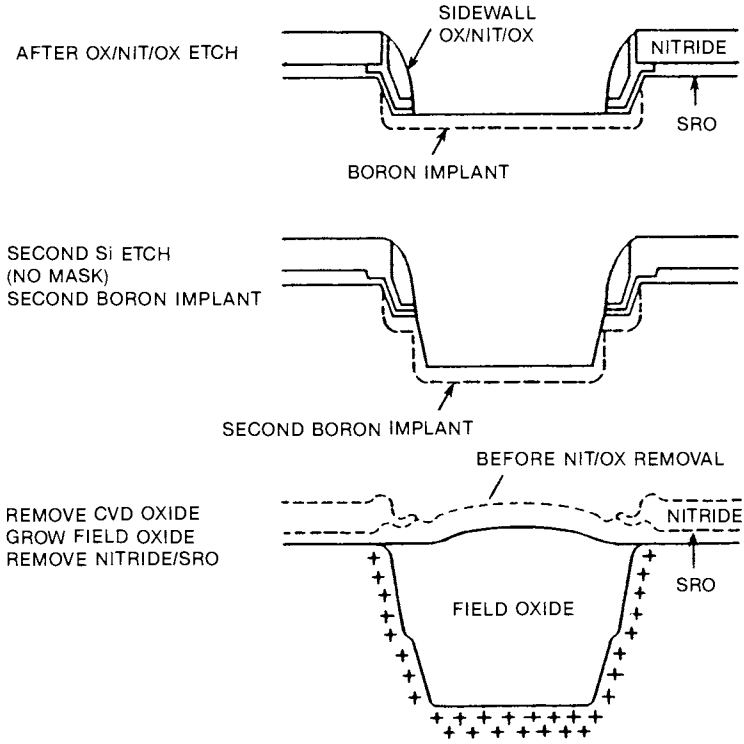


Fig. 4.5. Process flow of the doubly recessed MF^3R technique. (After Teng *et al.* [31]. Copyright 1985 by the IEEE.)

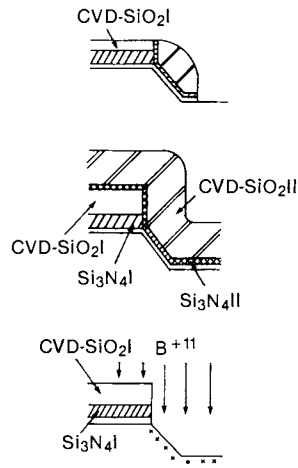


Fig. 4.6. Key features of the STOMI process. (After Sawada *et al.* [32]. Copyright 1985 by the IEEE.)

the spacer; and (3) increased implant protection of the active area during field implantation due to a thicker stack. Jointly, these advantages improve both manufacturability and yield, as demonstrated by applying STOMI to the fabrication of a 256-K DRAM [32].

If surface planarity is waived, STOMI can be further simplified by eliminating the Si recess and related etch, which is hard to control. This approach is implemented in the laterally sealed LOCOS isolation, resulting in a bird's beak $< 0.2\ \mu m$ and diode leakage similar to LOCOS [33].

Also trying to avoid the complications of the Si etch in SWAMI, but unwilling to forgo surface planarity, other researchers have proposed a modification of SWAMI, called fully recessed oxide (FUROX) field isolation technology [34,35]. The major difference is the method of forming the Si recess, which is based on a sacrificial field oxidation after patterning the nitride–oxide stack. Since the field oxide is partially countersunk, its removal leaves a uniform, well-defined Si recess. However, since this field oxidation preceeds the formation of the nitride spacer, a bird's beak will form unless prevented by appropriate action. For this purpose, the stress relief oxide is nitrided in pure ammonia gas to suppress the bird's beak using the SILO (Sealed interface local oxidation) technique [20]. The added process complexity, including nitridation and two field oxidations, is indicative of how much effort is considered acceptable to improve SWAMI, particularly in terms of manufacturability. Nevertheless, FUROX's excessive complexity defeats this goal, especially since it includes delicate steps, such as direct nitridation of silicon, which can easily generate silicon defects.

3. SILO

SILO's approach to bird's beak reduction consists of reducing to zero the thickness of the LOCOS pad oxide in order to seal the silicon interface under the LOCOS stack [20,36,37]. This eliminates the need for a perimeter nitride seal, as used in SWAMI, and allows for a simpler process.

SILO is easy to justify theoretically by analyzing the influence in LOCOS of the pad-oxide thickness on the oxidation rate at various distances from the nitride edges. Figure 4.7 shows this dependence with a family of curves for pad-oxide thickness ranging from 1 Å to 800 Å. Analytically, the local oxidation rate decays exponentially with the distance from the nitride edge, where the coefficient of the distance is inversely proportional to the square root of the pad-oxide thickness [20]. Thus, a thickness reduction increases the coefficient and the decay slope with attendant decrease of lateral oxidation. From a physical point of view, the thinner the pad oxide, the smaller is the lateral O_2 flux that diffuses under the nitride. Since O_2 is simultaneously consumed by Si oxidation during lateral diffusion, a

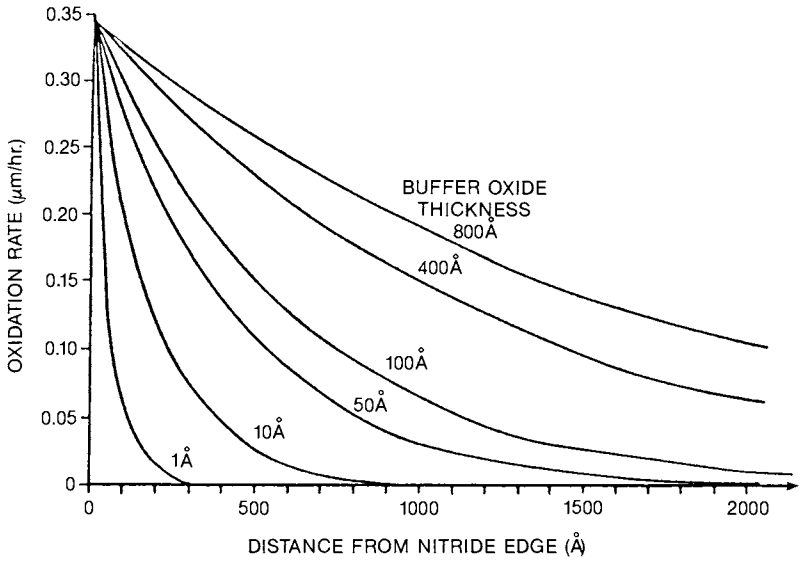


Fig. 4.7. Local oxidation rate as a function of the distance from the stack nitride edge for various thicknesses of the pad oxide. The curves are derived for wet oxidation at 950°C. (From Hui *et al.* [20]. Copyright 1982 by the IEEE.)

smaller flux is depleted faster and produces a smaller bird's beak. At the limit, if the pad oxide were totally eliminated, the bird's beak should not occur at all.

Unfortunately, reducing the bird's beak is not enough, because this must be done without increasing the Si defect density. In SILO the problem is solved by forming an active area stack with two different nitride layers, one very thin in direct contact to silicon and one much thicker on top of the stack. Between them, the usual pad oxide is retained for stress relief purposes. The thin nitride, 100–200 Å thick, is just for sealing the interface and is not intended to be a full-fledged oxidation mask, because the second nitride performs this function. The key to prevention of Si defects is the extremely small thickness of the sealing nitride, which limits the compressive stress induced in Si to values below the plastic deformation threshold [37].

The SILO structure is illustrated in Fig. 4.8 in two fabrication phases, before and after local oxidation [37]. Besides LPCVD deposition of Si_3N_4, energy-assisted techniques have been used for forming the nitride seal layer to improve its direct contact with Si and overcome the effect of native oxide. These techniques include, for example, nitrogen implantation [20,38] and plasma-enhanced thermal nitridation[20].

After extensive structural and electrical characterization, an optimized

choice of layer thickness has been proposed [37], consisting of 130 Å for the first (interface seal) nitride, 400 Å for the intermediate oxide, and 1000 Å for the second nitride. For a 0.5-μm-thick field oxide grown in wet O_2 at 950°C, these conditions generate few dislocations and a nearly zero bird's beak, as verified by TEM cross-sectional views. Histograms of leakage current from finger-type and gate-controlled diodes made with this SILO process produced results statistically equivalent to those of LOCOS control samples [37].

4. Non-planar Techniques

An important reason for the widespread acceptance of LOCOS is improved surface planarity compared to older nonplanar techniques. Nevertheless, novel nonplanar techniques have been proposed as late as 1987

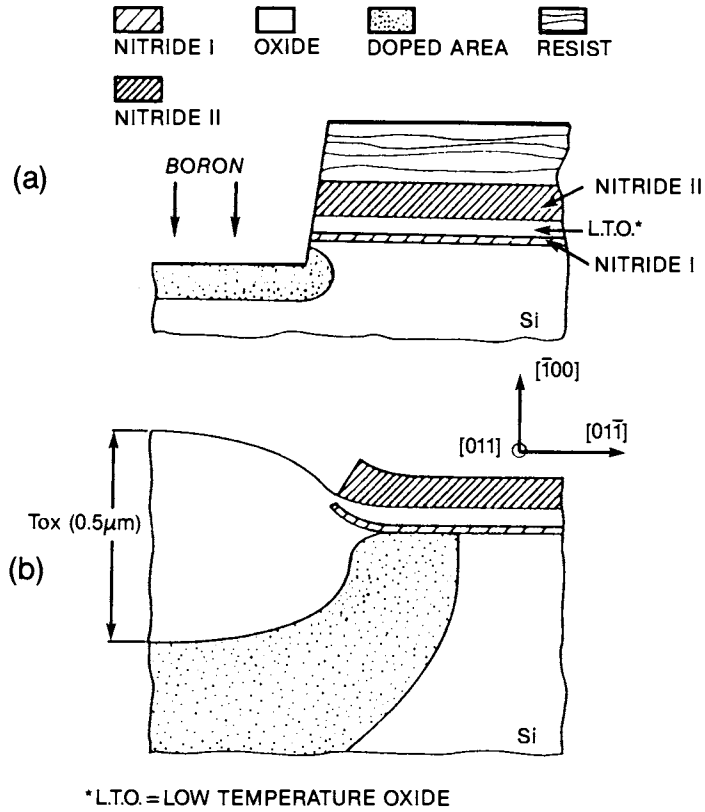


Fig. 4.8. Sample structure of SILO (a) before local oxidation and (b) after local oxidation. (After Deroux-Dauhphin and Gonchond [37]. Copyright 1985 by the IEEE.)

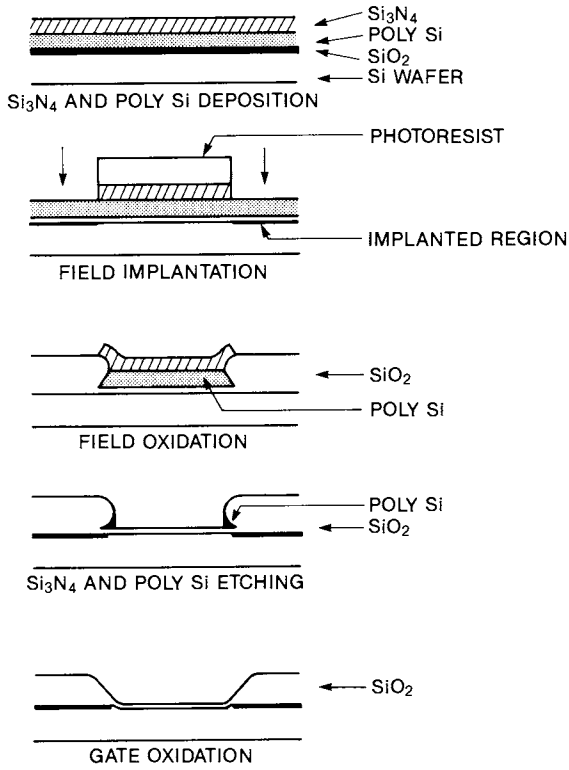


Fig. 4.9. Process sequence of SEPOX technology. (After Matsukawa *et al.* [41]. Copyright 1982 by the IEEE.)

[39] on the basis that they eliminate Si edge defects, which are often found in advanced LOCOS techniques, such as SWAMI and SILO. Though the bird's beak is missing, in its place there is a gently sloped transition region between field oxide and active area to ensure step coverage of subsequent films. Thus, these techniques do not improve density, and they clash with the trend toward higher resolution. This explains their limited success to the point that only one of them, SEPOX (selective polysilicon oxidation), has been extensively used in manufacturing [40,41].

SEPOX technology is actually similar to LOCOS, except that instead of growing the field oxide by selective oxidation of silicon, it uses a deposited layer of polysilicon for that purpose. The SEPOX fabrication sequence is illustrated in Fig. 4.9. An SiO_2 layer, 500 Å thick, is grown first. Then a polysilicon layer, 4000 Å thick, is deposited by LPCVD and is followed by a 3000 Å thick Si_3N_4 masking film. After patterning the active area, nitride is etched in the field, and, with the resist still in place, the field is implanted

to raise the threshold voltage. Next, the polysilicon is fully oxidized in the unmasked regions. A bird's beak is not formed, because the nitride mask and the polysilicon layer are in direct contact. The edge profile between field oxide and polysilicon is partially reentrant at the bottom, providing a convenient recess for leaving a polysilicon spacer while removing it with RIE in the active area. This spacer is then oxidized, and because of volumetric expansion of Si to SiO_2, the final field-oxide edge assumes a monotonic slope, as desired. With SEPOX, it has been reported that the actual active area edge is within 0.15 μm of the designed edge [41], compared to ≈ 0.75 μm for LOCOS using the same field-oxide thickness of 5500 Å. SEPOX is targeted for 2-μm technology and can be pushed to 1.5 μm with an adequate margin of tolerance. Theoretically, the minimum isolation spacing is 1.0 μm, as determined from leakage measurements of parasitic field transistors of variable gate length.

Though SEPOX shares with LOCOS the selective oxidation approach, this may be omitted in nonplanar isolation techniques, as demonstrated by the direct moat isolation (MOAT) technology [42]. Basically, this is an improved version of the old field-oxide-cut scheme, which was widely used in MOS aluminum gate fabrication. The MOAT process sequence is illustrated in Fig. 4.10. The improvements consist of (1) shaping the field-oxide contour with a straight line at $\approx 45°$ inclination for better step coverage, and (2) precisely defining the active area edges at their design position for full active area utilization. To achieve these results, the etch rate of the field oxide is increased in the top layer using argon implantation to generate a

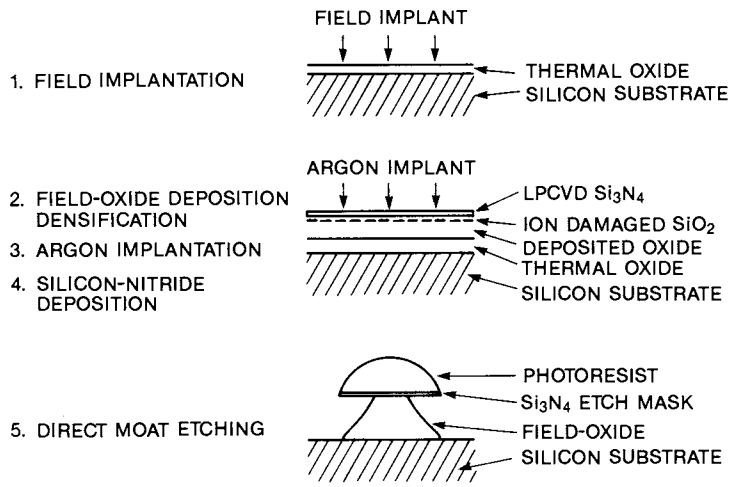


Fig. 4.10. Direct moat isolation process sequence. (After Wang *et al.* [42]. Copyright 1982 by the IEEE.)

carefully controlled oxide damage profile. Then, a diluted buffered HF solution is used to provide the inclined shape of the field-oxide edge contour, while RIE sets precisely the edge position in registration with an overlying nitride mask.

The weakest aspect of the MOAT technology is the channel-stop formation. Because of the lack of active area masking, there is no opportunity to form a self-aligned channel-stop, as in LOCOS or SEPOX. However, the issue is set aside by claiming that similar implant conditions can produce a channel-stop in the field region and simultaneously suppress the punch-through leakage current in the active area. Hence, a uniform implant across the entire surface should be sufficient and could be done before active area patterning. In practice, the trade-offs and design constraints are too restrictive for adequate process design latitude, especially in CMOS, where these adjustments must be made for both polarities. This reason and the fact that even theoretically the MOAT technology is only qualified for up to 2-μm isolation have limited its application to a very narrow technology window.

On the other hand, a novel nonplanar technology just published discloses a way to form a self-aligned channel-stop without a LOCOS approach. This process, called SAIL (self-aligned isolation using thin metal lift-off), is illustrated in Fig. 4.11 [43]. A field oxide is thermally grown first. Next, a clear field active area photoresist pattern is formed to selectively implant the field region. Then, a thin aluminum layer is evaporated over the photoresist pattern in preparation for lift-off. By immersion in a solvent, the aluminum pattern is removed over the active area, where it covers the photoresist, while it is left over the field oxide. Hence, the field oxide is protected during RIE of the oxide to open the active area window. After aluminum stripping, the structure comprises a self-aligned channel stop lying directly under the field oxide, while the active area still maintains the original doping. Because of the independent choice of channel-stop implant dose, there is wider latitude than with MOAT for choosing the field-oxide thickness. For instance, a reduced thickness could counteract the negative effect on step coverage due to sharp edges formed by RIE. Small circuits have been fabricated with SAIL isolation using 1-μm design rules, showing the capabilities of this technology.

5. Buried Oxide Isolation

Buried oxide (BOX) isolation addresses the needs of IC technology with feature size of 0.5 μm or below. At this level of resolution, neither LOCOS nor its advanced modifications are expected to provide the required surface planarity, field-oxide thickness, edge contour, and channel-stop character-

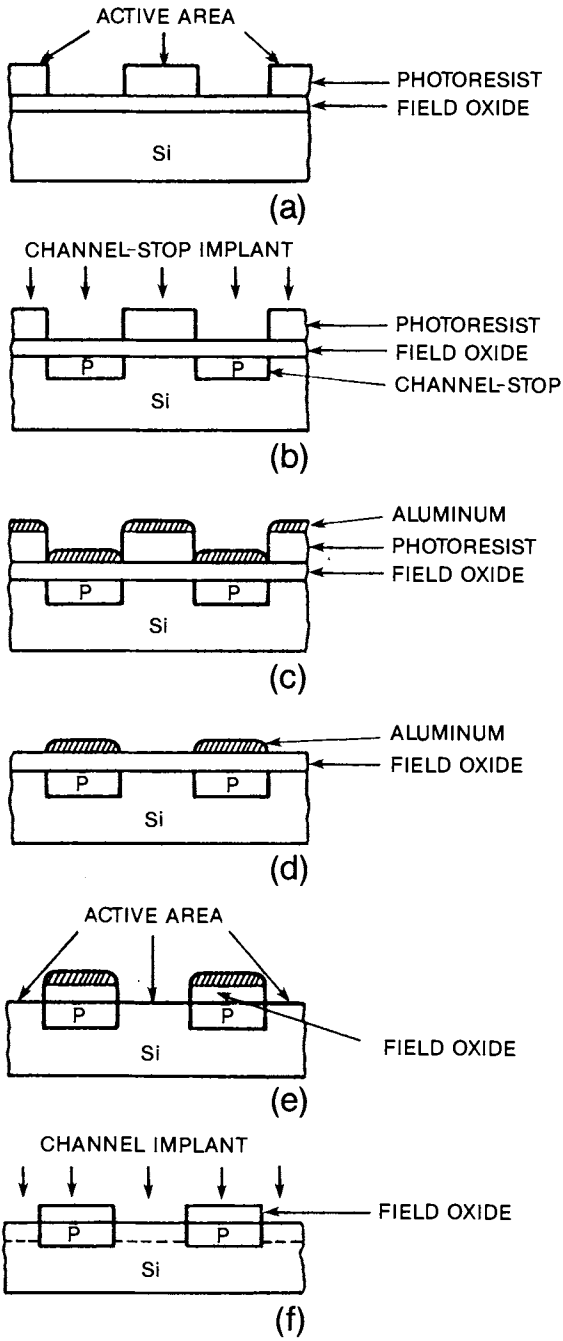


Fig. 4.11. SAIL process sequence. (After Lee *et al.* [43]. Copyright 1987 by the IEEE.)

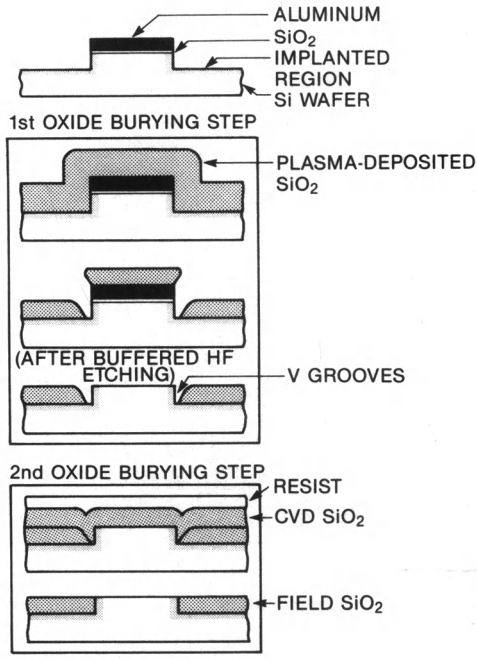


Fig. 4.12. Fabrication sequence of BOX-I. (After Kurosawa *et al.* [15]. Copyright 1981 by the IEEE.)

istics. On the other hand, the BOX isolation has the potential to fulfill these needs, though in practice it has to be further developed to become a robust manufacturing method.

The fabrication process of the original BOX isolation (BOX-I) is presented in Fig. 4.12 [15]. The key process steps are listed here.

1. Form an active area stack with a thin thermal oxide, topped by an aluminum film.
2. Pattern and RIE etch Si mesas, $\approx 0.7\ \mu$m deep.
3. Implant the self-aligned channel-stop.
4. Selectively refill the recessed field regions with plasma-deposited oxide.
5. Preferentially etch with buffered HF the oxide on the mesa sidewalls.
6. Lift off the active area stack with aluminum etch.
7. Deposit a planarizing oxide to fill the voids around the Si mesas.
8. Planarize the surface with a polymer, usually photoresist.
9. Etch back the polymer and the oxide with equal etch rates until the Si is exposed, in order to duplicate in the oxide the smooth polymer surface topology.

Even from this brief description of the BOX-I process it is easy to realize the manufacturing problems. Foremost, the aluminum lift-off is unreliable, especially in combination with unmasked preferential etching of the plasma-oxide sidewalls. Next, it is difficult to control the uniformity of the field-oxide edge contour at the Si mesa corners. This might leave some Si corners exposed, causing crowding of the equipotential lines, a high electric field, and subthreshold leakage [26]. Finally, the grazing incidence of the implant beam on the nearly vertical sidewalls may compromise the channel-stop formation on all sides of the mesa. This could cause *n*-channel subthreshold leakage, due to weak inversion from positive fixed charge in the oxide near the sidewalls, as illustrated in Fig. 4.13, which depicts a

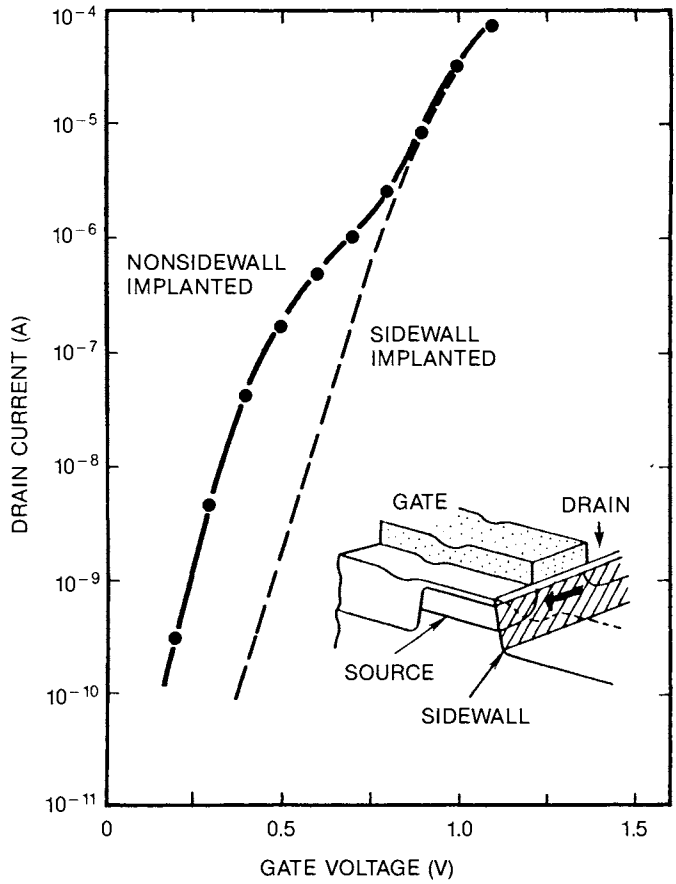


Fig. 4.13. Subthreshold characteristics of a transistor with $W = 20\ \mu m$, $L = 3\ \mu m$. A hump is observed in the curve of transistors with nonimplanted sidewalls. (After Kurosawa *et al.* [15]. Copyright 1981 by the IEEE.)

typical "hump" of the subthreshold curve if the sidewalls are not adequately implanted.

Despite these problems, test circuits using BOX confirmed the attractive features of this technology, as demonstrated by these results:

1. The layout density increased by 80% compared to LOCOS using 1.0-μm design rules.
2. The drawn active area dimensions were faithfully reproduced in silicon because of the total absence of the bird's beak.
3. The final surface was free from sharp edges and steps.
4. The threshold voltage was constant as a function of channel width because of the elimination of the narrow width effect found in LOCOS.
5. The parasitic capacitance was reduced because of a decrease of its diffusion edge and interconnection-to-substrate components.

These results proved to be a powerful incentive to improve the fabrication process. Hence, a second version of BOX, referred to as BOX-II, was developed and is schematically illustrated in Fig. 4.14 [44]. To ensure adequate sidewall doping, the mesas are etched with an orientation-dependent Si etch, which provides an inclination of $\approx 60°$. Instead of using lift-off, the surface is planarized with a double-resist process and an additional noncritical masking step. This step is needed because a polymer-based planarization is a short-range phenomenon, which, although it is excellent for filling narrow voids with high spatial frequency, is nevertheless incapable of planarizing extensive recessed regions. Hence, the first resist is used to build up the height of the polymer in the field-recessed regions with a complementary active area pattern. Then the second resist fills in the crevices at the joints between the oxide-covered Si mesas and the

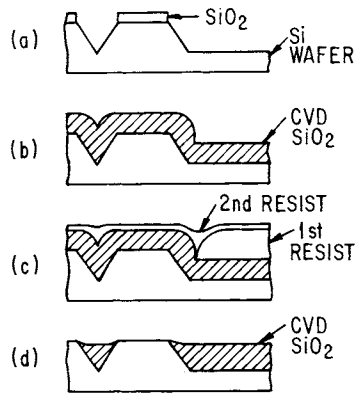


Fig. 4.14. Fabrication sequence of BOX-II. (After Shibata *et al.* [44]. Copyright 1983 by the IEEE.)

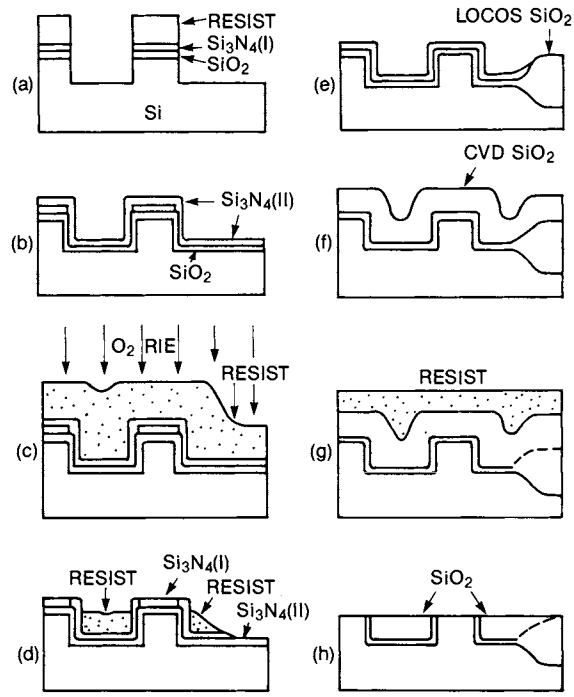


Fig. 4.15. New isolation process sequence combining BOX and LOCOS. (After Fuse *et al.* [45]. Copyright 1987 by the IEEE.)

first resist to form the planarized surface as in BOX-I. By virtue of these improvements, BOX-II is easier to manufacture, but it is still afflicted by process control problems, such as field isolation uniformity at mesa corners, double-resist processing, and registration of the first-resist pattern. Moreover, the adoption of sloped sidewalls contrasts with the trend toward higher resolution.

In a renewed effort to solve the remaining buried oxide isolation problems, a hybrid technique, which includes the best features of both BOX and LOCOS, has been proposed according to the process flow of Fig. 4.15 [45]. Instead of using lift-off or an additional resist pattern to selectively refill large field regions, this method turns to advantage a planarization problem, namely, the inability of keeping a uniform polymer level over a surface topology with a low spatial frequency. Specifically, the decrease of resist height away from the mesas is used to automatically determine where the field oxide must be built-up to remove the troublesome low spatial frequency dependence of the active area pattern. As seen in Fig. 4.15, after partial etching, the photoresist is left only near the mesa pattern and is totally removed elsewhere. Therefore, by placing a composite stack of

nitride and oxide under the resist, a natural edge bias is formed around the mesas after nitride etching. This prevents interference of the LOCOS bird's beak with the BOX structure during the following selective field oxidation. In addition, since with LOCOS the growth of field oxide is symmetrical with respect to the silicon surface, the field-oxide thickness in the LOCOS area is twice that of the BOX, further reducing the interconnection capacitance to the substrate. The crevices around the mesas are then filled by depositing a second oxide layer, which is planarized with a second resist film, followed by etch back, to expose the mesas' top silicon surface and the surrounding coplanar oxide isolation surface.

In this technique, the problem of low sidewall doping is resolved by implanting each sidewall with the same tilt angle, independently of its orientation on the IC layout. Accordingly, the wafers are implanted in four different positions, each rotated by 90°, to establish a common inclination of the beam with respect to the sidewall surface and to eliminate shadowing from the opposite sidewalls. With this procedure, the uniformity of the sidewall doping below the channel edges is guaranteed, yielding NMOS devices with low subthreshold leakage.

A discussion of buried oxide isolation would not be complete without mentioning a very elegant isolation technique, called PHOTOX™ [46], whose sequence is shown in Fig. 4.16. This method is based on the deposition of oxide at very low temperatures (100°C or less), to be compatible with photoresist processing. Since an energy-assisted CVD deposition process is needed at this temperature, UV light is used to promote the chemical reaction for the deposition of SiO_2, explaining the origin of the name *PHOTOX™*.

With reference to Fig. 4.16, the *PHOTOX* process flow is briefly described as follows. After active area patterning, a silicon recess is formed over the field region and later refilled with *PHOTOX™* oxide, which is simultaneously deposited over the resist on top of the active area. Using lift-off, both the resist and the overlying *PHOTOX* oxide are removed, leaving the active area bare and ready for continuing with the CMOS device fabrication. Instead of forming the channel-stop as usual while the active area is still masked, the channel-stop is implemented with a retrograde approach, exploiting the different heights between the active area surface and the field-oxide–silicon interface. Therefore, using a high-energy "blanket" implant, the concentration peak in the field could be positioned at the SiO_2–Si interface, while in the active area it would be formed well below the surface without possibility of interference with the MOS device characteristics. This approach is now popular with the retrograde-well CMOS technology and has the advantage of providing a high doping level on the sidewalls without impurity losses due to thermal redistribution, as in LOCOS.

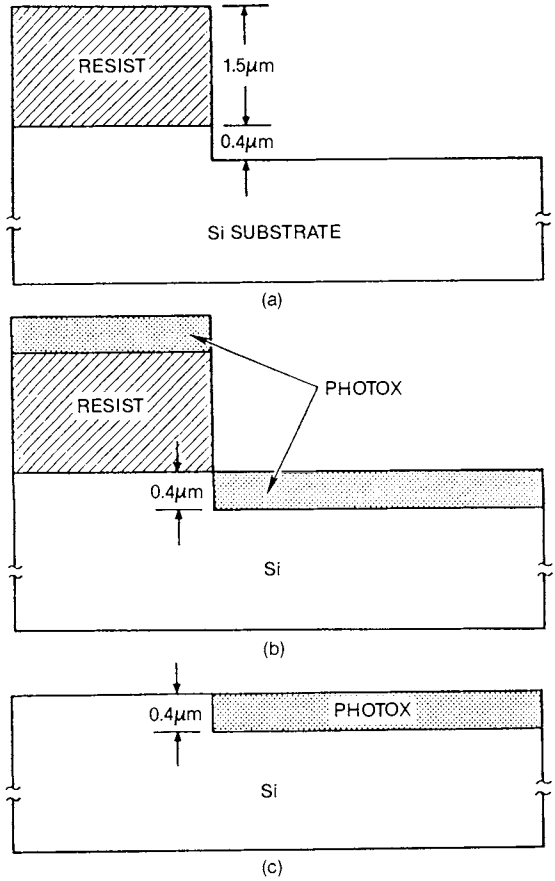


Fig. 4.16. Process sequence for PHOTOX isolation. (After Chen *et al.* [46]. Copyright 1982 by the IEEE.)

6. *Trench Isolation*

The use of trenches for field isolation represents an evolution from the former use of grooves [47]. As etching technology improved, the grooves became deeper and deeper with a larger aspect ratio, and eventually turned into trenches. In CMOS/BULK, trenches can be used for field isolation, well isolation, or special features, such as trench capacitor for DRAM. Though the distinction between these categories is somewhat blurred, we will concentrate here on field isolation [6,48].

The major objective of trenches is to achieve high density without suffering an increase of isolation leakage or having to reduce the supply voltage. This is achieved by folding the silicon surface across the isolation

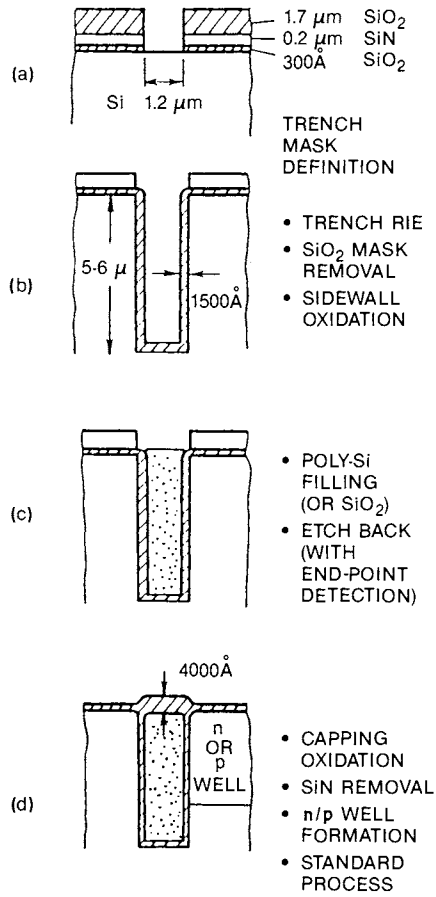


Fig. 4.17. Schematic description of the trench isolation process sequence. (After Rung *et al.* [6]. Copyright 1982 by the IEEE.)

spacing to form a deep, narrow barrier, which increases the current path many times with respect to the spacing. Moreover, by digging the trench so deep that its bottom lies in heavily doped silicon, such as the substrate of an epitaxial wafer, the isolation leakage suppression is virtually complete.

Figure 4.17 illustrates the basic process sequence for trench isolation using typical dimensions [6]. A stack is formed first, which consists of 1.7 μm SiO_2, 0.2 μm Si_3N_4, and 300 Å thermal oxide. Using a special mask, the trench is defined by opening in the photoresist a line pattern, whose width is equal to the minimum feature size, for example, 1.2 μm. The stack is etched by RIE stopping at silicon. Next, the resist is stripped, preferring to use the underlying oxide as a mask during RIE of Si to form

the trench. The trench depth is $\approx 5-6\ \mu$m. Notice that this is a key step in the process, because it requires excellent anisotropy to produce a deep trench with nearly straight sidewalls. A gas chemistry, based on chloro-fluorinated methanes, such as dichloro-difluoro-methane, is usually employed. The remaining SiO_2 is stripped and, using the underlying nitride as a mask, the trench surfaces are thermally oxidized to grow about 1500 Å oxide. This oxide provides dielectric isolation and a diffusion barrier for source/drain and well impurities. Then the trench is refilled with undoped LPCVD polysilicon and, since the process is not selective, the flat oxide surface is covered as well. Alternatively, LPCVD oxide is employed instead of polysilicon for trench refilling, depending on the applications. To remove the extra polysilicon layer an etch back is required using end-point detection to avoid overetching. Finally, with the nitride still in place, the trench is sealed by selectively oxidizing the top polysilicon surface to ≈ 4000 Å. Since the nitride is no longer needed, it is removed before resuming the normal CMOS process.

An early application of the trench method was to isolate adjacent NMOS and PMOS devices to save "real estate" in the layout of basic inverters and logic gates through a reduction of $n+$ to $p+$ spacing requirements [6]. Because of the superior electrical isolation provided laterally by the trench, the related design rule could be reduced in a 1.25-μm CMOS process from the usual $7-8\ \mu$m to only $1-2\ \mu$m, achieving ≈ 4 times reduction. Since, in this application, one side of the trench faces the well while the other side faces the substrate (or the tub of opposite polarity), the trench provides well isolation in addition to field isolation. The well lateral diffusion is eliminated, because it is blocked by the oxide-coated trench sidewalls, henceforth maintaining the well edges at their lithographically defined position. Moreover, the trench eliminates the lateral space charge region around the well, which in diffused wells has been a large component of this spacing. In addition to the well, similar conditions apply to the source/drain regions, although their impact on the spacing is smaller because of shallower junctions. Finally, the bird's beak is reduced, because the field oxidation consumes polysilicon from the trench surface. Unfortunately, these advantages are offset by problems, mostly in the form of parasitic leakage paths, as will be discussed [6].

Figure 4.18 illustrates the two main parasitic leakage paths, both affecting the NMOS device and located along the trench sidewall interface: 1. horizontal, under the channel edges, between source and drain; 2. vertical, across the p well, from the $n+$ source (or drain) to the n substrate. These paths operate like parasitic MOS transistors, which are turned on by the positive charge on the oxide sidewall interface and by the positive supply voltage of the n substrate on the other side of the trench. Moreover, the

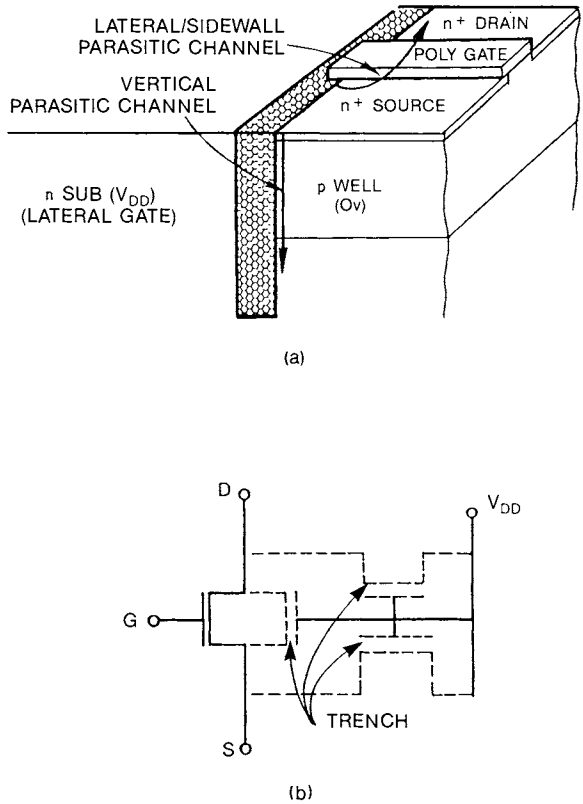


Fig. 4.18. (a) Parasitic channels in the trench process. (b) Schematic of *n*-channel device plus parastics. (After Rung *et al.* [6]. Copyright 1982 by the IEEE.)

horizontal path is enhanced by the corner effect, which may lead to crowding of electric field lines and attendant higher carrier concentration at the corner near the channel edges, as determined by 2-D device modeling [48]. Pulling back the source/drain from the trench isolation is one way to provide a safety margin. Other ways include higher doping of the *p* well and careful control of the sidewall fixed charge density [48]. For instance, using the 2-D device simulator GEMINI, it was determined that to avoid inversion with a 1-μm wide SiO_2 trench the fixed charge density had to be less than 5×10^{10} cm^{-2} [7]. Selecting a $\langle 100 \rangle$ orientation for the trench sidewalls instead of the $\langle 110 \rangle$ presently used could help to lower the fixed charge density and to equalize the sidewall and bottom oxide thickness [49]. This has been proven with flat-band voltage measurements of trench capacitors with a variable sidewall to bottom surface ratio.

Sometimes it is impractical to raise uniformly the doping level to ensure that it is sufficiently high around the trenches. In the case of trench capacitors for DRAMs, a deep implant in the memory region solves this problem without perturbing the more lightly doped peripheral epi layer, where CMOS circuits are laid out without trenches [50].

Particularly for trench capacitors, the trench oxidation properties are important to ensure high dielectric strength, high breakdown voltage, and low leakage current of the capacitor dielectric. Fortunately, these properties were found to be only slightly inferior to those of planar surfaces [51], with the weakest spot being at the corners, where the oxidation rate is smaller [52]. However, using a sacrificial oxidation, these corners can be rounded off erasing any difference between the trench and the planar oxide properties [53].

Finally, we note that trench isolation is expanding beyond its original purpose to provide other less obvious properties. For instance, its use has been reported to isolate bit lines and double-polysilicon FAMOS transistors in a high-density EPROM memory. Because of decreased parasitic capacitance, a higher coupling efficiency between control and floating gates has been realized resulting in enhanced programmability [54].

7. Selective Epitaxial Growth

This method can be considered the inverse of BOX or trench isolation. Instead of refilling with oxide a silicon recess over the field, a deep trench is etched anisotropically over the active area cutting into a previously grown thick oxide layer until the silicon interface is exposed. Then, using this silicon surface as a catalyst, the trench is selectively refilled with single-crystal epitaxial silicon, leaving a planar surface ready for CMOS device fabrication. Though there are many variants of this technique, all have in common the key process step of selective epitaxial growth (SEG).

Early reports of "selective" epitaxial growth date back to 1962 [55], but the silicon quality and selectivity of the original method were inadequate for silicon device fabrication. Nevertheless, more than a decade later, silicon epitaxial growth was utilized for refilling oxide trenches [56,57], but because of poor selectivity, a polysilicon layer was concurrently deposited over the top oxide surface, requiring a planarization process for its removal and therefore complicating the isolation sequence. Fortunately, in 1982 it was discovered that truly selective epitaxial growth occurs at reduced pressure and temperature by adding a small percentage of HCl to the SiH_2Cl_2-H_2 gas system [58,59]. This discovery, which came in time for submicron technology development, will probably enable further down-scaling by lifting limits, which are intrinsic to other isolation techniques.

Indeed, recent reports suggest that an isolation width as small as 0.25 μm can be obtained with SEG [60].

SEG isolation structures are simple, with few deviations from the basic method of oxide etch and silicon refill. Often, the differences are relegated to the method of well formation, since SEG allows for buried implanted layers and retrograde twin tubs. Figure 4.19 shows a fairly typical CMOS isolation sequence using SEG [61]. In this implementation, a 2-μm-thick SiO_2 is grown on (100)-oriented $p+$ silicon substrate. Different from the usual, the active area mask is aligned along the $\langle 100 \rangle$ direction instead of the commonly used $\langle 110 \rangle$ to reduce faceting at the corners of the epi surface [62]. Then the oxide is reactive ion etched down to silicon in the active area openings to form trenches with nearly vertical sidewalls, with the attendant damage being annealed during a short sacrificial oxidation. Using a noncritical n-tub mask, a buried $n+$ layer is formed at the bottom of the n tub leading to a retrograde well profile. After stripping the thin sacrificial oxide in diluted HF, the SEG process is carried out at 50 Torr and 950°C and is followed by a 5-hr anneal at 950°C in N_2 before resuming the normal CMOS process.

Through both SEG and BOX yield structures with similar oxide/silicon contours, SEG presents several advantages. First, SEG trenches are deeper, because, without the need for planarization, their depth is not a factor

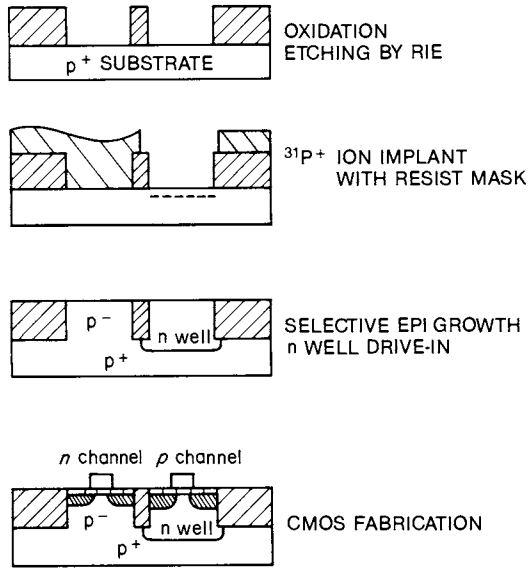


Fig. 4.19. CMOS isolation sequence using SEG and buried n-well structure. (After Endo *et al.* [61]. Copyright 1986 by the IEEE.)

during trench refilling by SEG, henceforth improving the electrical isolation characteristics. Concurrently, the avoidance of planarization, which on the other hand is an integral part of BOX, leads to a simpler process with SEG, which has the additional advantage of being insensitive to spatial frequency variations of the trench pattern during silicon growth, thereby improving the surface planarity [59]. Second, SEG isolation is naturally suited for forming highly conductive buried layers at the bottom of the well using low-energy implants. In conjunction with low-temperature processing, these layers produce deep retrograde wells, which further promote latch-up resistance. Indeed, it has been reported that with a trigger current of 2.3 mA and an isolation width of 4 μm, the holding voltage for a SEG structure was 10 V compared to 1.2 V for a LOCOS structure with a conventional *n* well [61]. Third, the sidewall *p*-well isolation leakage, which is also present in BOX, can be reduced with SEG by lowering the temperature of epitaxial growth to 875°C and therefore decreasing the interface defect density [63]. Actually, for CMOS technology at 0.5 μm or below, this effect is totally eliminated because of a necessary increase of doping level under the channel for avoiding punch-through leakage between source and drain [64].

Since SEG isolation relies on buried layers implanted through the active area openings to form a well, this practice may lead to an excessive number of well contacts and to decreased layout density. However, this can be overcome by electrically connecting adjoining wells through their buried layers and their lateral diffusion extensions [65]. If additional SEG regions are needed to provide continuity, but are not utilized for active devices, they may be passivated with field oxide, as shown in Fig. 4.20.

At the leading edge of this technology, SEG has demonstrated an incredibly short active area spacing across the well of only $\frac{1}{4}$ μm, despite the use of a CMOS technology with a 0.5-μm minimum feature size, as defined by its lithographic resolution capability [60,66]. The reason for this apparent contradiction is due to the fact that in this SEG version the n-tub trench is not opened in a thick oxide, as described earlier, but in the *p* tub itself, which is formed by the silicon substrate. Since the isolation between the tubs is formed by coating the trench with a 0.25-μm-thick dielectric "spacer" before being refilled by SEG, there is no need to use lithography to define the width of this spacing, explaining the reason for its exceedingly small value [60].

This approach has many merits besides producing the narrowest isolation spacing ever reported. First, by making the NMOS devices in the original substrate (*p* tub) and forming the oxide–silicon interface by thermal oxidation, the defect density is very low and the sidewall leakage is minimized. Second, the mask count is reduced, because the *p* tub may be

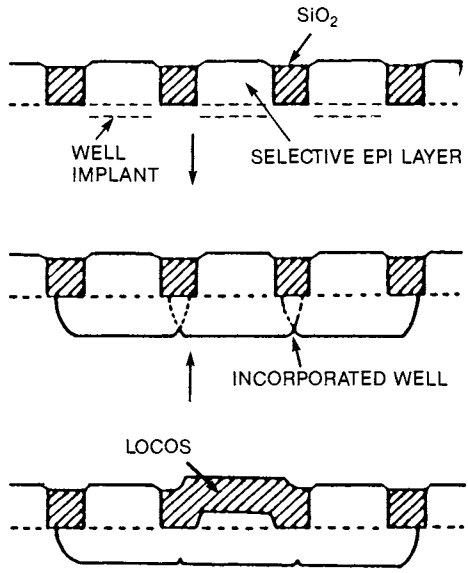


Fig. 4.20. Formation of single well with multiple active areas using SEG isolation. (After Nagao *et al.* [65]. Copyright 1986 by the IEEE.

doped with blanket implants, since their effect on *n*-tub doping is erased as a result of silicon removal in the *n*-tub trench, followed by replacement with virgin SEG-grown silicon.

The rapid progress experienced by SEG has resulted in a growing number of publications and improvements in SEG fabrication equipment, leading one to believe that this technique will play a major role in CMOS submicron isolation at 0.5 μm and below.

III. WELL ISOLATION

A. Features, Issues, and Applications

The challenge of CMOS technology is to fabricate side by side on the same chip both NMOS and PMOS transistors merging their respective process steps. This implies that silicon regions of opposite polarity must coexist on the wafer to provide local substrates for the complementary transistors. Since a bulk wafer already has a definite polarity, the region of opposite polarity is commonly referred to as the well or tub [3].

But, besides being of the right polarity, the well and the substrate must be electrically isolated from each other under every operating condition. For

this reason, they are reverse biased with the supply voltage, which, being the largest voltage in the circuit, ensures the maintenance of the reverse-bias condition during normal voltage variations at internal circuit nodes. Nevertheless, under special circumstances, over-voltage conditions may occur in the well or the substrate, temporarily forwarding a part of the junction and starting a silicon controlled rectifier (SCR) action, commonly referred to as latch-up [10]. Effective methods of eliminating latch-up have been devised and involve either circuit techniques or special process features, which will be discussed at length later. With these techniques, latch-up has been nearly eliminated in modern CMOS circuits despite increased susceptibility due to scaling.

There are many issues regarding well isolation besides latch-up that affect the entire process architecture. These can be considered as criteria for selecting among a variety of implementation options, such as single well or twin tub, *n* well or *p* well, retrograde or diffused well, bulk or epitaxial substrate, and the use or not of trench isolation at the well edge.

The single-well approach has been universally used in the past and is still popular because of its simplicity [67,68]. Normally, with this technique, the well is formed at the beginning of the CMOS process by ion implantation, followed by deep diffusion at high temperature and selective field oxidation. The advantages of this approach start to vanish as the minimum feature size decreases, because it becomes progressively more difficult to simultaneously optimize both NMOS and PMOS devices. For instance, assuming a CMOS *p*-well technology, the surface doping of the well must be at least five times that of the substrate to reproducibly define the well characteristics, particularly its depth and surface concentration. Hence, a lightly doped *n*-type substrate is needed, although it is susceptible to source/drain punch-through leakage in short-channel PMOS devices. The solution is to implant the PMOS active area with a deep phosphorus punch-through control implant, followed by a shallow boron implant for threshold adjustment, in order to compensate for the negative work function component of the $n+$ polysilicon gate. Considering these complications, the alternative twin-tub approach becomes more appealing for submicron CMOS [69,70], because it presents an integral solution to these problems.

Another reason why the single-well approach is not suited for submicron CMOS technology is that the well lateral diffusion does not scale at the same rate as the planar dimensions, with this problem becoming worse at higher resolution. Consequently, the density increase predicted by scaling cannot be realized. Moreover, since the substrate is more lightly doped than the well, a wide space charge region surrounds it and must be added to the well lateral diffusion for determining the minimum spacing between

NMOS and PMOS, unless advanced isolation techniques, for example, trenches, are employed. Finally, since constant voltage scaling is often used, a high field threshold voltage must be maintained despite reduced field-oxide thickness, requiring a higher doping level at the field-oxide–silicon interface. While on the well side this can be achieved with increased well doping, on the substrate side a field implant is needed. This in turn requires an extra mask to avoid counterdoping of the well during field implant. These complications, which represent piecemeal solutions to the problems of extending the single-well technology beyond its natural design boundaries, are indicative that this technology has run its course at about 1-μm resolution and that trying to push it below this level is economically counterproductive.

On the other hand, in the twin-tub technology, each well is formed separately by doping it at a higher concentration than the substrate, henceforth reducing the substrate's main functions to mechanical support and bias polarity selection. Two definitions are commonly given for the twin-tub technology. The less stringent definition requires that the *n* tub and *p* tub be symmetrically formed and that NMOS and PMOS FETs be independently optimized, although the substrate polarity is fixed and the process flow is optimized for this choice of polarity. On the other hand, under the most aggressive definition, this restriction is lifted, allowing that the choice of substrate be determined purely by circuit applications without changes in the process flow. From a manufacturing viewpoint, this is an ideal situation, because it may reduce the number of fabrication lines without sacrificing circuit needs, although the engineering effort is greater. These considerations will become apparent from a detailed description of the twin-tub process, which will be given in a following section.

Another important advantage of the twin-tub approach is the symmetric treatment of the well polarity. Since the process architecture is independent of the substrate polarity, the choice of the well type is reduced to a matter of either connecting the well to ground and the substrate to V_{DD}, or vice versa. Thus, the advent of the twin-tub technology has put to rest the controversy over the choice of *n*-well versus *p*-well technology, which was fought for many years with process- and device-related arguments, and contributed to the development of separate *n*-well and *p*-well processes within the same company. Now, the choice of well type (or more appropriately substrate type) is recognized to be application dependent, adding further value to the twin-tub technology by virtue of its higher circuit design flexibility.

To gain an historical perspective, we will briefly mention the major arguments of *n*-well and *p*-well supporters during the CMOS evolution, and then we will consider the application aspects, which are still valid. The

use of *p*-well technology was often justified on these grounds: (1) better matching of NMOS and PMOS "gain" constants, since the carrier mobility reduction associated to higher well doping could be compensated by forming the NMOS in the *p* well because of the electrons higher mobility; and (2) higher NMOS field threshold voltage, since the enhanced well doping could compensate for boron segregation into the field oxide and the effect of fixed oxide charge. Before the availability of ion implantation this was a preeminent concern, because surface inversion under the NMOS field could not be easily eliminated in CMOS without recourse to area-consuming "guard rings."

The *n*-well approach was proposed late, in 1979 [71], principally to extend the benefits of CMOS to large families of existing NMOS circuits with only minor alterations of both process and design. Besides this pragmatic justification, other arguments presented in favor of this technology have been (1) the speed is higher at comparable resolution, because *n*-well circuits are predominantly composed of NMOS devices, which are fully optimized in this technology; (2) the substrate current due to electron impact ionization is outside the well, therefore improving latch-up immunity by eliminating this major source of lateral voltage gradient within the well (by contrast, the hole impact ionization is several orders of magnitude smaller and its effect on latch-up can be neglected); and (3) the transition to Bipolar CMOS (BICMOS) technology is easier because of the presence of an *n* well, which can serve as an electrically isolated collector [71].

As mentioned earlier, circuit considerations now prevail in the selection of an *n* well or a *p* well within the context of a twin-tub environment, as explained in the following discussion.

High-density SRAM circuits usually employ the *p*-well configuration to satisfy simultaneously these requirements: (1) to keep the memory array within the well for minimizing charge collection from the substrate due to alpha particles; (2) to use NMOS devices in the memory array for achieving high switching speed and large sensing current; (3) to utilize polysilicon load resistors in 4-transistor memory cells and buried contacts; and (4) to achieve higher layout density through the use of $n/n+$ epi substrates, by exploiting their lower latch-up susceptibility compared to $p/p+$ because of thinner epi layers and heavier substrate doping [72].

In DRAMs, because of the need for soft error immunity from alpha particles, the memory array is usually formed within the well. However, because of the dynamic nature of charge storage, the major requirement is to minimize cell leakage to reduce the refresh frequency. For this reason, PMOS devices are preferred in the array, since they exhibit 3 to 4 orders of magnitude less drain to substrate leakage due to lower impact ionization of holes versus electrons. As a consequence, the typical DRAM memory is

implemented with an *n*-well configuration, though the actual process is often twin tub on *p*-type or *p*/*p*+ substrate. The substrate properties also favor *n* well, since the leakage current has been found to be statistically smaller for *p*/*p*+ versus *n*/*n*+ by as much as two orders of magnitude [73]. Recently, the use of trenches has lifted some of these restrictions because of the growing trend of storing charge on the inner side of the trench, which is isolated from the substrate and hence relatively immune to logic upset from alpha particles [74]. This supersedes the need of enclosing the memory array in a well, with a consequent change of device type for the cell access transistors from PMOS to NMOS and the use of the *n* well only in the peripheral circuits.

Nonvolatile Electrically Programmable (EPROM) or Electrically Erasable Programmable (EEPROM) memories definitely benefit from *n*-well technology, because only NMOS transistors are suitable for programming and only the substrate has the resiliency to voltage and current disturbances, which occur during write/erase operations. Specifically, the reasons are slightly different in EPROMs and EEPROMs, as will be explained. Programming of EPROM is based on generation of hot electrons in *n*-channel devices. The resulting substrate current also influences the programming efficiency. If a *p* well is used, the substrate current flowing into the *p*-well contact will result in a lateral potential drop along the *p* well, which will be larger for devices further away from the *p*-well contact. This location dependence of the potential drop will result in nonuniform programming characteristics. Therefore the *n*-channel devices in the memory array must be placed in a *p*-substrate, and consequently, *n* wells have to be used.

In the case of EEPROMs, assuming the conventional floating gate and tunneling oxide approach, *n*-channel devices are required for obtaining electron tunneling. In the erase operation, high voltage has to be applied to an *n*+ diffusion. The combination of high voltage and *n*-channel devices leads to impact ionization and high substrate current, which in the case of a *p* well may upset the logic state of memory cells and eventually produce latch-up.

Gate arrays, standard cells, and custom logic are generally designed with a symmetric layout methodology, which is well matched to the twin-tub process, while the choice of substrate polarity is of secondary importance. However, for easier insertion of memory macrocells or BICMOS circuits, the twin-tub *n*-well technology is preferred.

There is just one type of application, where the *p* well has maintained an edge despite the trend toward the *n* well and twin tub. This is in high-density, radiation-hardened CMOS/BULK circuits. Since these circuits are more susceptible to latch-up and to NMOS field inversion, a *p*-well retro-

grade well [16,75] helps in many ways. First, the high concentration at the bottom of the well reduces the gain of the parasitic vertical *n-p-n* transistor and the attendant positive feedback. Second, since the well is shallow, the epi layer can be very thin, further contributing to latch-up suppression. This is further enhanced by an $n/n+$ substrate, because up-diffusion of the epi–substrate interface is minimal using low diffusivity *n*-type dopants, such as antimony or arsenic. Third, the NMOS field region near the interface can be doped at a higher level than with a conventional diffused well, because (1) boron segregation into the field oxide is avoided by a high-energy *p*-well implant after field oxidation, and (2) the implant range can be adjusted to position the peak at the interface. Consequently, the initial NMOS field threshold voltage can be high enough to compensate for the large negative shifts induced by ionizing radiation.

With the recent introduction of MeV implanters, retrograde *n*-well processes have been demonstrated [76], opening the way to doubly retrograde twin-tub processes [77]. Since these will retain the high radiation tolerance of the *p*-well retrograde process, and in addition will offer more flexibility on the choice of substrate polarity and further increase the layout density, we expect them to become a trend setter for submicron CMOS in the high-performance, limited volume IC market, which serves system needs, such as aerospace applications.

In addition to retrograde wells and epitaxial substrate, trenches have been introduced to improve well isolation and increase layout density [24]. These trenches are built along the well border, while retaining field-oxide isolation around the active areas to avoid sidewall leakage problems, as discussed earlier. The major benefit is enhanced latch-up immunity despite minimum spacing of the active devices from the well edge. This is because of a major gain reduction of the parasitic "lateral" bipolar transistor by changing its configuration from planar to vertical and by forcing the parasitic current to flow along the trench sidewall over a corrugated path. Hence, the base width is tremendously increased and is controlled by the trench depth instead of the surface spacing. Moreover, the injected carriers are intercepted by self-aligned guard rings, which are easily formed by ion implantation at the bottom of the trench. Though well trench isolation is technically appealing, its use has been limited because of higher production costs and possible yield losses.

B. Latch-Up and Methods of Prevention

Since the invention of CMOS, latch-up has been recognized as a possible cause of catastrophic failures, justifying the thorough analysis and exten-

sive reviews of this topic [10,78]. Since it is important to avoid latch-up for long-term CMOS reliability, new methods of prevention are constantly proposed through modifications of MOS device design, process architecture, or circuit layout rules. Since these methods are intimately related to the isolation techniques, they will be reviewed in this section after a short explanation of the latch-up phenomenon.

The source of latch-up is intrinsic to the well junction isolation. As illustrated in Fig. 4.21, two parasitic bipolar transistors are easily identified in the cross section of an *n*-well CMOS structure [79]. One is a vertical *p-n-p*, which extends from a *p*+ region (emitter), across the *n*-well (base), and into the substrate (collector). The other is a lateral *n-p-n*, which is formed by an *n*+ region outside the well (emitter), the substrate (base), and an adjacent *n*-well region (collector). Unfortunately, the base and collector of these transistors are interconnected, forming a thyristor or SCR. Normally, the SCR is not conducting, because the well to substrate junction is reverse biased, forcing the SCR to be in the blocking state. However,

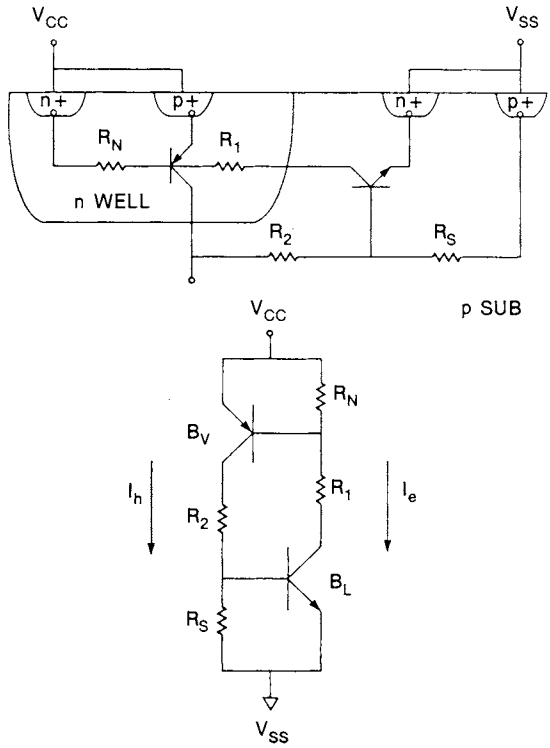


Fig. 4.21. Circuit model of the parasitic *p–n–p–n* SCR. (After Yu *et al.* [79]. Copyright 1981 by the IEEE.)

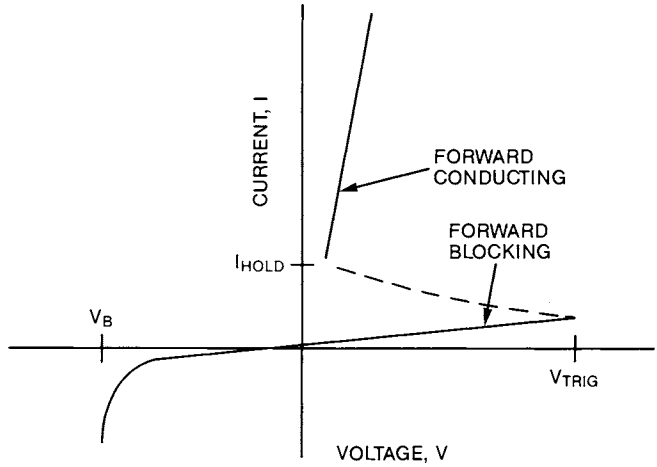


Fig. 4.22. Current–voltage characteristics for the parasitic SCR. (After Holton and Cavin [3]. Copyright 1986 by the IEEE.)

because of regenerative feedback, if conduction is triggered by some mechanism, the SCR turns on and remains in the conducting state after the triggering stimulus is removed. The SCR behavior is characterized mainly by two points, which are easily derived from the current–voltage (I–V) characteristics: (1) the triggering point, which corresponds to the maximum voltage or current allowed to keep the blocking state; and (2) the holding point, which represents the minimum voltage required to sustain the forward conducting state. Figure 4.22 illustrates the position of these points in the I–V characteristics [3].

The SCR behavior is strongly influenced both by the gain of the parasitic transistors and by the distributed resistance along the current paths. However, for simplicity, a lumped element equivalent circuit model is used for latch-up analysis, as shown in Fig. 4.21. Recently, very sophisticated SCR models have been reported to match the latch-up results of 2-D simulation programs and to study inverter circuits [80,81]. In addition, for modeling transient conditions and determining the triggering point, capacitances must be considered, and in particular, the well to substrate capacitance [82,83]. Finally, because of scaling, three-dimensional effects can no longer be neglected [84,85], since narrow-width phenomena can play a dominant role in determining the CMOS latch-up performance.

Basic conditions must be satisfied for latch-up to occur and to hold. First, according to elementary SCR theory, the product of the two parasitic bipolar gains must be larger than one to allow regenerative feedback [86]. This represents the worst case, which excludes the effect of the well and substrate resistance on the holding current. As illustrated by the lumped

element model, these resistors shunt a portion of the SCR current from the emitter-base junctions, so that only a fraction of the total holding current is active in the SCR loop. The lower the resistance, the higher will be the holding current and the holding voltage. Unless the external circuit is capable of exceeding these limits, latch-up cannot occur, thereby setting a second condition. Hence, a coveted goal of the process architect is to increase by design the holding voltage above the supply voltage specification to ensure latch-up-free operation. Finally, since latch-up turn-on is a dynamic event, a third condition is represented by the minimum pulse duration required for triggering latch-up [87,88].

Based on this description, it is understandable that the methods of eliminating latch-up fall into one of these categories: (1) reduce the gains of the bipolar parasitic transistors, (2) decrease the resistance of the well and the substrate, (3) decouple the parasitic transistors, and (4) prevent the latch-up stimulus from reaching susceptible portions of the circuit [89]. Since the last item is primarily a circuit technique, we will not dwell on it, except for mentioning that input protection circuits are very effective to screen power supply overvoltage or undervoltage, which often cause latch-up. Since these circuits occupy only a small fraction of the chip area, their design can be optimized to absorb without damage these network disturbances by utilizing larger design rules and special latch-up protection features, such as guard rings. By contrast, the other latch-up prevention techniques are built into the CMOS technology and hence are preferable because they protect every portion of the circuit against all causes of latch-up, including radiation exposure [90]. In exchange, they have to be planned in the original process design and often increase the process complexity and the manufacturing costs.

To reduce the gain of the parasitic bipolar transistors, several approaches are available, which consist of decreasing the emitter injection efficiency, reducing the minority carriers lifetime in the base region, and increasing the base width and doping level, or, in other words, the transistor Gummel number [91]. Early efforts of latch-up prevention concentrated on "killing" the minority carriers lifetime by introducing deep levels in the silicon band gap with either heavy metals diffusion, such as gold [92], or crystal lattice damage, generated by neutron irradiation [93]. Unfortunately, the reduced carrier lifetime also increased the leakage of the MOS source/drain and well junctions, creating a major deterrent to the use of this method. To some extent, similar considerations also apply to the use of oxygen precipitates for minority carriers recombination, although better control of their position is possible with internal oxygen gettering methods, leaving the junctions free from precipitates and with low leakage [94].

Emitter efficiency may be substantially reduced using Schottky barriers

at the source–drain junctions [95,96]. However, their use is restricted to PMOS devices, because only these transistors form low barrier-height junctions with the silicides normally employed in IC processing, such as $TiSi_2$ or $PtSi_2$. Unfortunately, the attendant MOSFET characteristics are badly degraded, resulting in reduced transconductance and higher leakage, thereby offsetting the latch-up benefits of the Schottky junctions and reducing the attractiveness of this technology.

Methods based on reduction of the well and substrate shunt resistance have been much more successful and are widely used in fine-dimension CMOS technology [97]. Since these resistances are in parallel with the emitter-base junctions of the parasitic bipolar transistors, the lower their values, the higher will be the current required for forward biasing these transistors. Consequently, the holding current will increase, thereby reducing latch-up susceptibility. An effective mean of reducing the substrate resistance is to use epitaxial wafers as starting material. The effectiveness of this technique is enhanced by reducing the thickness of the epi layer as far as possible without inducing vertical punch-through and by increasing the doping level of the substrate [98]. Accordingly, the use of $n/n+$ material shows enhanced latch-up resistance compared to $p/p+$, because the reduced diffusivity of n-type dopants, such as antimony or arsenic, provides a sharper epi–substrate interface by bringing the heavily doped substrate closer to the active devices [72]. The p-well advantage is confirmed in Fig. 4.23, which compares measurements of critical (triggering) current for n- and p-well CMOS, with and without the epi layer [72]. On the other hand, as previously discussed, the choice of p well is unsuitable for many applications. Moreover, the carrier lifetime has been found to be consistently higher in $p/p+$ material [99], which is an important advantage for circuits using dynamic charge storage because of their low leakage requirements. These conflicting considerations are a typical example of the trade-offs required in CMOS process design, since it seldom happens that all the desirable electrical parameters are associated with a single process option.

Reducing the well resistance is another useful tool for preventing latch-up. This cannot simply be obtained by increasing the total dose of the dopant in the well followed by thermal redistribution. Under these conditions the surface concentration would exceed the optimal MOS channel doping level, which is defined by threshold voltage requirements. Hence, the well profile must peak near the bottom of the well, generating a so-called "retrograde" shape, which may be implemented either by means of high-energy implants [16,100] or buried layers [101].

Layout techniques have been found effective for latch-up reduction. A major benefit is derived by placing the substrate and well contacts closer to the well edge than the source/drain regions. In other words, the contacts to

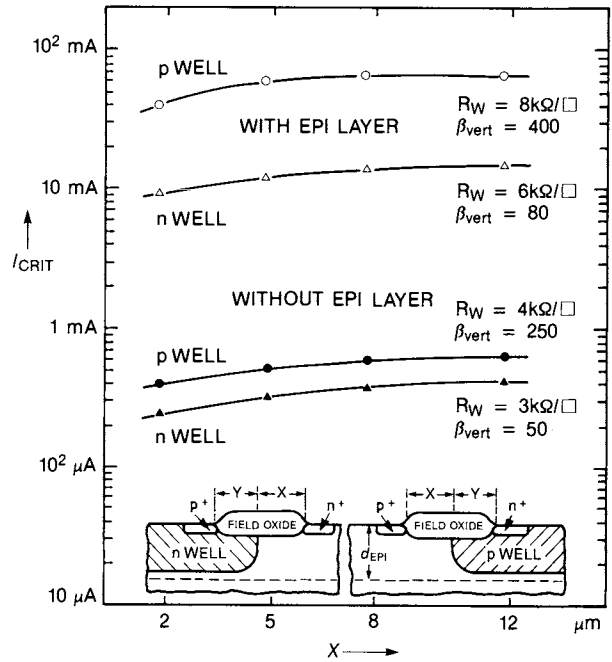


Fig. 4.23. Critical (triggering) current I_{crit} as a function of emitter to well spacing x for n-well and p-well CMOS with and without epi layer at $V_{GS} = 0$. (After Takacs *et al.* [72]. Copyright 1983 by the IEEE.)

the base of the vertical and lateral bipolar transistors should be inserted, whenever possible, within the distance between the corresponding emitters and the well edge [102]. With this configuration, any current across the well junction would pass through the well and substrate contacts without contributing to the emitter-base bias, since the emitters are located outside the current path. If the substrate/well contacts are extended all around the active devices, then they form guard rings [10,103], which have often been used in the past for latch-up protection, but are no longer popular because of loss of layout density.

An elegant variant of this technique is to merge the substrate (or well) contact with the MOS source region, forming a butted contact oriented with the drain furthest away from the well edge, as shown in Fig. 4.24 [104]. This technique is also useful to improve layout density, but it is questionable if it can be extended to 0.8-μm contacts or below, because it requires a rectangularly shaped contact and a precise alignment of the $n+$ and $p+$ select masks to ensure that both polarities appear in the butted contact. Specifically, this violates the trend of allowing in submicron ICs

only one equal-size, square-shaped contact to avoid variable and unpredictable pattern distortions during lithography, which may cause design rules violations. Moreover, improved latch-up immunity can be achieved by supplementing the topside contact with a backside contact, particularly in the case of an epi layer on a heavily doped substrate, because the current injected through the vertical transistor would flow directly into the substrate and would bypass the lateral transistor [102].

The spacings between the terminal regions and the well edge, either on the well side or the substrate side, have a major effect on latch-up, although this effect is of opposite type depending if the source/drain regions or the well/substrate contacts are involved. In the source/drain case, the larger these spacings, the higher is the latch-up hardness. This is not only due to a decrease of the lateral transistor gain but also to reduced surface leakage of the field-oxide transistors as short-channel effects are eliminated by larger spacings [98]. Figure 4.25 shows the characteristic increase of holding voltage with $n+$ to $p+$ separation across the well [104]. On the other hand, a larger distance between the well/substrate contacts and the well edge (or the source/drain regions) reduces latch-up hardness because of a related increase of lateral resistance and voltage drop along this distance, with the attendant possibility of forward biasing the emitter-base junction of the parasitic bipolar transistors and triggering latch-up. This is demonstrated

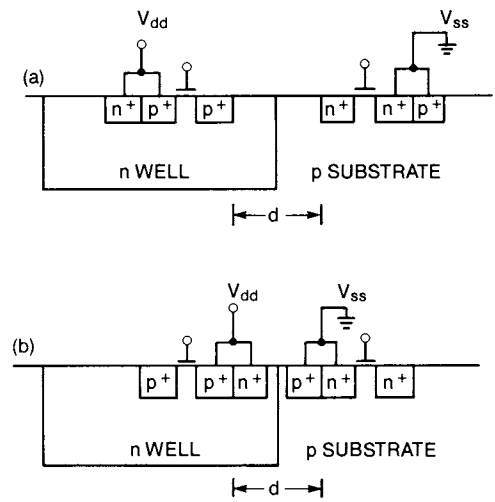


Fig. 4.24. Cross section of latch-up structure with butted source/well (or substrate) contact: (a) with well/substrate diffusions as far away as possible from well edge; and (b) with well/substrate diffusions closer to the well edges than other diffusions. The last configuration increases latch-up hardness. (After Hu and Bruce [104]. Copyright 1984 by the IEEE.)

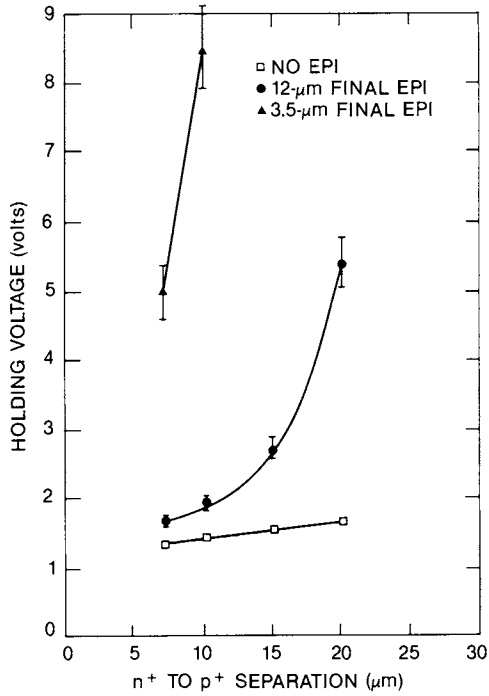


Fig. 4.25. Measurements of holding voltage versus n+ to p+ separation for various epitaxial layer thicknesses. (After Hu and Bruce [104]. Copyright 1984 by the IEEE.)

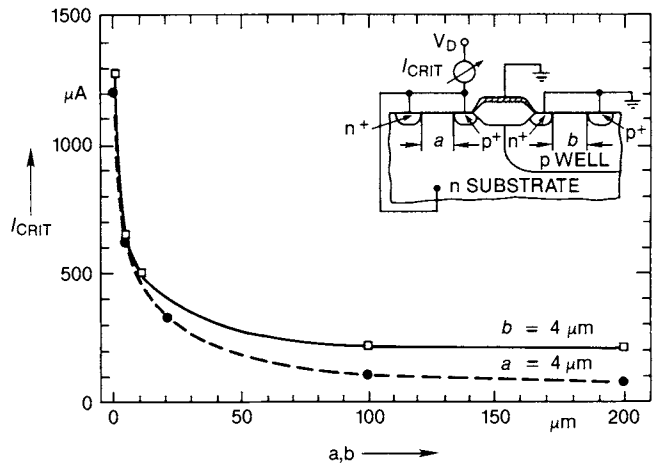


Fig. 4.26. Critical (triggering) current as a function of substrate, *a*, and well, *b*, contact spacing for *p*-well CMOS technology. (After Takacs *et al.* [72]. Copyright 1983 by the IEEE.)

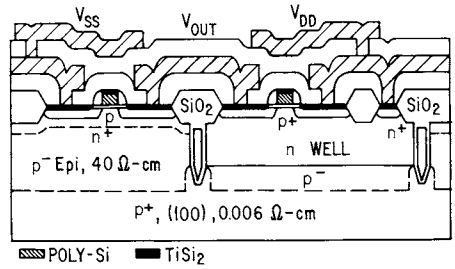


Fig. 4.27. Device cross section of advanced *n*-well CMOS process with deep-trench well isolation. (After Yamaguchi *et al.* [105]. Copyright 1984 by the IEEE.)

in Fig. 4.26 with measurements of triggering current as a function of well/substrate contact spacing from the well edge [72].

Since high density has become a major objective of advanced CMOS technology, many new approaches have been tried for decreasing NMOS to PMOS separation without compromising latch-up hardness. Without resorting to full dielectric isolation, which will be discussed later, the use of deep trenches at the well edges has provided excellent results [24,105]. A device cross section of an advanced CMOS inverter cell made with this technology is shown in Fig. 4.27. A comparative test of latch-up and electrical isolation, conducted on identically processed wafers except for the difference of a 6-μm-deep trench, has demonstrated that without the trench the isolation failed with an $n+$ to n-well separation of 4 μm, while with the trench this distance could be safely reduced below 3 μm [105]. Moreover, if the deep trench was combined with an epitaxial layer, the devices were unable to latch up even with a trigger current of 200 mA [24].

C. Twin-Tub Technology

The concept of twin-tub CMOS and its advantages for high-resolution VLSI have been presented earlier in the introductory remarks on well isolation. The fabrication aspects will be described next with reference to Fig. 4.28, which shows the sequence of device cross sections in the original CMOS twin-tub implementation [69].

Assuming a p-well process, the starting material consists of $n/n+$ epitaxial wafers. A sandwich of SiO_2 and Si_3N_4 is formed over the p tub to selectively mask it during the n-tub phosphorus implant and the subsequent thick oxide growth, for example, 5000 Å. The purpose of this oxide is to form a complementary mask in order to implant the p tub with boron using a self-aligned technique. To further enhance masking selectivity during this implant, the nitride layer is removed before the p-tub implant

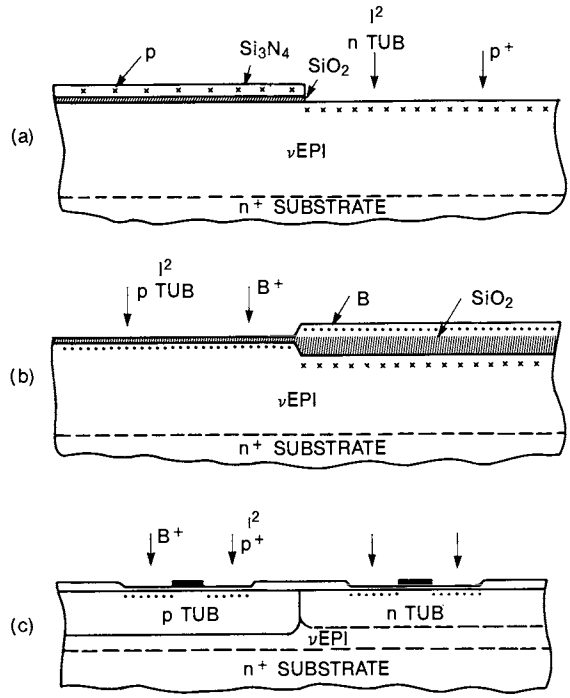


Fig. 4.28. Device cross section during the twin-tub fabrication sequence. (After Parrillo *et al.* [69]. Copyright 1980 by the IEEE.)

in order to use a lower implant energy, which is more readily stopped by the thick oxide over the *n* tub.

The tub implants are then thermally driven in by diffusion, forming wells with deep gaussian profiles, but nearly vertical lateral junctions, due to joint diffusion of one lateral profile into the other from opposite sides of the tub and concomitant impurity compensation.

After twin-tub formation, the surface is stripped of remaining oxide to start the active area isolation, followed by conventional CMOS processing. Fortunately, even on a bare Si surface, the twin-tub sequence leaves an edge mark along the well border, which is needed for the next mask alignment. This edge originates from the conversion of Si into SiO_2 during selective thick oxidation, which affects only the unmasked tub. However, if the step height of this mark is too small for quick recognition by automatic alignment systems, a separate alignment mark must be engraved in silicon at the beginning of the process through the use of another masking step, which is preferable to avoid for simplicity and cost reasons.

As mentioned, in the twin-tub process, the two tubs laterally diffuse into

each other, causing possible isolation problems with field inversion, lateral punch-through, and latch-up [9,106]. Figure 4.29 illustrates the net doping fall-off on both sides of the well as a function of lateral distance from the well edge [69]. The slope of these curves is sharper than for a conventional p-well, but since the doping change occurs on both sides of the junction compared to a single side for the p well, these two effects compensate each other so that both technologies require similar spacings between $n+$ and $p+$ across the well. However, in the twin tub, the well border is located close to its drawn position, while in the single well, it is displaced by the well lateral diffusion, which must be compensated either by design rules or mask biasing. Lately, a more abrupt doping transition between the two tubs has been realized by (1) adding arsenic to phosphorus in the n-tub implant to ensure a higher net n doping near the well edge because of lower arsenic diffusivity compared to phosphorus; (2) adding a second shallow

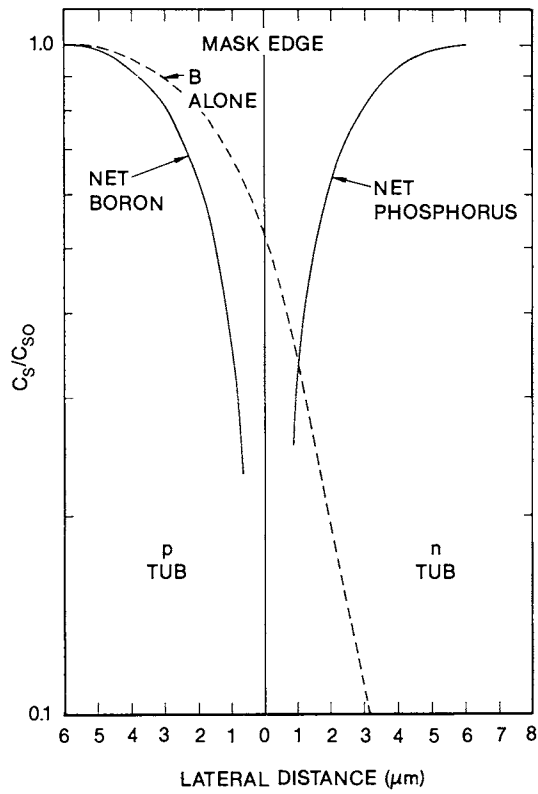


Fig. 4.29. Lateral diffusion profiles of boron and phosphorus at the twin-tub border. (After Parrillo *et al.* [69]. Copyright 1980 by the IEEE.)

boron implant to the p tub after well drive-in to increase the surface concentration near the tub edge and henceforth compensate for the loss of boron due to lateral diffusion; and (3) using high-pressure oxidation for field-oxide growth since it can provide the desired field-oxide thickness with a smaller thermal budget and consequently less dopant redistribution [107].

The field inversion and punch-through leakage near the tub borders can be associated with parasitic transistors, as shown schematically in Fig. 4.30 [106]. The isolation characteristics of these transistors are usually determined as a function of the distance from the active area to the tub edge, as illustrated in Fig. 4.31 [107]. In these plots, the ordinate scale represents the voltage required to induce a predefined maximum source/drain leakage current in these transistors, for example, 1 pA/μm width. To include in this test the effect of an energized polysilicon or metal wire crossing the field isolation spacing, the drain voltage is also applied to the field-gate electrode during these measurements. This is illustrated in the testing diagrams, shown in the insets of Fig. 4.31. Observing these curves, it is easy to derive the minimum distance needed for safe operation at the supply voltage, usually V_{DD} plus 10% safety margin. However, since this distance also affects latch-up, the final design rule must exceed both the value obtained from this plot and the one necessary for avoiding latch-up.

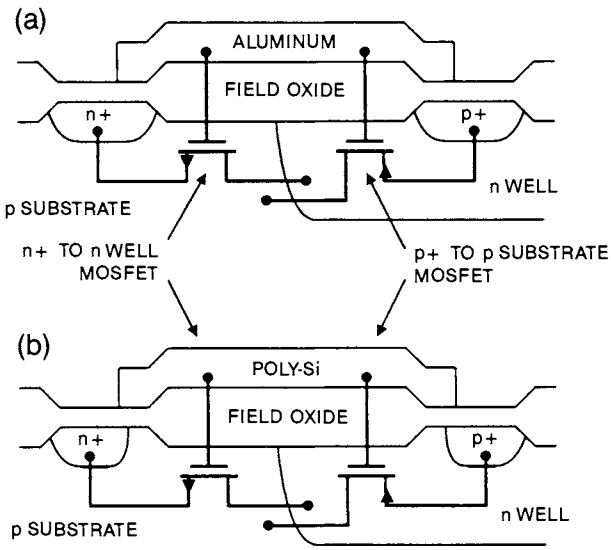


Fig. 4.30. Parasitic MOS transistors formed at n-well edge: (a) aluminum gate; (b) polysilicon gate. (After Lewis *et al.* [106]. Copyright 1987 by the IEEE.)

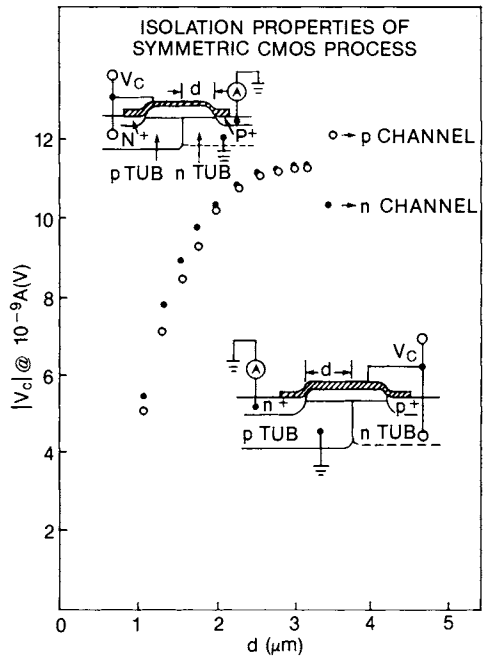


Fig. 4.31. Isolation characteristics of tub-edge parasitic transistors as a function of the distance from tub edge to active area edge. (After Hillenius *et al.* [107]. Copyright 1986 by the IEEE.)

The interdependence of doping profiles, thermal cycles, and device dimensions requires a careful optimization of these parameters during process design. To avoid many experimental runs for finding the optimal configuration, two-dimensional process and device modeling has been successfully utilized following a procedure described here [9]. The first step is to use a two-dimensional process simulator for finding the isoconcentration impurity contours of the isolation structure, as shown in Fig. 4.32. Inserting these results into a two-dimensional device analyzer, various potential contours can be plotted for the bias conditions required to analyze field inversion or lateral punch-through.

For the field inversion analysis, it is useful to plot the surface potential versus lateral distance under the field-oxide isolation at various gate voltages, V_G, as shown in Fig. 4.33. Notice that this figure corresponds to the layout of Fig. 4.32, where the left side of the field oxide overlaps the p tub, while the right side overlaps the n tub. Since field inversion occurs if the minority carrier density exceeds $\approx 1 \times 10^{11}\ cm^{-3}$ (corresponding to a surface potential of 51 mV at 25°C), then a safe operational range for this structure with 2-μm spacing is $-7.0\ \text{V} < V_G < 21.4\ \text{V}$. This can be visually

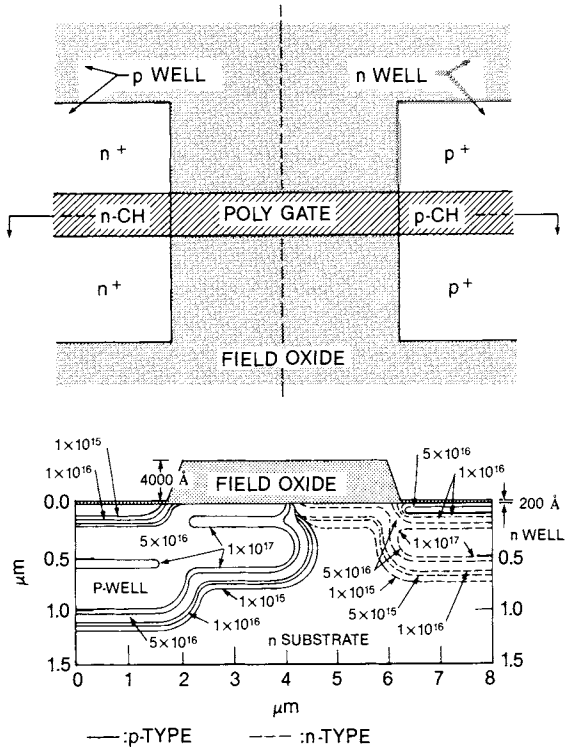


Fig. 4.32. Planar and cross-sectional view of CMOS isolation region with 2-D isoconcentration impurity contours. (After Chen and Snyder [9]. Copyright 1986 by the IEEE.)

derived from Fig. 4.33 by reading the V_G label of the contours, whose maximum (minimum) potential is close to zero over the n-tub (p-tub) section of the graph.

For studying lateral punch-through, the voltage applied to the well junction and to the neighboring source/drain regions must be considered in the simulations. Moreover, the effect of the field-gate voltage must also be included, since it can indirectly affect punch-through by controlling the underlying surface potential and carrier density. An example of these simulations is given in Fig. 4.34, which shows 2-D equipotential contours for 2-μm isolation with 1.5-μm $p+$ to p-tub spacing and 0.5-μm $n+$ to n-tub spacing. The n tub is tied to the supply voltage, 5V, while the p tub and the substrate are at ground. The graph at the top shows good electrical isolation with a gate voltage of 10 V due to the undepleted portion of the p-tub surface, which is characterized by zero surface potential. However, if the gate voltage is raised to 20 V, then punch-through occurs from $n+$ to n tub, because the p-tub surface becomes fully depleted or inverted, as

demonstrated by the fact that it is at positive potential. By analyzing a matrix of bias conditions and spacings, the isolation design can be optimized even before processing a single silicon lot with great savings of development costs and time.

D. Retrograde Wells

Though the twin-tub approach has brought the advantages of independent optimization of NMOS and PMOS devices coupled to a symmetric treatment of both wells, the density has not improved because of the long well drive-in and consequent lateral diffusion. Moreover, the problem of boron segregation during field oxidation was left unsolved, causing low-field threshold voltage on the p tub unless the surface concentration was enhanced with an additional shallow boron implant after drive-in. But, besides the complication of an extra step, this implant resulted in lateral boron diffusion at the NMOS channel edges, limiting the effective channel width and presenting another obstacle to scaling. In addition to seeking a

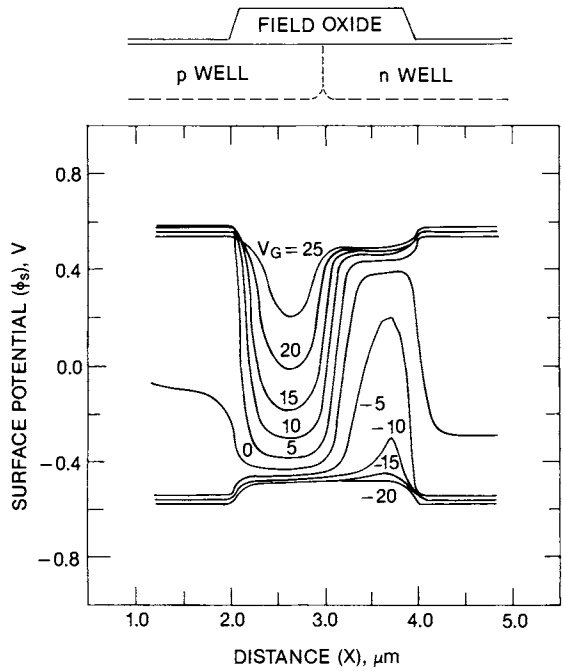


Fig. 4.33. Surface potential for 2-μm isolation region versus lateral distance as a function of gate voltage V_G. (After Chen and Snyder [9]. Copyright 1986 by the IEEE.)

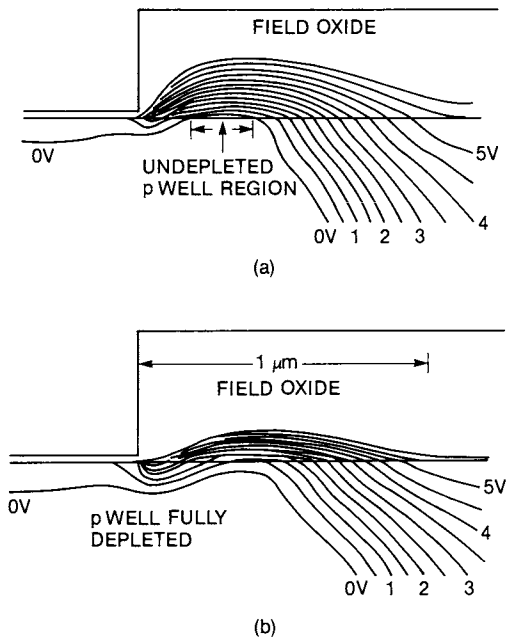


Fig. 4.34. Two-dimensional equipotential contours for 2-μm isolation with 1.5-μm $p+$ to p-tub spacing and 0.5-μm $n+$ to n-tub spacing. $V_{n\,tub} = 5$ V, $V_{p\,tub} = V_{Sub} = 0$ V. (a) No punch-through at $V_G = 10$ V. (b) Punch-through at $V_G = 20$ V. (After Chen and Snyder [9]. Copyright 1986 by the IEEE.)

solution to these problems, there was also the need to improve latch-up. This implied the use of a shallow well to reduce the epi layer thickness and bring the heavily doped substrate closer to the active devices. Moreover, to lower the well sheet resistance with reduced junction depth, the well profile could no longer be gaussian, since this would have resulted in a high surface concentration, which is incompatible with the gate threshold voltage of ≈ 0.8 V. The answer to all these requirements came in the form of the retrograde well approach [16,75,77,100].

Retrograde wells are formed by high-energy ion implantation after the active area isolation is already in place, as illustrated in Fig. 4.35 [16], thereby avoiding the problem of impurity segregation during field oxidation, which occurs in conventional CMOS. After the retrograde implant, the thermal treatments are reduced to a minimum in order to retain the original implant profile as much as possible. For instance, instead of the long drive-in cycle at high temperature, only a short anneal at relatively low temperature is used to activate the implant. Among other reasons, the low thermal budget is necessary to avoid upward diffusion from the profile

peak toward the surface, because it could interfere with the MOS threshold adjustment. For the same reason, a high-energy implant must be used to lengthen the implant range and minimize the attendant doping increase near the active area surface. It is estimated that, in order to avoid excessive spread of the electrical device parameters, this concentration should be at least five times smaller than its final value, which is set later by shallower, low-dose implants, specifically designed for threshold and punch-through control. The above-mentioned spread would principally affect threshold voltage, source to drain leakage, and hot electron lifetime.

The simulated concentration profiles of a retrograde *p*-well process are presented in Fig. 4.36 [16], where the upper graph shows the NMOS channel (solid line) and source/drain (dotted line) profiles, while the lower graph provides the *p*-tub field profile. Notice that the peak of the field profile is near the oxide–silicon interface, since the field oxide absorbs a large fraction of the implant energy. This is very beneficial, because a large surface impurity concentration results in high-field threshold voltage and low surface leakage, acting as a self-aligned guard ring. The fact that the NMOS source/drain profile intersects the channel profile away from the peak, at lower concentration, is important to prevent an increase of junction capacitance and, consequently, of propagation delay, compared to conventional CMOS.

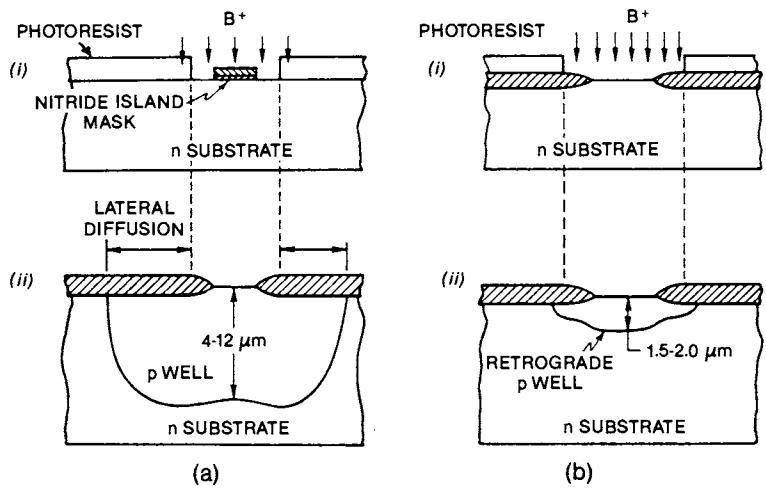


Fig. 4.35. Comparison of *p*-well CMOS fabrication between conventional and retrograde processes. (a) Conventional method: (i) *p*-well implant; (ii) after drive-in and field oxidation. (b) Retrograde method: (i) *p*-well implant with field oxidation already in place; (ii) after short thermal activation anneal. (After Rung *et al.* [16]. Copyright 1981 by the IEEE.)

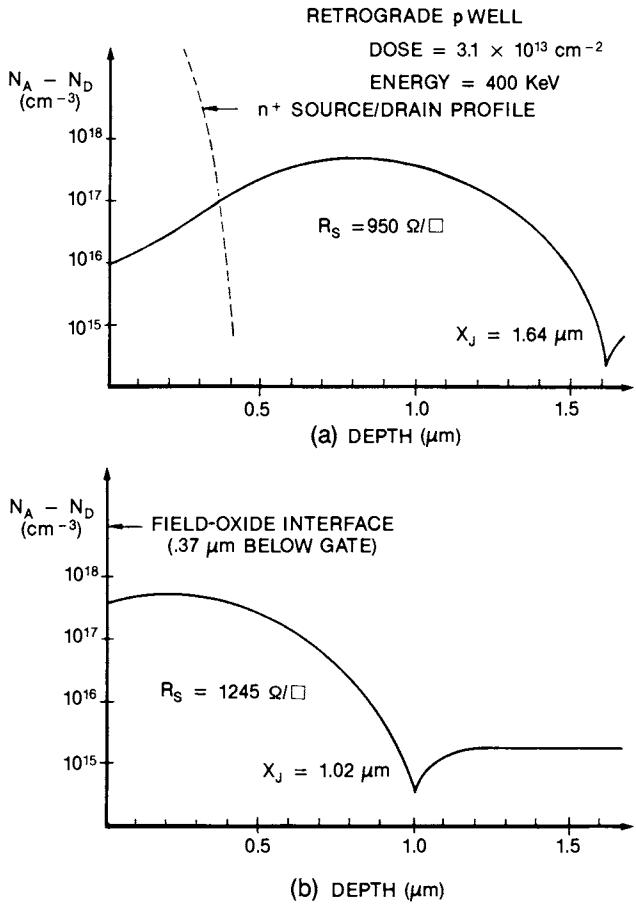


Fig. 4.36. Simulated net doping concentration profiles for retrograde *p*-well process in the *p*-tub region. (a) NMOS channel (solid line), NMOS source/drain (dotted line); (b) NMOS field. (After Rung *et al.* [16]. Copyright 1981 by the IEEE.)

The major benefit of retrograde wells is a significant increase in density compared to diffused wells because of the elimination of the well lateral diffusion. Considering that the lateral diffusion is about 0.7 times the junction depth, a typical diffused well of 3 to 4 μm results in a lateral spread of ≈ 2 to 3 μm. By contrast, the lateral implant straggle is ≈ 0.7 μm and barely exceeds one micron after all the thermal processing [16]. Moreover, the small radius of curvature of the implanted well edge leads to a narrower space charge region at the surface, further reducing the isolation spacing of the well from adjacent diffusions. Hence, for a 1.2-μm CMOS

process with a retrograde p well, the $n+$ to $p+$ spacing across the well can be as small as 5 μm compared to 8–9 μm for a diffused twin-tub process.

Although the retrograde process was originally used only for p-well technology because of the higher penetration power of boron ions and the limited energy range of existing implanters [16,75], it has since been successfully employed for both wells in conjunction with the twin-tub approach [108]. This seems to be a growing trend for submicron CMOS because of the wider availability and production worthiness of MeV implanters in recent times. On the other hand, the alternative method of using double-charged species with medium-energy implanters can give rise to problems, because it has been found that a large fraction of double-charged ions, as high as 20%, can interact with gas molecules at the entrance of the acceleration column and become single-charged ions because of partial neutralization. These ions would only penetrate in silicon about half of the desired range, distorting the profile and causing abnormalities of the threshold voltage and other electrical parameters. Thus, for good agreement with the process design, if a MeV implanter is not available, at least the medium-energy implanter should be of the post-analysis type to eliminate single-charged ions before they reach the wafers [108]. With these precautions, retrograde well CMOS processes have now been transferred to production with excellent results, particularly for custom logic and advanced system applications, since these technologies can absorb the higher costs of retrograde well processing in exchange for higher density and performance.

IV. DIELECTRIC ISOLATION

A. Features and Applications

There are some CMOS applications where a combination of field-oxide isolation and junction isolation is not sufficient. The solution in these cases is full dielectric isolation, which is often implemented with silicon-on-insulator technology [4,5]. This technology is appealing when total isolation is required among active devices and between them and the substrate. This is an important factor for CMOS operation in harsh environments, resulting from transient radiation pulses, high temperature, or high voltage. All these conditions determine a rapid increase of junction leakage current, which eventually leads to excessive power dissipation and, perhaps, to latch-up [10,109,110].

Though at present this may be the premier reason for continued development of silicon-on-insulator technology, historically other justifications were given that are still regarded as long-term prospects. The major contender with CMOS/SOI has always been CMOS/BULK with both technologies coexisting side by side nearly from their inception in the early 1970s [111]. At that time, the only viable SOI technology was SOS [112], which enjoyed large popularity and was regarded by many proponents as the technology of choice for high-performance, high-density applications based on the arguments described next [113].

Since, in SOI, silicon islands are formed over an insulating substrate, there is no need for well isolation and attendant space charge regions to isolate complementary devices. Hence, the spacing between adjacent islands can be as close as allowed by lithographic and etching capabilities, resulting in potentially higher density. On the other hand, CMOS/BULK supporters contend that the well isolation spacing between *p*- and *n*-type islands is not wasted, since it is utilized for interconnections, and therefore the density difference in actual circuits is minimal, as demonstrated by laying out the same circuits in both technologies.

Higher performance is another recurring claim of SOI. This is based on the reduction of parasitic capacitance compared to BULK, which results in smaller propagation delay at the gate level and along interconnections. However, in the case of SOS, the advantage has been reduced with downscaling, because the interconnection to substrate capacitance has become a smaller fraction of the total capacitance because of the reduction of the conductor's width to thickness ratio. As a consequence, the interline capacitance has become more prominent, reducing the benefit of an insulating substrate. At the gate level, SOI exhibits smaller parasitic capacitance than BULK, because the source and drain regions interface directly with the insulating substrate, so that their capacitance is mainly due to the lateral junctions at the channel edges. On the other hand, BULK has the advantage of higher current drive because of higher electron mobility in bulk silicon, though the difference has been reduced recently with new treatments of the epitaxial silicon film in SOS wafers [114,115]. To evaluate the relative importance of these conflicting effects on speed between these technologies, experimental 1-μm CMOS/SOS and CMOS/BULK devices and basic circuits were made using nearly the same process. The propagation delay of a chain of inverters was measured and was found to be 60% lower for SOS than BULK [116]. Of course, the validity of this comparison is tempered by the fact that a process optimized for SOS might yield less than optimal results for BULK, suggesting that the real comparison between these technologies can only be done on competing products.

In addition to improvements in circuit performance and packing den-

sity, SOI advocates claim major simplifications in the CMOS fabrication process [113]. Since the well (or tub) is not required, the corresponding mask is eliminated together with the thermal processing needed to drive in the well. Field oxidation is also optional, since device isolation is already provided by the substrate. A typical CMOS/SOS process for a 2-μm self-aligned polysilicon gate is shown in Fig. 4.37 [117].

At the beginning of the process, silicon islands are defined by the active area mask, followed by silicon etching to create electrically isolated mesas. Anisotropic etching is preferred to form sloped sidewalls, which reduce step coverage problems for interconnections. After channel doping, which is done separately for NMOS and PMOS transistors, the CMOS/SOS process very closely resembles a corresponding BULK process. However, every single operation in SOS requires more care, as will be demonstrated

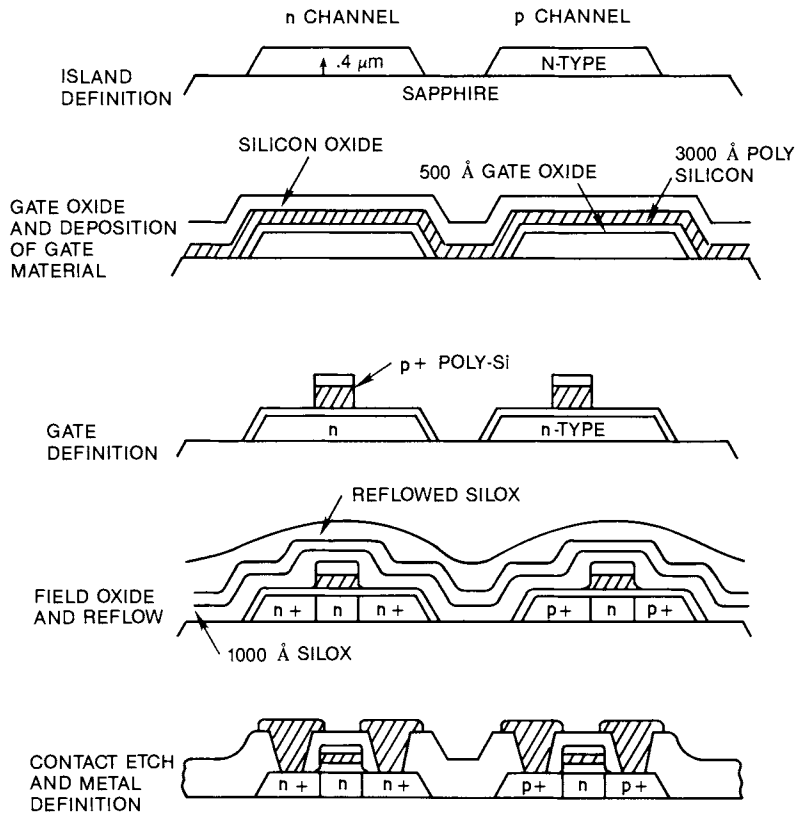


Fig. 4.37. Fabrication sequence for 2-μm CMOS/SOS process. (After Splinter [117]. Copyright 1978 by the IEEE.)

for specific operations: (1) longer ramping time for thermal treatments because of a difference in thermal expansion coefficients of sapphire and silicon; (2) careful handling, since sapphire is more fragile than silicon; (3) antireflective coatings for the support pedestal of optical exposure stations because of sapphire transparency; and (4) use of an electron flood gun in ion implanters to avoid charge buildup because of the electrical isolation from the pedestal. Some of these problems are avoided with SOI techniques using silicon as a substrate, but these techniques are not yet mature for production. Hence, claims of manufacturability advantages for SOI appear premature at this stage.

There is an area of SOI that offers exciting and unique opportunities. This is the area of three-dimensional integrated circuits. Since polysilicon can be converted to single-crystal silicon by zone melting crystallization without seeding, active devices can be fabricated over the substrate in a multilevel structure. Applications abound where 3-D ICs can simplify the chip architecture, particularly if parallel signal processing is required [11]. A three-dimensional SRAM has been fabricated with double active layers using the laser recrystallization technique [12]. This demonstrated the potential savings of the die area, since the decoders, sense amplifiers, and input/output buffer circuits were stacked on the upper silicon layer, leaving all the underlying area available for memory cells. Figure 4.38 illustrates this approach. More recently, a DRAM has also been implemented in a triple-layer structure using laser recrystallized polysilicon [118]. The purpose was to demonstrate a novel approach for solving the recurrent problem of down scaling DRAMs, which consists of retaining a high storage capacitance while reducing the memory cell size. The leakage current per cell, due to the access transistor and stacked capacitor, measured less than 0.1 pA, indicating excellent silicon lattice quality. Consequently, the refresh cycle time was in excess of a millisecond. These examples demonstrate the vigorous research now underway in 3-D ICs.

B. Techniques of Dielectric Isolation

As a consequence of the great interest recently placed in dielectric isolation, many new techniques have been developed with different characteristics. A useful classification [4] distinguishes three main categories according to Fig. 4.39:

(a) silicon layer/insulating substrate
(b) silicon layer/insulator/silicon substrate
(c) restricted silicon layer/insulator/silicon substrate

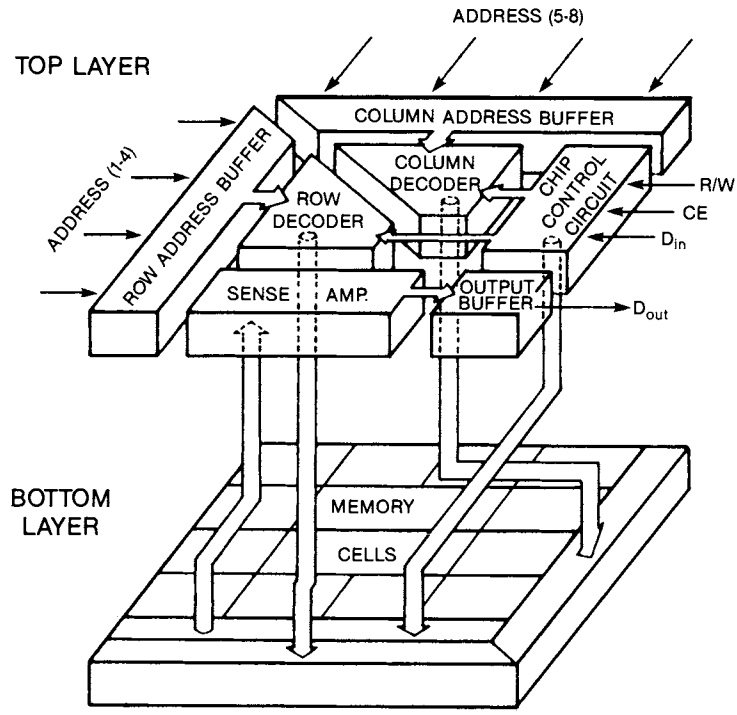


Fig. 4.38. Circuit diagram of a 3-D static RAM. (After Inoue *et al.* [12]. Copyright 1986 by the IEEE.)

In the first category, the entire substrate is made of insulating material several hundreds micrometers thick to provide both dielectric isolation and mechanical support. Sapphire is most widely used, but in the past, other materials, such as spinels, have also been employed. Then, a thin silicon layer usually in the range of 0.2 to 0.6 μm is deposited by heteroepitaxy for use in device fabrication.

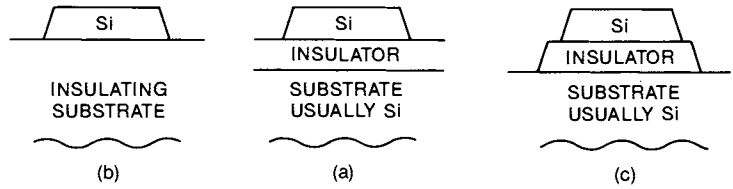


Fig. 4.39. Contending techniques for SOI. Classes of SOI substrate. (a) Thick insulator (> 1 μm); (b) Thin insulator ($\lesssim 1$ μm); (c) Restricted structure. (After Partridge [4]. Copyright 1986 by the IEEE.)

In the second category, SOI wafers are produced starting from normal silicon wafers. This ensures compatibility with automated manufacturing equipment made for silicon bulk processing. The SOI structure is then completed by forming a single-crystal silicon film over an insulating layer, with both films being $\approx 0.5-1.0$-μm thick. Many techniques may be employed, such as oxygen ion implantation, zone melting recrystallization, and back-to-back wafer bonding and etching.

The third category includes methods where restrictions are imposed on the size of the silicon islands, their mutual spacing, and their placement on the wafer. These restrictions are intrinsic with these methods and are due either to avoidance of imperfectly crystallized regions, for example, silicon boundaries, or to lateral oxidation limits for pinching off tiny surface portions of silicon to form islands.

1. Silicon Layer over Thick Insulating Substrate

The most mature technology in this category is represented by SOS [112], since CMOS/SOS circuits are commercially available, though they are more expensive than BULK. The major problem is the formation of high-quality epitaxial silicon films over sapphire with electrical properties similar to BULK. However, neither material quality nor higher cost have deterred the use of SOS in military applications, since the primary consideration has been the excellent radiation hardness characteristics of SOS.

In conjunction with the downscaling trend, there has been a progressive reduction in SOS of silicon epi layer thickness from 0.6 μm in 1979 [122] to ≈ 0.3 μm in 1987 [116,123]. This has been partially motivated by the desire of continuing the traditional practice in thin films MOSFETs of extending the source/drain to the silicon–sapphire interface to reduce parasitic capacitance and, hence, increase performance. However, the only way to realize this type of structure with shallow junctions is to reduce the thickness of the epitaxial silicon film. The shallow junctions are in turn required to control the channel length and avoid source/drain encroachment with dangerous consequences on punch-through leakage. Unfortunately, a thinner epi layer reduces the effective *n*-channel mobility, since the microtwin defect density increases rapidly in proximity of the sapphire interface [114]. This is confirmed by Table 4.1 [116], which shows low-field *n*-channel mobility of only 281 cm^2/V sec in 0.3-μm-thick SOS compared to 520 cm^2/V sec in BULK. On the other hand, the *p*-channel mobility is scarcely affected by SOS, independently of silicon film thickness.

To overcome *n*-channel mobility degradation with scaling, various techniques of solid-phase recrystallization were developed to provide high-

TABLE 4.1
Low-Field Carrier Mobilities[a]

Substrate	*n* channel	*p* channel	Units
0.3-μm silicon on sapphire	281 ± 12	193 ± 4	cm^2/V sec
0.4-μm silicon on sapphire	296 ± 23	199 ± 3	cm^2/V sec
Bulk silicon	520	200	cm^2/V sec

[a] After Brassington *et al.* [116].

quality, thin (0.3 μm or less) SOS films [123,124,125], figuring prominently among them solid phase epitaxy and regrowth (SPEAR) [126] and double solid phase epitaxy (DSPE) [127]. Both processes employ heavy dose silicon implants to amorphize the silicon layer near the sapphire interface and then use solid phase epitaxy (SPE) at low temperature to recrystallize the entire layer from the surface down, utilizing the relatively undamaged surface as a seed. The difference is that in SPEAR the initial layer is very thin and after recrystallization the final silicon thickness is obtained by epitaxial CVD deposition of silicon. On the other hand, in DSPE, a second silicon implant is used after recrystallization of the bottom layer to amorphize the surface layer, which is then recrystallized from the bottom up. The enhanced mobility of SPEAR–SOS compared to as-grown SOS is demonstrated by a family of current–voltage characteristics for NMOS and PMOS devices of $W = 100$ μm and $L_{eff} = 0.8$ μm, as shown in Fig. 4.40.

Though SPEAR and DSPE have greatly improved the quality of thin silicon SOS, they have worsened another well-recognized SOS problem, namely, back-channel leakage [128], which is caused by electrically active defects in the sapphire near the interface. Since these defects are positively charged, they induce a negative image charge on the silicon side of the interface through accumulation of electrons and formation of a parasitic back-channel, which escapes control of the gate because of its remote location. As a result, enhancement-mode NMOS FETs in SOS exhibit higher subthreshold leakage than in BULK, even after other process-related leakage components, for example, island sidewall leakage, have been eliminated. Unfortunately, the mobility improvements of SPEAR and DSPE enhance the back-channel effect in comparison to conventional SOS, since higher mobility implies also higher current regardless of its nature. Actually, in conventional SOS, the large mobility gradient across the Si film favored the surface channel in comparison to the back-channel, while this advantage is lost in the advanced SOS material because of the elimination of this gradient.

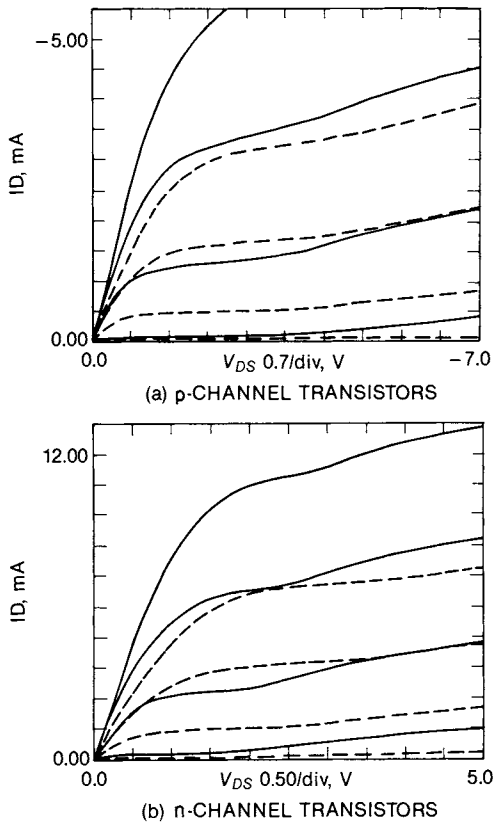


Fig. 4.40. Current–voltage characteristics for NMOS and PMOS devices fabricated in as-deposited (--------) and SPEAR (——) SOS wafers. (After Vasudev [123]. Copyright 1987 by the IEEE.)

Since the material improvements in SOS over more than a decade have eliminated some problems, but have not helped others, there is a growing perception that this technology has leveled off, and radically new concepts are needed to provide a comprehensive solution to dielectric isolation for the next decade. These novel SOI approaches will be presented next.

2. *Silicon Layer/Thin Insulating Layer/Silicon Substrate*

Of the three categories of SOI technologies, this one appears to be the most promising, because technologies adopting this structure are immune from the problems associated with a thick insulating substrate, for example, SOS, and are free from layout restrictions, since the silicon film extends uninterrupted over the entire wafer surface. These attractive fea-

tures and the possibility of application to 3-D circuit fabrication have accelerated the development pace of this technology, leading to the successful demonstration of large circuits [12,119,129]. Typically, the thin insulating layer separating the silicon film from the silicon substrate consists of silicon dioxide, but recently CaF_2 has also been used with equally good results [130].

Next, we will review the most popular implementations of this dielectric isolation approach, including SIMOX (separation by implanted oxygen) [119], ZMR (zone melting recrystallization) [113,120,121], and SDB (silicon wafer direct bonding) [131,132,133].

a. SIMOX. In SIMOX a buried oxide layer is produced by a deep, heavy dose, oxygen implant [119], as illustrated in Fig. 4.41. Though the silicon surface is heavily damaged by this implant, it retains enough crystallinity to permit nearly complete restoration of the original lattice with thermal annealing. Though SIMOX was proposed in 1978 by Izumi *et al.* [134,135], there have been practical problems of implementation, particularly because of the need of new ion implantation equipment [136] for achieving reasonable throughput of SIMOX wafers with oxygen implants at 150–200 keV and ion dose of $1.2-2.2 \times 10^{18}$ ions/cm^2 [119].

Another problem was adequate retention of surface crystallinity during the oxygen implant to provide the seed for the following lattice restoration during anneal. This problem was solved by heating the wafers at $\approx 400°C$ during the implant to induce a partial anneal [137]. The crystal perfection is then fully restored either with a high-temperature anneal above 1150°C

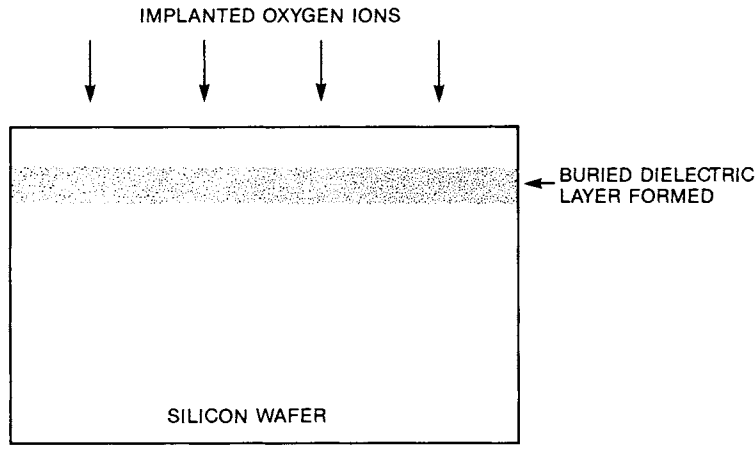


Fig. 4.41. Oxygen implantation for the formation of SIMOX wafers. (After Lam [119]. Copyright 1987 by the IEEE.)

in an inert ambient [138] or by adding to this anneal a thin silicon epitaxial layer, provided that the increased thickness is acceptable [139].

SIMOX has the advantage of conforming to the scaling needs, since it can produce controllably and uniformly thin silicon films with mobilities similar to BULK without epitaxial regrowth [140]. Moreover, since the silicon film is only 1000–2000 Å thick, the silicon region below the channel will deplete completely and avoid the problem of floating substrate and attendant "kink" in the NMOS I–V characteristics [141]. As a result, SIMOX/CMOS devices with gate length as low as 0.25 μm have been fabricated with excellent characteristics [142].

There is still a lingering problem with SIMOX that needs to be solved, and that is the high junction leakage current [143]. This is attributed to high levels of contaminants, particularly carbon and iron, which are entering into silicon during oxygen implantation. With anticipated improvements in the implanters, this problem will probably be solved in the near future, making the standby power dissipation of SIMOX/CMOS comparable to BULK/CMOS.

b. ZMR. Zone melting recrystallization produces a monocrystalline silicon film over an oxidized silicon substrate by localized polysilicon melting, followed by recrystallization [113]. Specific techniques for zone melting comprise: (1) laser scanning [144]; (2) movable graphite strip heating [145]; (3) CW Hg lamp scanning [146]; and (4) strip-window rf heating [120].

A schematic diagram of the ZMR process with graphite strip heaters is illustrated in Fig. 4.42. Typically, the process consists of growing a 0.5–1.0 μm thermal oxide over a silicon substrate. Then, a polysilicon layer is deposited by low-pressure CVD at 600°C to a thickness of 0.3–0.5 μm. For protection during zone melting, another silicon oxide layer, about 2 μm thick, is deposited by LPCVD to encapsulate the entire wafer. During processing, the wafer temperature is at first raised uniformly using the lower heater until it is 100–200°C below the silicon melting point. Next, the top graphite-strip heater is moved over the wafer, melting a narrow strip of polysilicon and sweeping it across the wafer to leave behind enlarged recrystallized grains of 20–50 μm in size [113].

Though ZMR produces silicon of excellent quality within a single grain domain, the grain boundaries and subboundaries (low-angle grain boundaries) may cause problems during device fabrication because of lattice discontinuities. For instance, enhanced diffusion along these boundaries can increase MOS leakage and create shorts between source and drain [147]. One way to reduce this anomalous diffusion is to use low-temperature processing. However, this precaution does not solve the problem of exces-

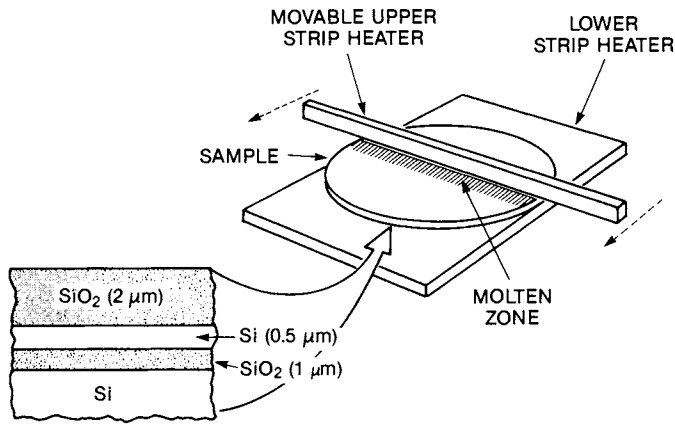


Fig. 4.42. Schematic diagram of sample and graphite strip heaters used in zone melting recrystallization of SOI films. (After Tsaur [113]. Copyright 1987 by the IEEE.)

sive spread of the threshold voltage and other MOS electrical parameters, which are sensitive to the subboundary defect density. Therefore, the long-term solution consists of either making subboundary-free films or confining the subboundaries to a restricted area, where active devices are excluded. Both approaches have been demonstrated to be feasible. The first one requires a new capping technique for protecting the SOI structure during zone melting [148]. The second one requires an array of nitride stripes to act as antireflection coating and hence reduce the temperature of the underlying silicon. As a result, the stripes become the preferred nucleation site for subboundaries, leaving a perfect silicon crystal between the stripes [149].

Devices and circuits have been made in ZMR silicon with excellent characteristics in the 1.2-μm–submicrometer regime. With ZMR RF and a 1.2-μm gate length, the subthreshold leakage current was in the range 0.01–1 pA/μm width for both n- and p-channel transistors and the propagation delay measured 130 psec at 3.5 V [120]. Similarly, with ZMR graphite strip heating and a 0.8-μm gate length, the subthreshold leakage current was <0.1 pA/μm width for both n- and p-channel devices and the propagation delay measured 95 psec at 5 V [147]. By comparison, in BULK, the propagation delay is about 20–30% higher.

Of all the SOI techniques, ZMR has the advantage that it can be easily applied to three-dimensional integrated circuits [121]. However, as the number of active layers increases, the ZMR process becomes more complicated, because the device characteristics of the underlying active levels must not be altered. This narrows the latitude of the recrystallization process, since the ZMR thermal budget must be more carefully controlled

[149]. Therefore, as of 1986, the 3-D technology for large-scale integrated circuits is available for first and second active levels but requires improvements for adding a third level of similar quality [149].

c. SDB. Wafer bonding consists of joining together two oxidized silicon wafers front to front and then etching back one of the wafers until a very thin silicon layer is left [131,132]. For yielding devices with a narrow distribution of electrical parameters, this film must have uniform thickness and doping level. For this reason, the film is often derived by etching off the heavily doped substrate of an epitaxial wafer, utilizing as an etch-stop the epi–substrate interface. The large etch-rate differential between high and low doping levels provides an excellent etch-stop, while the thickness control and uniformity of the epi layer guarantees reproduction of these characteristics in the bonded film.

To promote bonding, the surface of both wafers has to be made hydrophilic, either by soaking it in a diluted H_2SO_4 solution [132] or by growing a thermal oxide [131]. This agrees with the belief that bonding is attributable to the polymerization of silanol bonds to form a siloxane network [131]. Obviously, the wafers should be very flat (less than 3–4 μm) and free from imperfections, particularly localized depressions, which will generate gaps after bonding. The ambient properties are also critical, both with respect to cleanliness (class 2 is suggested in [132]) and gas choice. Oxygen is preferred to an inert gas, because residual oxygen can be converted into SiO_2 during annealing, thereby eliminating any gap due to trapped gas [131]. After the wafers are joined together, they are annealed at $\approx 1100°C$ for a few hours to increase the bonding strength and eliminate voids [132], often using high pressure during annealing. The process sequence is illustrated in Fig. 4.43.

Compared to other SOI techniques, wafer bonding allows a flexible choice of oxide thickness, which can be important for high-voltage applications [143]. On the other hand, it is difficult to accurately control the

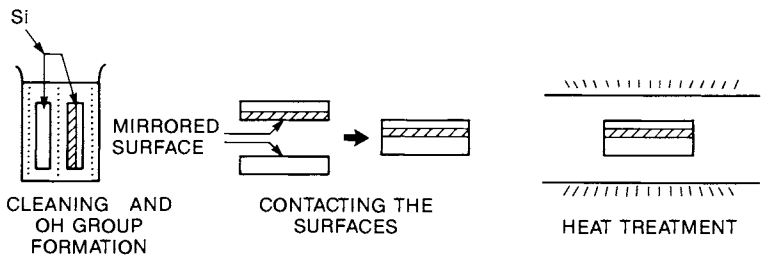


Fig. 4.43. Outline of the SDB process. (After Ohashi *et al.* [133]. Copyright 1986 by the IEEE.)

silicon film thickness below 1 μm because of the tolerance margin of the etch back process. Thus, SIMOX or ZMR appear preferable for submicrometer CMOS/SOI circuits, since they can produce films of only a few thousand angstroms.

One of the best features of direct wafer bonding is the very low junction leakage current, nearly the same as BULK. A typical value at 5 V is 10 nA/cm^2, compared to 400 nA/cm^2 for SIMOX, as measured on large $n+/p$-diodes [143]. Hence, wafer bonding may be used for advanced bipolar applications [150], where long lifetime is required.

3. Restricted Silicon Layer/Insulator/Silicon Substrate

This category includes SOI methods where the size and placement of silicon islands are restricted by properties that are intrinsic to the technique being used. From the designer's point of view, this is a disadvantage, because it limits the layout freedom, complicates the design, and reduces the packing density. The major reasons for these restrictions are (1) exclusion of defective areas, such as boundaries, or electrically not isolated regions, such as seed-holes; and (2) maximum extension of lateral oxidation, which is used for island pinch-off. This extension would limit the width of a fully isolated island, which is formed by the encroachment of lateral oxidation from opposite edges of the island. Restrictions of the first type are usually associated with polysilicon seeding [151,152], while those of the second type are found in anodic oxidation [153,154] or selective thermal oxidation techniques [155].

a. Seeded Channel MOS Technology. The classification of this technology as SOI is borderline. The reason is that the MOS channel is in electrical contact with the BULK substrate [151], as shown in Fig. 4.44. As a result, this technology can be considered a hybrid between BULK and SOI, as it tries to incorporate the best features of both.

The mobility is similar to BULK, since the channel is formed in high-quality silicon, immediately above the seed window. Floating body effects are eliminated because of the electrical connection to the substrate. The parasitic capacitance is many folds smaller than in BULK, since the source/drain regions are formed over thick oxide. Frameless contacts are possible because a contact hole misalignment would only expose the thick oxide around the channel, without provoking catastrophic consequences. By comparison, in BULK, this could lead to shorting of the source/drain to the substrate.

The seeded channel process starts with silicon wafers of resistivity and type, selected according to CMOS needs. A combination of the gate and active area mask is used to define the channel regions, in order to protect

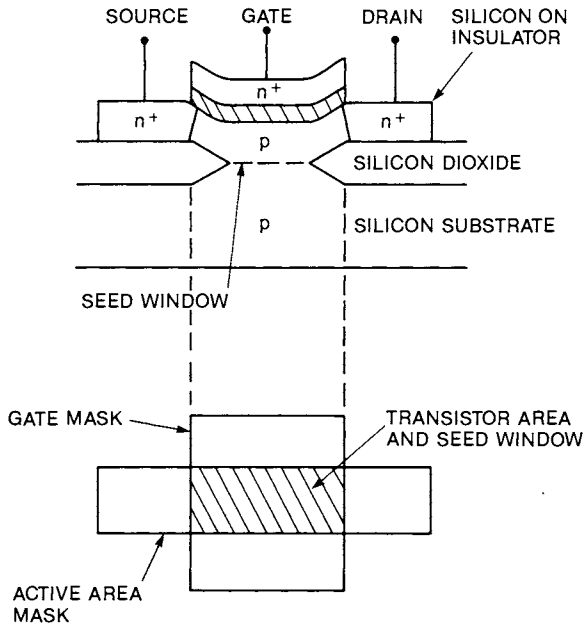


Fig. 4.44. Cross section and top view of the seeded channel SOI transistor. (After Baerg *et al.* [151]. Copyright 1985 by the IEEE.)

them from field oxidation. After exposing the silicon in the channel, a polysilicon layer is deposited, covered by an oxide cap, and recrystallized using a scanned argon laser [151]. Next, the cap is etched off and the normal CMOS process is carried out.

Despite the simplicity of this technology, it is seldom used for reasons which can be speculated as follows. First, if the technology were applied literally, both NMOS and PMOS transistors would make contact to the substrate and thus would require tubs and be subject to latch-up, as in BULK. Second, carriers generated by large radiation transients would not be isolated from the transistors, as in conventional SOI. Third, the recrystallization step needs to be an integral part of the CMOS process, because it occurs in the sequence after the initial masking step. This could be a problem for many CMOS processing facilities, which are not equipped for laser recrystallization of silicon.

b. FIPOS. FIPOS (full isolation by porous oxidized silicon) employs lateral anodic oxidation to form isolated silicon islands over a silicon substrate [153,156]. This approach is appealing for its simplicity and the low cost of equipment if compared, for example, with the expensive ion implanter required by SIMOX. However, in early developments, lateral

oxidation could only be extended to a few micrometers without forming excessively thick porous oxide films, which would cause warpage and later interfere with the rest of the process. For these reasons, the original FIPOS process has not been introduced into manufacturing, although it is probable that after further developments this process might become more manufacturable.

Recently, an improved FIPOS approach was developed. It is known as the ISLANDS method [154] and its process sequence is illustrated in Fig. 4.45. Starting from silicon, a heavily doped $n+$ layer is formed by epitaxy using H_2Cl_2Si and AsH_3. On top of it, a second n-epitaxial layer is deposited with the desired resistivity. Then, Si_3N_4 and SiO_2 are deposited to form the masking stack. Trenches are patterned along the active area edges and reactive ion etched to a depth of a few microns. Using a computer-controlled anodization station, the porous oxide is formed preferentially along the $n+$ epitaxial layer, electrically isolating the top n-silicon layer from the substrate. Finally, the trenches are refilled with oxide and planarized.

The ISLANDS technique has removed many of the manufacturability obstacles of the original FIPOS process, as seen from the following characteristics. The maximum size of an isolated feature is 42 μm in width and unlimited in length. The minimum pitch is 2.8 μm and consists of a 1-μm line and a 1.8-μm gap. The porous oxide thickness is uniformly controlled to 4900 $\pm$ 300 Å. Measured electron mobilities are equivalent to those of BULK. The subthreshold leakage current is low ($\approx$0.1 pA/μm width at

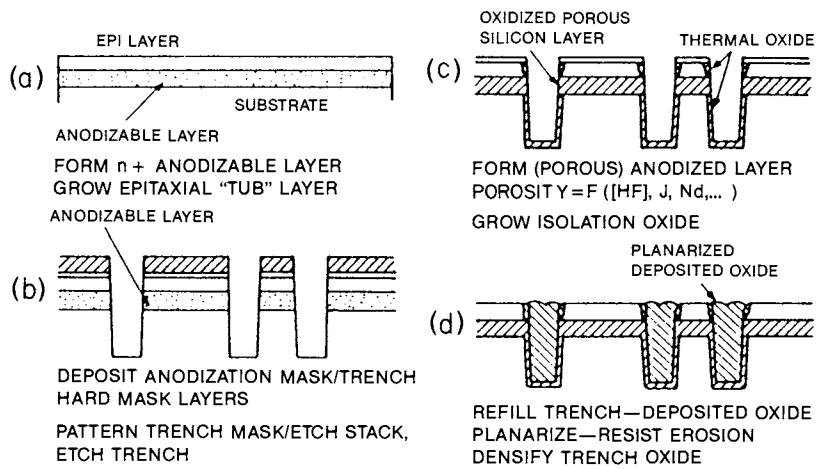


Fig. 4.45. Process overview of the ISLANDS method. (After Zorinsky *et al.* [154]. Copyright 1986 by the IEEE.)

5 V), demonstrating complete elimination of the back channel. On the basis of these results, the ISLANDS version of FIPOS appears a promising contender for the SOI race in submicrometer VLSI.

c. Selective Oxidation Beneath the Top Silicon Layer. This technique is a modification of FIPOS, where anodic oxidation has been replaced by thermal oxidation to use a standard IC process step [155]. The fabrication sequence is illustrated in Fig. 4.46. Using a Si_3N_4/SiO_2 double mask, trenches are formed to define the silicon islands. A second layer of Si_3N_4 is conformally deposited and etched anisotropically by RIE to leave a protection over the island sidewalls while exposing the silicon at the bottom of the trenches. Using selective oxidation in wet ambient, oxide is grown laterally under the edges of the silicon islands until they are electrically isolated from the substrate by oxide encroachment. To obtain a planar surface, the trenches are then refilled with polysilicon and CVD SiO_2.

Although devices made with this process have good characteristics, the

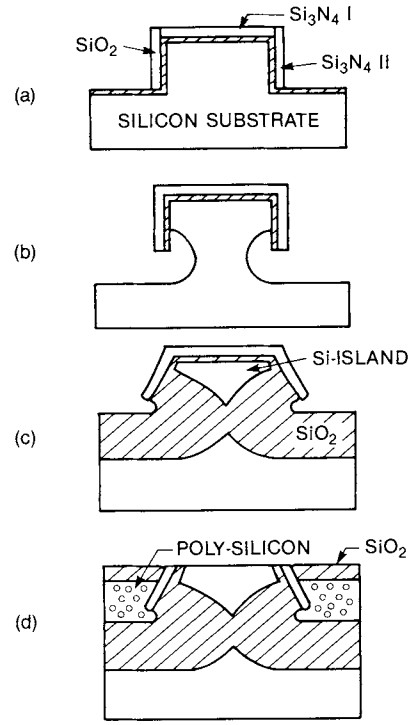


Fig. 4.46. Process sequence of CMOS/SOI process with selective oxidation. (After Kubota *et al.* [155]. Copyright 1986 by the IEEE.)

width of the islands is very small, 1.1 – 1.7 μm [155], compared to > 40 μm with FIPOS using the ISLANDS method [154]. This width is too small for adequate layout flexibility, and because of the intrinsic properties of thermal oxidation, it is unlikely that large lateral-to-vertical oxidation ratios can be achieved. Consequently, this technique appears too restrictive for transition to manufacturing.

REFERENCES

1. J. A. Appels, E. Kooi, M. M. Paffen, J. J. H. Schatorje, and W. H. C. G. Verkuylen, Local oxidation of silicon and its applications in semiconductor device technology, *Philips Research Reports* **25**,118 – 132 (1970).
2. E. Bassous, H. N. Yu, and V. Maniscalco, Topology of silicon structures with recessed SiO_2, *J. Electrochem. Soc.* **123**(11), 1729 – 1737 (1976).
3. W. C. Holton and R. K. Cavin, III, A perspective on CMOS technology trends, *Proc. IEEE* **74**(12), 1646 – 1668 (1986).
4. S. L. Partridge, The current status of silicon-on-insulator technologies—a comparison, *IEDM Tech. Dig.*, pp. 428 – 430, Los Angeles, California, December, 1986.
5. H. T. Weaver, Overview—SOI technology, *IEEE Circuits and Devices Magazine* **3**(4), 3 – 5 (1987).
6. R. D. Rung, H. Momose, and Y. Nagakubo, Deep trench isolated CMOS devices, *IEDM Tech. Digest, Abs. 9.6,* pp. 237 – 240, San Francisco, California, December, 1982.
7. K. M. Cham and S. Y. Chiang, A study of trench surface inversion problems in the trench CMOS technology, *IEEE Electron Device Lett.* **EDL-4**(9), 303 – 305 (1983).
8. K. Kasama, F. Toyokawa, M. Sakamoto, and K. Kobayashi, A radiation-hardened insulator for MOS LSI device isolation, *IEEE Trans. Nucl. Sci.* **NS-32**(6), 3965 – 70 (1985).
9. J. Y. Chen and D. E. Snyder, Modeling device isolation in high-density CMOS, *IEEE Electron Device Lett.* **EDL-7**(2), 64 – 65 (1986).
10. D. B. Estreich, The physics and modeling of latch-up and CMOS integrated circuits, Stanford University, Tech. Rep. G-201-9, November, 1980.
11. Y. Akasaka and T. Nishimura, Concept and basic technologies for 3-D IC structure, *IEDM Tech. Digest Abs. 18.6,* pp. 488 – 491, Los Angeles, California, December, 1986.
12. Y. Inoue, K. Sugahara, S. Kusunoki, M. Nakaya, T. Nishimura, Y. Horiba, Y. Akasaka, and H. Nakata, A three-dimensional static RAM, *IEEE Electron Device Lett.* **EDL-7**(5), 327 – 329 (1986).
13. D. M. Brown, M. Ghezzo, and J. M. Pimbley, Trends in advanced process technology —submicrometer CMOS device design and process requirements, *Proc. IEEE* **74**(12), 1678 – 1702,(1986).
14. S. P. Tay, J. P. Ellul, and M. I. H. King, Applications of high-pressure technology to ULSI fabrication, Materials Issues in Silicon Integrated Circuit Processing Symposium, Mater. Res. Soc. 1986, Palo Alto, California, 15 – 18 April, 1986.
15. K. Kurosawa, T. Shibata, and H. Iizuka, A new bird's beak free field isolation technology for VLSI devices, *IEDM Tech. Digest, Abs. 16.4,* pp. 384 – 387 Washington, D. C., December, 1981.

16. R. D. Rung, C. J. Dell'Oca, and L. G. Walker, A retrograde p-well for higher density CMOS, *IEEE Trans. Electron Devices* **ED-28**(10), 1115–1119 (1981).
17. K. Shibata and K. Taniguchi, Generation mechanism of dislocations in local oxidation of silicon, *J. Electrochem. Soc.* **127**(6), 1383–1387 (1980).
18. K. Y. Chiu, J. L. Moll, and J. Manoliu, A bird's beak free local oxidation technology feasible for VLSI circuits fabrication, *IEEE Trans. Electron Devices* **ED-29**(4), 536–540 (1982).
19. K. Y. Chiu, R. Fang, J. Lin, J. L. Moll, C. Lage, S. Angelos, and R. Tillman, The SWAMI—a defect-free and near-zero bird's-beak local oxidation process and its application in VLSI technology, *IEDM Tech. Digest, Abs. 9.3,* pp. 224–227, San Fransisco, California, December, 1982.
20. J. C. Hui, T. Y. Chiu, S. S. Wong, and W. G. Oldham, Sealed-interface local oxidation technology, *IEEE Trans. Electron Devices* **ED-29**(4), 554–561 (1982).
21. J. Hui, P. Vande Voorde, and J. Moll, Scaling limitations of submicron local oxidation technology, *IEDM Tech. Digest, Abs. 14.7,* pp. 392–395, Washington, D.C., December, 1985.
22. E. Kooi, J. G. Van Lierop, and J. A. Appels, Formation of silicon nitride at a Si-SiO_2 interface during local oxidation of silicon and during heat-treatment of oxidized silicon in NH_3 gas, *J. Electrochem. Soc.* **123**(7), 1117–1120 (1976).
23. N. Endo, K. Tanno, A. Ishitani, Y. Kurogi, and H. Tsuya, Novel device isolation technology with selective epitaxial growth, *IEDM Tech. Dig.* pp. 241–244, December, 1982; also in *IEEE Trans. Electron Devices* **ED-31,** 1283–1288 (1984).
24. T. Yamaguchi, S. Morimoto, G. H. Kawamoto, H. K. Park, and G. C. Eiden, High-speed latchup-free 0.5-μm-channel CMOS using self-aligned TiSi_2 and deep-trench isolation technologies, *IEDM Tech. Dig.* p. 522, December 1983; also in *IEEE Trans. Electron Devices* **ED-32**(2), 184–193 (1985).
25. P. Balk, P. J. Burkhardt, and L. V. Gregor, Orientation dependence of built-in surface charge on thermally oxidized silicon, *Proc. IEEE (Correspondence),* **53,** 2133 (1965).
26. T. Iizuka, K. Chiu, and J. Moll, Double threshold MOSFET's in bird's beak free structure, *IEDM Tech. Dig., Abs. 16.3,* pp. 380–383, Washington, D.C., December, 1981.
27. M. C. Peckerar and R. E. Neidert, High-speed microelectronics for military applications, *Proc. IEEE* **71,** 657–666 (1983).
28. G. F. Derbenwick and B. L. Gregory, Process optimization of radiation-hardened CMOS integrated circuits, *IEEE Trans. Nucl. Sci.* **NS-22,** 2151–2156 (1975).
29. A. Bohg and A. K. Gaind, Influence of film stress and thermal oxidation on the generation of dislocations in silicon, *Appl. Phys. Lett.* **33**(10), 895–897 (1978).
30. W. R. Hunter, L. Ephrath, W. D. Grobman, C. M. Osburn, B. L. Crowder, A. Cramer, and H. E. Luhn, 1 μm MOSFET VLSI technology: part V-A single-level polysilicon technology using electron-beam lithography, *IEEE Trans. Electron Devices* **ED-26**(4), 353–359 (1979).
31. C. W. Teng, G. Pollack, and W. R. Hunter, Optimization of sidewall masked isolation process, *IEEE Trans. Electron Devices* **ED-32**(2), 124–131 (1985).
32. S. Sawada, T. Higuchi, T. Mizuno, S. Shinozaki, and O. Ozawa, Electrical properties for MOS LSI's fabricated using stacked oxide SWAMI technology, *IEEE Trans. Electron Devices* **ED-32**(11), 2243–2248 (1985).
33. M. Ghezzo, M. J. Kim, J. F. Norton, and R. J. Saia, Laterally sealed LOCOS isolation, *J. Electrochem. Soc.* **134**(6), 1475–1479 (1987).
34. H.-H. Tsai, C.-L. Yu, and C.-Y. Wu, A bird's beak reduction technique for LOCOS in VLSI fabrication, *IEEE Electron Device Lett.* **EDL-7**(2), 122–123 (1986).

35. H.-H. Tsai, S.-M. Chen, and C.-Y. Wu, A new fully recessed-oxide (FUROX) field isolation technology for scaled VLSI circuit fabrication, *IEEE Electron Device Lett.* **EDL-7**(2), 124–126 (1986).
36. J. Hui, T. Y. Chiu, S. Wong, and W. G. Oldham, Selective oxidation technologies for high density MOS, *IEEE Electron Device Lett.* **EDL-2,** 244–247 (1981).
37. P. Deroux-Dauhphin and J. P. Gonchond, Physical and electrical characterization of a SILO isolation structure, *IEEE Trans. Electron Devices* **ED-32,** 2392–2398 (1985).
38. M. J. Kim and M. Ghezzo, *J. Electrochem. Soc.* **131,** 1934 (1984).
39. J. Y. Lee, C. W. Slayman, H. L. Garvin, R. E. Kastris, and M. C. Montes, A new self-aligned VLSI isolation process using thin-metal lift-off, *IEEE Electron Device Lett.* **EDL-8**(7), 309–311 (1987).
40. J. Matsunaga, N. Matsukawa, H. Nozawa, and S. Kohyama, Selective polysilicon oxidation technology for defect free isolation, *IEDM Tech. Dig., Abs. 22.4,* pp. 565–568, Washington, D.C., December, 1980.
41. N. Matsukawa, N. Nozawa, J. Matsunaga, and S. Kohyama, Selective polysilicon oxidation technology for VLSI isolation, *IEEE Trans. Electron Devices* **ED-29**(4), 561–567 (1982).
42. K. L. Wang, S. A. Saller, W. R. Hunter, P. K. Chatterjee, and P. Yang, Direct moat isolation for VLSI, *IEEE Trans. Electron Devices* **ED-29**(4), 541–547 (1982).
43. J. Y. Lee, C. W. Slayman, H. L. Garvin, R. E. Kastris, and M. C. Montes, A new self-aligned VLSI isolation process using thin-metal lift-off, *IEEE Electron Device Lett.* **EDL-8**(7), 309–311 (1987).
44. T. Shibata, R. Nakayama, K. Kurosawa, S. Onga, M. Konaka, and H. Iizuka, A simplified BOX (buried-oxide) isolation technology for megabit dynamic memories, *IEDM Tech. Dig., Abs. 2.3.* pp. 27–30, Washington, D.C., December, 1983.
45. G. Fuse, M. Fukumoto, A. Shinohara, S. Odanaka, M. Sasago, and T. Ohzone, A new isolation method with boron-implanted sidewalls for controlling narrow-width effect, *IEEE Trans. Electron Devices* **ED-34**(2), 356–360 (1987).
46. J. Y. Chen, R. C. Henderson, J. T. Hall, and E. W. Yee, A fully recessed field isolation technology using photo-CVD oxide, *IEDM Tech. Dig., Abs. 9.5,* pp. 233–236, San Francisco, California, December, 1982.
47. Y. Tamaki, T. Kure, T. Shibata, and H. Higuchi, U-groove isolation for high density bipolar LSI's, *Jpn. J. Appl. Phys.* **21**(Suppl. 21-1), 37 (1981).
48. H. P. Vyas, R. S. Lutze, and J. S. T. Huang, A trench-isolated submicrometer CMOS technology, *IEEE Trans. Electron Devices* **ED-32**(5), 926–931 (1985).
49. C. Gonzalez and J. P. McVittie, A study of trenched capacitor structures, *IEEE Electron Device Lett.* **EDL-6**(5), 215–218 (1985).
50. P.-L. Chen, A. Selcuk, and D. Erb, A double-epitaxial process for high-density DRAM trench-capacitor isolation, *IEEE Electron Device Lett.* **EDL-8**(11), 550–552 (1987).
51. D. A. Baglee, R. R. Doering, M. Elahy, M. Yashiro, D. Clark, S. Crank, and G. Armstrong, Properties of trench capacitors for high density DRAM applications, *IEDM Tech. Dig., Abs. 14.5,* pp. 384–387, Washington, D.C., December, 1985.
52. R. B. Marcus and T. T. Sheng, *J. Electrochem. Soc.* **129,** 1278 (1982).
53. K. Yamabe and K. Imai, Nonplanar oxidation and reduction of oxide leakage currents at silicon corners by rounding-off oxidation, *IEEE Trans. Electron Devices* **ED-34**(8), 1681–1687 (1987).
54. A. L. Esquivel, A. T. Mitchell, J. L. Paterson, M. Douglas, H. L. Tigelaar, B. R. Riemenshneider, T. M. Coffman, M. Gill, R. Lahiry, D. McElroy, and P. Shah, A novel trench-isolated buried $n+$ FAMOS transistor suitable for high-density EPROM's, *IEEE Electron Device Lett.* **EDL-8**(4), 146–147 (1987).

55. B. D. Joyce and J. A. Baldrey, Selective epitaxial deposition of silicon, *Nature (London)* **195,** 485–486 (1962).
56. S. Iwamatsu, S. Meguro, and S. Shimizu, A new isolation structure for high density LSI, *IEDM Tech. Dig.* pp. 244–247, Washington, D.C., December, 1973.
57. O. Shinchi and J. Sakurai, The buried-oxide MOSFET—a new type of high-speed switching, *IEEE Trans. Electron Devices* **ED-23,** 1190–1191 (1976).
58. K. Tanno, N. Endo, H. Kitajima, Y. Kurogi, and H. Tsuya, *Jpn. J. Appl. Phys. Lett.* **21,** L564–566 (1982).
59. H. J. Voss and H. Kurten, Device isolation technology by selective low-pressure silicon epitaxy, *IEDM Tech. Dig., Abs. 2.5,* pp. 35–38, Washington, D.C., December, 1983.
60. N. Kasai, N. Endo, A. Ishitani, and H. Kitajima, ¼-μm CMOS isolation technique using selective epitaxy, *IEEE Trans. Electron Devices* **ED-34**(6), 1331–1336 (1987).
61. N. Endo, N. Kasai, A. Ishitani, H. Kitajima, and Y. Kurogi, Scaled CMOS technology using SEG isolation and buried well process, *IEEE Trans. Electron Devices* **ED-33**(11), 1659–1666 (1986).
62. J. O. Borland and C. I. Drowley, Advanced dielectric isolation through selective epitaxial growth techniques, *Solid State Technology,* **28**(8), 141–148 (1985).
63. J. O. Borland, Novel device structures by selective epitaxial growth, *IEDM Tech. Dig., Abs. 2.1,* pp. 12–15, Washington, D.C., December, 1987.
64. J. Manoliu and J. O. Borland, A submicron buried layer twin well CMOS SEG process, *IEDM Tech. Dig., Abs. 2.3,* pp. 20–23, Washington, D.C., December, 1987.
65. S. Nagao, K. Higashitani, Y. Akasaka, and H. Nakata, Application of selective epitaxial growth for CMOS technology, *IEEE Trans. Electron Devices* **ED-33**(11), 1738–1744 (1986).
66. T. I. Kamins and S.-Y. Chiang, CMOS device isolation using the selective-etch-and-refill-with-EPI (SEREPI) process, *IEEE Trans. Electron Device Lett.* **EDL-6**(12), 617–619 (1985).
67. O. Minato, T. Masuhara, T. Sasaki, Y. Sakai, M. Kubo, K. Uchibori, and T. Yasui, A high-speed low-power Hi-CMOS 4K static RAM, *IEEE Trans. Electron Devices,* **ED-26**(6), 882–885, (1979).
68. T. Ohzone, H. Shimura, K. Tsuji, and T. Hirao, Silicon-gate n-well CMOS process by full ion-implantation technology, *IEEE Trans. Electron Devices* **ED-27**(9), 1789–1795 (1980).
69. L. C. Parillo, R. S. Payne, R. E. Davis, G. W. Reutlinger, and R. L. Field, Twin-tub CMOS—a technology for VLSI circuits, *IEDM Tech. Dig., Abs. 29.1,* pp. 752–755, Washington, D.C., December, 1980.
70. L. C. Parillo, L. K. Wang, R. D. Swenumson, R. L. Field, R. C. Melin, and R. A. Levy, Twin-tub CMOS II—an advanced VLSI technology, *IEDM Tech. Dig., Abs. 29.3,* pp. 706–709, San Francisco, California, December, 1982.
71. W. C. Black, Jr., R. H. McCharles, and D. A. Hodges, CMOS process for high-performance analog LSI, *IEDM Tech. Dig., Abs. 14.4,* pp. 331–334, Washington, D.C., December, 1976.
72. D. Takacs, J. Harter, E. P. Jacobs, C. Werner, U. Schwabe, J. Winner, and E. Lange, Comparison of latch-up in p- and n-well CMOS circuits, *IEDM Tech. Dig.,* p. 159, Washington, D.C., December, 1983.
73. J. Kiely, The impact of epitaxial silicon on CMOS VLSI/ULSI device processing, Applied Materials Seminar on Innovations in VLSI/ULSI CMOS Technology, San Jose, California, December, 5, 1986.
74. P. Chatterjee and 4Mb dRAM Team, Trench and compact structures for dRAMs, *IEDM Tech. Dig., Abs. 6.1,* pp. 128–131, Los Angeles, California, December, 1986.

75. S. R. Combs, Scaleable retrograde p-well CMOS technology, *IEDM Tech. Dig., Abs. 15.1,* pp. 346–349, Washington, D.C., December, 1981.
76. R. A. Martin and J. Y. Chen, Optimized retrograde n-well for one micron CMOS technology, *Proc. IEEE Custom Integrated Circuit Conf.,* p. 199 Portland, Oregon, May 20–23, 1985.
77. J. Y. Chen, Quadruple-well CMOS for VLSI technology, *IEEE Trans. Electron Devices* **ED-31**(7), 910–919 (1984).
78. R. R. Troutman, "Latch-up in CMOS Technology." Kluwer Academic Publishers, Norwell, Massachusetts, 1986.
79. K. Yu, R. J. C. Chwang, M. T. Bohr, P. A. Warkentin, S. Stern, and C. N. Berglund, HMOS-CMOS—a low-power high-performance technology, *IEEE J. Solid-State Circuits* **SC-16**(5), 454–459 (1981).
80. A. W. Wieder, C. Werner, and J. Harter, Design model for bulk CMOS scaling enabling accurate latchup prediction, *IEEE Trans. Electron Devices* **ED-30**(3), 240–245 (1983).
81. J. E. Hall, J. A. Seitchik, L. A. Arledge, and P. Yang, An improved circuit model for CMOS latchup, *IEEE Electron Device Lett* **EDL-6**(7), 320–322 (1985).
82. K. Y. Fu, "Transient latchup in bulk CMOS with a voltage-dependent well-substrate junction, *IEEE Trans. Electron Devices* **ED-32**(3), 717 (1985).
83. R. R. Troutman and H. P. Zappe, A transient analysis of latch-up in bulk CMOS, *IEEE Trans. Electron Devices* **ED-30**(2), 170–179 (1983).
84. A. G. Lewis, R. A. Martin, T. Y. Huang, and J. Y. Chen, Three-dimensional effects in CMOS latch-up, *IEDM Tech. Dig., Abs. 10.4,* pp. 248–251, Los Angeles, California, December, 1986.
85. E. Sangiorgi, B. Ricco, and L. Selmi, Three-dimensional distribution of CMOS latch-up current, *IEEE Electron Device Lett* **EDL-8**(4), 154–156 (1987).
86. D. B. Estreich and R. W. Dutton, Modelling latch-up in CMOS integrated circuits, *IEEE Trans. on Computer-Aided Design of Integrated Circuits and Systems* **CAD-1,** 157 (1982).
87. R. D. Rung and H. Momose, DC holding and dynamic triggering characteristics of bulk CMOS latch-up, *IEEE Trans. Electron Devices* **ED-30,** 1647–1655 (1983).
88. S. Odanaka, M. Wakabayashi, and T. Ohzone, The dynamics of latchup turn-on behavior in scaled CMOS, *IEEE Trans. Electron Devices* **ED-32**(7), 1334–1340 (1985).
89. R. R. Troutman, Recent developments and future trends in latch-up prevention in scaled CMOS, *IEEE Trans. Electron Devices* **ED-30,** 1564 (1983).
90. N. Shiono, Y. Sakagawa, T. Matsumoto, and Y. Akasaka, A 64K SRAM with high immunity from heavy ion induced latch-up, *IEEE Electron Device Lett.* **EDL-7**(1), 20–22 (1986).
91. H. K. Gummel, Measurement of the number of impurities in the base layer of a transistor, *Proc. IRE* **49,** 834 (1961).
92. W. R. Dawes, Jr., and G. F. Derbenwick, Prevention of CMOS latch-up by gold doping, *IEEE Trans. Nucl. Sci.* **NS-23,** 2027 (1976).
93. J. R. Adams and R. J. Sokel, Neutron irradiation for prevention of latch-up in MOS integrated circuits, *IEEE Trans. Nucl. Sci.* **NS-26,** 5069 (1979).
94. C. N. Anagnostopoulos, E. T. Nelson, J. P. Lavine, K. Y. Wong, and D. N. Nichols, Latch-up and image crosstalk suppression by internal gettering, *IEEE Trans. Electron Devices* **ED-31**(2), 225–231 (1984).
95. M. Sugino, L. A. Akers, and M. E. Rebeschini, Latch-up free Schottky-barrier CMOS, *IEEE Trans. Electron Devices* **ED-30**(2), 110–118 (1983).

96. S. Swirhun, E. Sangiorgi, A. Weeks, R. M. Swanson, K. C. Saraswat, and R. W. Dutton, Latch-up free CMOS using guarded Shottky barrier technology, *IEDM Tech. Dig.*, p. 402, San Francisco, California, December, 1984.
97. A. G. Lewis, Latchup suppression in fine-dimension shallow p-well CMOS circuits, *IEEE Trans. Electron Devices* **ED-31**(10), 1472-1481 (1984).
98. D. Takacs, C. Werner, J. Harter, and U. Schwabe, Surface induced latchup in VLSI CMOS, *IEEE Trans. Electron Devices* **ED-31**(3), 279 (1984).
99. J. O. Borland, M. Kuo, J. Shibley, B. Roberts, R. Schindler, and T. Dalrymple, An intrinsic gettering process to improve minority carrier lifetimes in MOS and bipolar silicon epitaxial technology, *Semiconductor Processing, ASTM STP 850,* Dinesh C. Gupta, ed. pp. 49–62. American Society for Testing and Materials, 1984.
100. R. Jerdonek, M. Ghezzo, J. Weaver, and S. Combs, Reduced geometry CMOS technology, *IEDM Tech. Dig., Abs. 17.1,* pp. 450–453, San Francisco, California, December, 1982.
101. D. B. Estreich and A. Ochoa, Jr., An analysis of latch-up in CMOS IC's using an epitaxial-buried layer process, *IEDM Tech. Dig.*, p. 230, Washington, D.C., December, 1978.
102. R. R. Troutman and H. P. Zappe, Layout and bias considerations for preventing transiently triggered latchup in CMOS, *IEEE Trans. Electron Devices* **ED-31**(3), 315–321 (1984).
103. L. Herman, Controlling CMOS latchup, *VLSI Des.*, p. 100, April 1985.
104. G. J. Hu and R. H. Bruce, A CMOS structure with high latchup holding voltage, *IEEE Electron Device Lett.* **EDL-5**(6), 211–214 (1984).
105. T. Yamaguchi, S. Morimoto, G. H. Kawamoto, and J. C. DeLacy, Process and device performance of 1 μm-channel n-well CMOS technology, *IEEE Trans. Electron Devices* **ED-31**(2), 205–214 (1984).
106. A. G. Lewis, J. Y. Chen, R. A. Martin, and T.-Y. Huang, Device isolation in high-density LOCOS-isolated CMOS, *IEEE Trans. Electron Devices* **ED-34**(6), 1337–1345 (1987).
107. S. J. Hillenius, R. Liu, G. E. Georgiou, R. L. Field, D. S. Williams, A. Kornblit, D. M. Boulin, R. L. Johnston, and W. T. Lynch, A symmetric submicron CMOS technology, *IEDM Tech. Dig., Abs. 10.5,* pp. 252–255, Los Angeles, California, December, 1986.
108. A. Stolmeijer, A twin-well CMOS process employing high-energy ion implantation, *IEEE Trans. Electron Devices* **ED-33**(4), 450–457 (1986).
109. K. Soliman and D. K. Nicols, Latch-up in CMOS devices from heavy ions, *IEEE Trans. Nucl. Sci.*, **NS-30,** 4514 (1983).
110. E. Sangiorgi, R. L. Johnston, M. R. Pinto, P. F. Bechtold, and W. Fichtner, Temperature dependence of latch-up phenomena in scaled CMOS structures, *IEEE Electron Device Lett.* **EDL-7**(1), 28–31 (1986).
111. E. J. Boleki, *RCA Review* **31,** 372 (1970).
112. A. C. Ipri, The properties of silicon-on-sapphire: substrates, devices, and integrated circuits, *Appl. Solid State Sci.* (Suppl. 2A), D. Kahng, ed., pp. 253–395. Academic Press, San Diego, California, 1981.
113. B.-Y. Tsaur, Zone-melting-recrystallization silicon-on-insulator technology, *IEEE Circuits and Devices Magazine* **3**(4), 12–16 (1987).
114. A. C. Ipri, Electrical properties of silicon films on sapphire using the MOS Hall technique, *J. Appl. Phys.* **43,** 2770 (1972).
115. D. C. Mayer, P. K. Vasudev, J. Y. Lee, Y. K. Allen, and R. C. Henderson, A

short-channel CMOS/SOS technology in recrystallised 0.3-μm-thick silicon-on-sapphire films, *IEEE Electron Device Lett.* **EDL-5,** 156 (1984).
116. M. P. Brassington, A. G. Lewis, and S. L. Partridge, A comparison of fine-dimension silicon-on-sapphire and bulk-silicon complementary MOS devices and circuits, *IEEE Trans. Electron Devices* **ED-32**(9), 1858-67 (1985).
117. M. R. Splinter, A 2-μm silicon-gate C-MOS/SOS technology, *IEEE Trans. Electron Devices* **ED-25**(8), 996–1004 (1978).
118. K. Ohtake, K. Shirakawa, M. Koba, K. Awane, Y. Ohta, D. Azuma, and S. Miyata, Triple layer SOI dynamic memory, *IEDM Tech. Dig., Abs. 6.6,* pp. 148–151, Los Angeles, California (1986).
119. H. W. Lam, SIMOX SOI for integrated circuit fabrication, *IEEE Circuits and Devices Magazine* **3**(4), 6–11 (1987).
120. Y. Kobayashi, A. Fukami, and T. Nagano, Characteristics of a 1.2-μm CMOS technology fabricated on an RF-heated zone-melting recrystallized SOI, *IEEE Electron Device Lett.* **EDL-7**(6), 350–352 (1986).
121. Y. Akasaka, Three-dimensional IC trends, *Proc. IEEE* **74**(12), 1703–1714 (1986).
122. S. Taguchi, H. Tango, K. Maeguchi, and L. M. Dang, Performance of downward scaled CMOS/SOS, *IEDM Tech. Dig., Abs. 25.4,* pp. 589–593, Washington, D.C., December, 1979.
123. P. K. Vasudev, Recent advances in solid-phase epitaxial recrystallization of SOS with applications to CMOS and bipolar processes, *IEEE Circuits and Devices Magazine* **3**(4), 17–19 (1987).
124. T. Yoshii, S. Taguchi, T. Inoue, and H. Tango, Improvement of SOS device performance by solid-phase epitaxy, *Jpn. J. Appl. Phys.* **21**(Suppl. 21-1), 175–179 (1982).
125. T. Yoshii, S. Taguchi, T. Inoue, and H. Tango, CMOS/SOS devices improved by silicon implantation and subsequent thermal annealing technique, pp. 195–202, in JARECT Vol. 8, Semiconductor Technologies, J. Nishizawa, ed. OHM/North-Holland, Tokyo/Amsterdam, 1983.
126. P. K. Vasudev and D. C. Mayer, Solid-phase recrystallization of SOS, *Proc. 1984 Materials Research Sympos. on Thin Film Transistors and Silicon on Insulators* **33,** 35 (1984).
127. T. Inoue and T. Yoshii, Double solid-phase epitaxy of SOS, *Appl. Phys. Lett.* **36,** 64 (1980).
128. D. J. McGreivy, On the origin of leakage currents in silicon-on-sapphire MOS transistors, *IEEE Trans. Electron Devices* **ED-24**(6), 730–738 (1977).
129. C.-E. Chen, T. G. W. Blake, L. R. Hite, S. D. S. Malhi, B.-Y. Mao, and H. W. Lam, SOI-CMOS 4K SRAM with high dose oxygen implanted substrate, *IEDM Tech. Dig.,* pp. 702–705, San Francisco, California, December, 1984.
130. H. Onoda, M. Sasaki, T. Katoh, and N. Hirashita, Si-gate CMOS devices on a Si/CaF_2/Si structure, *IEEE Trans. Electron Devices* **ED-34**(11), 2280–2290 (1987).
131. J. B. Lasky, S. R. Stiffler, F. R. White, and J. R. Abernathey, Silicon-on-insulator (SOI) by bonding and etch-back, *IEDM Tech. Dig., Abs. 28.4,* pp. 684–686, Washington, D.C., December, 1985.
132. M. Shimbo, K. Furukawa, K. Fukuda, and K. Tanzawa, Silicon-to-silicon direct bonding method, *J. Appl. Phys.* **60**(8), 2987–2989 (1986).
133. H. Ohashi, J. Ohura, T. Tsukakoshi, and M. Simbo, Improved dielectrically isolated devices integration by silicon-wafer direct bonding (SDB) technique, *IEDM Tech. Dig., Abs. 9.1,* pp. 210–213, Los Angeles, California, December, 1986.
134. K. Izumi, M. Doken, and H. Ariyoshi, *Electron. Lett.* **14,** 593 (1978).

135. K. Izumi, Y. Omura, and T. Sakai, SIMOX technology and its application to CMOS LSIs, *J. Electron. Mater.* **12,** 845 (1983).
136. K. Izumi, Y. Omura, and S. Nakashima, Promotion of practical SIMOX technology by the development of a 100 mA class high-current oxygen implanter, *Electron Lett.* **22**(15), 775 (1986).
137. H. W. Lam, R. F. Pinizzotto, H. T. Yuan, and D. W. Bellavance, *Electron. Lett.* **17,** 356 (1981).
138. B. Y. Mao, P.-H. Chang, H. W. Lam, B. W. Shen, and J. A. Keenan, *Appl. Phys. Lett.* **48,** 794 (1986).
139. H. W. Lam, Silicon-on-insulator epitaxy, "Epitaxy Silicon Technology," J. Baliga, ed., p. 269. Academic Press, San Diego, California, 1986.
140. J. R. Davis, K. J. Reeson, P. L. F. Hemment, and C. D. Marsh, High-performance SOI-CMOS transistors in oxygen-implanted silicon without epitaxy, *IEEE Electron Device Lett.* **EDL-8**(7), 291–293 (1987).
141. J.-P. Colinge, Reduction of floating substrate effect in thin-film SOI MOSFETs, *Electron. Lett.* **22**(4), 187 (1986).
142. J.-P. Colinge, K. Hashimoto, T. Kamins, S.-Y. Chiang, E.-D. Liu, S. Peng, and P. Rissman, *IEEE Electron. Dev. Lett.* **EDL-7**(5), 279–281 (1986).
143. W. A. Krull, J. F. Buller, G. V. Rouse, and R. D. Cherne, Electrical and radiation characterization of three SOI material technologies, *IEEE Circuits and Devices Magazine* **3**(4), 20–26 (1987).
144. G. E. Possin, H. G. Parks, S. W. Chiang, and Y. S. Liu, MOSFETs fabricated in laser-recrystallized silicon on quartz using selectively absorbing dielectric layers, *IEEE Trans. Electron Devices* **ED-31,** 68–74 (1984).
145. J. C. C. Fan, B.-Y. Tsaur, and M. W. Geis, Graphite-strip heater zone-melting recrystallization of Si films, *J. Cryst. Growth* **63**(3), 453 (1983).
146. T. Stultz, J. Sturm, and J. F. Gibbons, Beam processing of silicon with a scanning CW Hg lamp, *in* "Laser-Solid Interactions and Transient Thermal Processing of Materials," J. Narayan, W. L. Brown, and R. A. Lemons, eds., p. 463. North-Holland Publ., Amsterdam, 1983.
147. B.-Y. Tsaur and C. K. Chen, Submicrometer CMOS devices in zone-melting-recrystallized SOI films, *IEEE Electron Device Lett.* **EDL-7**(7), 443–445 (1986).
148. C. K. Chen, M. W. Geis, M. C. Finn, and B.-Y. Tsaur, A new capping technique for zone-melting recrystallization of silicon-on-insulator films, *Appl. Phys. Lett.* **48,** 1300 (1986).
149. K. Sugahara, T. Nishimura, S. Kusunoki, Y. Akasaka, and H. Nakata, SOI/SOI/Bulk-Si triple-level structure for three-dimensional devices, *IEEE Electron Device Lett.* **EDL-7**(3), 193–195 (1986).
150. A. Nakagawa, K. Watanabe, Y. Yamaguchi, H. Ohashi, and K. Furukawa, 1800 V bipolar-mode MOSFETs: a first application of silicon wafer direct bonding (SDB) technique to a power device, *IEDM Tech. Dig., Abs. 5.6,* pp. 122–125, Los Angeles, California, 1986.
151. W. Baerg, J. C. Sturm, T. L. Hwa, H. Y Lin, C. H. Ting, J. C. Tzeng, and J. F. Gibbons, A seeded-channel silicon-on-insulator (SOI) MOS technology, *IEEE Electron Device Lett.* **EDL-6**(12), 668–670 (1985).
152. H. W. Lam, R. F. Pinizzoto, and A. F. Tasch, Jr., Single crystal silicon on oxide by a scanning cw laser induced lateral seeding process, *J. Electrochem. Soc.* **128,** 1981–1986 (1981).
153. K. Imai, A new dielectric isolation method using porous silicon, *Solid-State Electron.* **24,** 159–164 (1981).

154. E. J. Zorinsky, D. B. Spratt, and R. L. Virkus, The ISLANDS method—a manufacturable porous silicon SOI technology, *IEDM Tech. Dig., Abs. 16.7,* pp. 431–434, December, 1986.
155. M. Kubota, T. Tamaki, K. Kawakita, N. Nomura, and T. Takemoto, New SOI CMOS process with selective oxidation, *IEDM Tech. Dig., Abs. 16.9,* pp. 814–816, Los Angeles, California, December, 1986.
156. K. Imai and H. Unno, FIPOS (full isolation by porous oxidized silicon) technology and its application to LSI ICs, *IEEE Trans. Electron Devices* **ED-31**(3), 297–302 (1984).

Chapter 5

Reliability

A primary advantage of electrical devices over their mechanical counterparts is long-term reliability. Upon inspection of a new automated wafer inspection system, one of our colleagues exclaimed succinctly, "I see moving parts. That means maintenance." Confidence in the continued, uninterrupted operation of any device or system is of great importance. Thus, adequate reliability assumes equal footing with considerations such as cost and performance in the definition of a new semiconductor technology.

Unfortunately, continued miniaturization in CMOS evolution directly exacerbates several reliability hazards. So one immediately encounters a decision point in process development. One must accept reduced reliability or work harder to conquer reliability problems. The latter option may require a more difficult or exacting fabrication process or, at a minimum, adds constraints to the process flow. The reliability issues we discuss in the context of CMOS process development are channel hot electron degradation, electromigration, and dielectric (silicon dioxide) wear-out.

I. CHANNEL HOT ELECTRON DEGRADATION

Channel hot electron (CHE) degradation [1,2] appears in the short-channel NMOS field-effect transistor operation. [We note that we employ the term "channel" hot electron degradation to differentiate this subject from "substrate" hot electron degradation [1,3]. As we shall discuss, we do not accept the "drain avalanche hot carrier" (DAHC) hypothesis [4–6] and thus do not list DAHC as a reliability issue separate from CHE.] To appreciate the physical mechanism of the CHE problem, we digress briefly

to the general subject of electron transport in a semiconductor. For the MOSFET operating above the threshold condition, as discussed in Chapter 2, electrons are injected into the channel from the source since the source–channel diode is forward biased by the effect of the gate bias. After injection, the electrons drift under the influence of the drain bias. The drain bias creates an electric field in the MOSFET channel with a polarity that pulls negatively charged electrons toward the drain.

The response of an electron within a semiconductor to an applied electric field is complex. An electron in free space would simply accelerate (indefinitely) in proportion to the Lorentz force of the electric field. [7]. Within a semiconductor, as well as many other materials, many other factors influence the electronic motion. Most common of these factors is electron scattering with atoms of the host lattice. Other factors include scattering with impurity atoms, other electrons, surfaces, and interfaces between dissimilar media, as well as impact ionization. Lattice scattering involves the transfer of kinetic energy from an electron to vibrational energy of the host lattice. This energy transfer occurs in discrete quanta of vibrational energy known as phonons [8–10]. Since the initial presence of lattice vibration distorts the lattice and actually enables the scattering process, one often speaks of electron–phonon scattering.

The result of all these scattering mechanisms is that an electron will not accelerate indefinitely in response to an electric field. Rather, scattering tends to redirect the electron momentum and, in many cases, dissipates the electron energy gained from the electric field. In a uniform electric field, then, the processes of energy gain (from the field) and energy loss (from scattering) balance, and the electron attains a constant velocity as opposed to the constant acceleration of the free space electron [11].

As this electric field strength increases, both the average electron velocity and the average electron energy increase. Of course, high electron velocity is desirable since current is proportional to velocity. Unfortunately, velocity tends to saturate with increasing field, and this tendency has negative, inescapable consequences for semiconductor device performance. High electron energy, on the other hand, can be detrimental to semiconductor devices. (The terms "hot" electron and "high energy" electron are synonymous.) For example, an electron with energy in excess of the conduction band minimum by more than the semiconductor bandgap can transfer its energy to a valence electron. Thus, the initial energetic conduction electron loses energy and generates an additional conduction electron and valence hole. This impact ionization process clearly adds more mobile charge carriers and suffers from positive feedback since the generated carriers will gain energy from the electric field and produce more ionization events. Excessive impact ionization is detrimental since the resulting high current

is generally not consistent with the circuit function of the device (a reverse-biased diode, for example), and this current may easily damage circuit components.

A more vexing and complicated fallout of high energy electrons arises in the presence of a semiconductor–dielectric interface. (We think specifically of the silicon–silicon dioxide interface since this is the only system that has been studied in this manner and is the only relevant system.) In an *n*-channel MOSFET, a dielectric layer separates the gate from the channel. Ideally, electrons flow in the channel since an energy barrier, about 3.2 eV [12] for silicon dioxide, suppresses electron injection into the gate dielectric. The presence of this barrier is important since, along most of the channel, there exists an electric field component pulling the negative electrons toward the gate. With the presence of the barrier, this field component merely acts to constrain the electrons to the silicon surface.

Since the average electron energy increases as the electric field strength increases, there will exist a point at which some significant fraction of electrons possesses enough energy to surmount the silicon–silicon dioxide energy barrier and jump into the gate oxide. This electron emission from the channel manifests itself in several ways. First, one may measure the electron current in the gate lead [1,13]. In all MOSFETs of practical importance, this current is negligibly small and thus does not impede proper circuit operation. In fact, this gate current is generally difficult to measure. When measurable, the gate current is one useful monitor of hot electron reliability.

Since the mean electron energy is a reasonably strong function of the electric field [14] and the fraction of hot electrons (with energy greater than the oxide barrier height) is an exponential function of this mean energy, one expects most of the hot electron emission into the dielectric where the channel electric field is greatest. Figure 5.1 plots the lateral (source-to-drain) component of the electric field in the channel of an *n*-channel MOSFET. The rapid rise of this field, determined by numerical simulation, near the drain is typical and clearly implies that we expect maximum electron emission near the MOSFET drain. A corollary to the observation that electron emission is greatest in the portion of the channel at which the electric field is maximized is that any modification of applied voltage or processing conditions that increases the electric field strength will also increase hot electron emission.

The dependence of gate current on gate and drain voltages is instructive. Figure 5.2 plots measured gate current as a function of gate bias with drain bias as a parameter. For fixed drain bias, the gate current peaks at a particular gate bias. The decline in gate current beyond the maximum is due to the decreasing channel electric field with increasing gate bias. (The

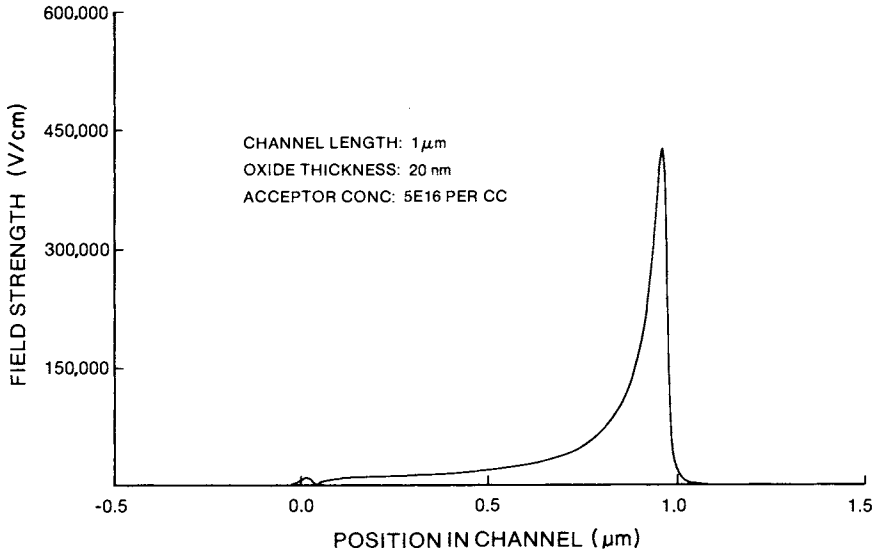


Fig. 5.1. Lateral component of the surface electric field strength within an *n*-channel MOSFET.

lateral component of the channel electric field decreases monotonically with increasing gate bias. One may consider this as a "field pinch" effect between the gate and drain equipotential contours.) As the channel field decreases beyond the maximum, so does the hot electron emission. Of course, we expect that electron emission would increase as gate bias decreases from the maximum point for this same reason. But there is a complicating factor. With a decreasing gate bias, electrons injected onto the gate oxide will experience a repulsive force. That is, the gate terminal is at lower potential than the drain so that electrons will jump over the energy barrier into the oxide and simply return to the MOSFET channel. Since only electrons that reach the gate are measured as gate current, this current will decrease as gate bias decreases even when electron emission at the silicon–silicon dioxide interface is increasing (because of the higher electric field). Thus, the mechanisms of increasing field and increasing barrier for gate collection compete to produce the single maximum. As drain bias increases, hot electron emission increases because of increased channel field. Negligible differences in the gate current at different drain biases and low gate bias are evidence of a moving electron injection "window" within the channel where the channel-to-gate potential barrier is nearly zero [13]. We emphasize that this window concept applies only to those electrons that eventually reach the gate. Even with very low gate bias (and high drain bias), the exceedingly low gate current indicates only that few electrons

reach the gate. But a prodigious number, driven by the high drain bias, may still surmount the energy barrier into the oxide without reaching the gate.

Another result of hot electron emission into the gate oxide is electron trapping within the oxide. Typical silicon dioxide films trap only about one percent of the total number of injected electrons [15,16]. Electron trapping, of course, results in a negative charge accumulation within the gate oxide and thus will eventually change the MOSFET behavior. Specifically,

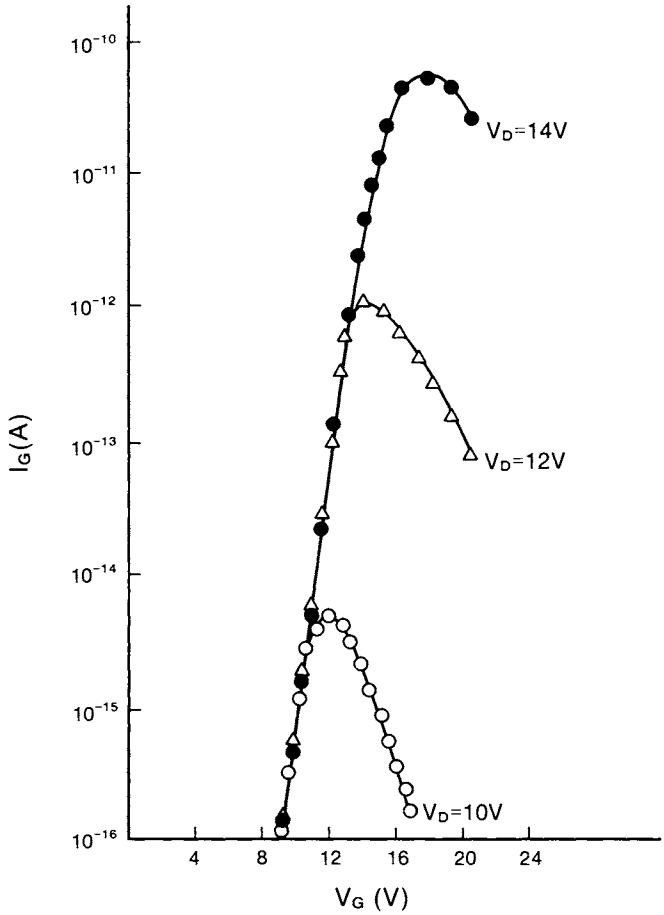


Fig. 5.2. Dependence of gate current on gate bias with drain bias as a parameter. In these measurements the FET gate length and width are 3.8 μm and 24 μm, respectively. The channel impurity concentration is 1.6×10^{16}/cc and the gate oxide is 800 A thick. (Reprinted from Eitan and Frohman-Bentchkowsky [13].) © 1981 IEEE.

the addition of negative charge raises the MOSFET threshold voltage since a more positive gate bias is required to overcome the negative oxide charge. Analytical representation of this threshold shift is nearly impossible since, as one might suspect, the trapped electrons are highly localized near the drain because of the strong dependence of injection on electric field strength. Increased threshold voltage is detrimental since the MOSFET current drive decreases. The degradation in MOSFET performance is minimal in the saturation (pentode) regime and much more pronounced either in the linear (triode) regime or with source and drain terminals reversed so that the hot electron damage is localized at the source end of the channel.

Quite intriguing is the well-known result that hot electron degradation consists also of interface states [17–21]. Interface states are electron energy states within the forbidden silicon band gap at the silicon–silicon dioxide interface. These states are detrimental since they can temporarily trap charge and thus change MOSFET characteristics or reduce charge carrier mobility within the channel. Charge pumping [18,19] and low-frequency noise measurements [21] have confirmed the localized nature of these surface states. The mechanism of this surface state generation is unknown, but there exist many speculations [19,22]. It is reasonable to suspect that surface state generation occurs here by the same mechanism as that observed in one-dimensional electron injection experiments [23]. In addition to increased threshold voltage, another result of hot electron degradation is reduced linear region transconductance. In fact, it is more common to report transconductance data since this parameter tends to degrade before one observes any changes in threshold voltage [24–26]. Transconductance degradation is often loosely explained by channel mobility reduction due to surface state generation, but modeling studies have shown that localized electron trapping can also explain the loss in transconductance [27].

The qualitative picture we have drawn of CHE reliability is straightforward. Electrons flowing from source to drain within the NMOS FET channel gain energy from the lateral electric field component. This energy gain permits simple emission over the energy barrier into the gate oxide with possible electron trapping within this gate dielectric or, by some unknown mechanism, the energetic electron generates interface states. One may find other accounts in the literature. For example, one proposal states that the channel electrons gain energy, produce electron–hole pairs by impact ionization, and the *generated* electrons then gain energy and jump into the gate dielectric [4–6]. The proponents of this DAHC theory note that CHE degradation and substrate (avalanche) current are maximized at roughly the same bias conditions. This observation then motivates the claim that it is actually the avalanche, not the primary, carriers that damage the MOSFET. This claim is specious and unsubstantiated since

there exists no physical basis for the expectation that secondary (avalanche) carriers could cause more degradation than primary carriers. In fact, just the opposite is true. The primary carriers are, on average, more energetic and outnumber the secondary carriers by at least an order of magnitude. It is not surprising that CHE degradation and substrate current, two manifestations of high electric fields within the MOSFET, would exhibit similar dependences on MOSFET parameters in the absence of any direct, causal connection.

Another school of thought emphasizes the importance of hot holes in NMOS FET degradation [18,22]. In the NMOS channel, holes arise by avalanche ionization and are thus found in far lower concentrations than are electrons. Holes will gain energy from the electric field but this energy gain is less efficient than is electron energy gain due to the lower hole mobility. Given the paucity of holes in the one-dimensional electron injection experiments [23] in which device degradation is observed, we believe it unlikely that hot holes are required to explain MOSFET CHE degradation. We must note, though, that extremely sensitive gate current measurements find net *hole* injection at sufficiently low gate bias [28]. Thus, hot holes do exist in the NMOS FET channel, but their relevance in CHE degradation is not clearly established. Perhaps the strongest argument in favor of a hot hole damage mechanism is the observation of enhanced MOSFET degradation under pulsed stress [29,30]. This experimental result is difficult to reconcile and the only model of which we are aware is the hot hole model of Weber *et al.* [29,30]. Fair and Sun [31] might argue that the holes are supplied by the substrate in the collapse of the depletion region as the gate bias switches low.

Channel hot electron degradation increases in severity as the *n*-channel MOSFET channel length decreases. As we noted previously, hot electron injection is a strong (exponential) function of the mean electron energy, and this mean energy increases with increasing electric field. If the design and process engineers adhered to the tenets of ideal, constant field (CE) [32,33] scaling discussed in Chapter 2 for the CMOS miniaturization, we would expect no increase in CHE degradation. But "real world" considerations often render pure CE scaling unattractive. Quite often, for example, the circuit power supply is set by system compatibility requirements and cannot be prescribed arbitrarily by scaling considerations. Noise margin requirements, discussed in Chapter 2, coupled with subthreshold leakage concerns, place a lower bound on acceptable power supply.

Consider the scaling regimen that approximates CE scaling with the important exception of a failure to reduce the power supply. This is constant voltage (CV) scaling [34]. One specifies reduced channel length, gate oxide thickness, and source–drain junction depth, as well as increased

channel acceptor impurity concentration. All of these modifications act to increase the maximum electric field in the channel and thus exacerbate the hot electron problem. While most obvious, but not most important, channel length reduction produces a shorter distance across which the source–drain potential difference must "drop" and hence increases the mean channel electric field. The maximum channel field does not increase as quickly as the inverse channel length since, independent of channel length, the channel field is nonuniform and is strongest near the drain. Figure 5.3 plots the lateral component of the channel field for two channel lengths with all other parameters constant.

Clearly, reduction of the gate oxide thickness will increase the transverse (gate-to-channel) electric field component. Less obvious is that the diminished oxide thickness also increases the lateral channel field component. Consider the portion of the MOSFET near the drain. The gate and drain terminals are separated only by the gate oxide and may be biased at different potentials. The gate and drain, because of high conductivity, form two equipotential contours. In the narrow separation between these equipotentials, one may draw an equipotential for any intermediate bias. These contours "bend down" into the silicon channel and the length separation between adjacent contours determines the local electric field. If one decreases the gate oxide thickness, the equipotential contours squeeze to-

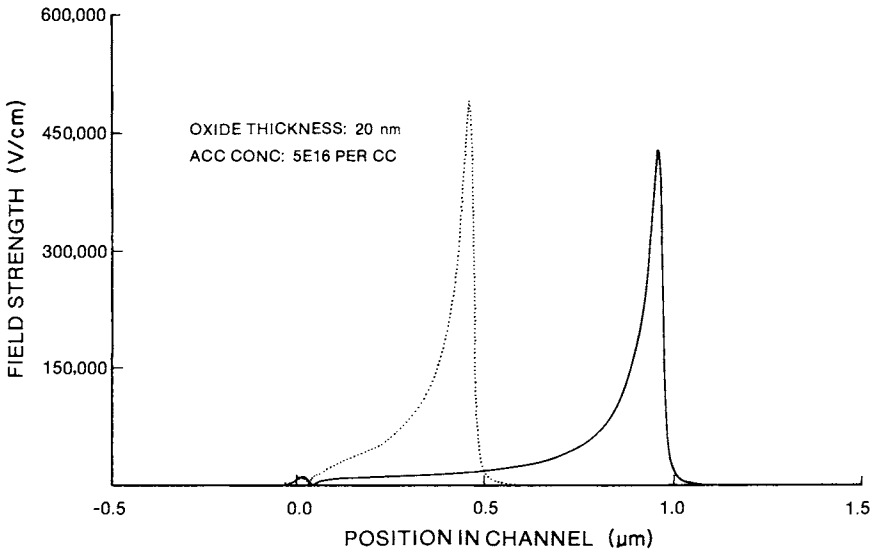


Fig. 5.3. Lateral component of the surface electric field strength within an *n*-channel MOSFET as a function of the channel length.

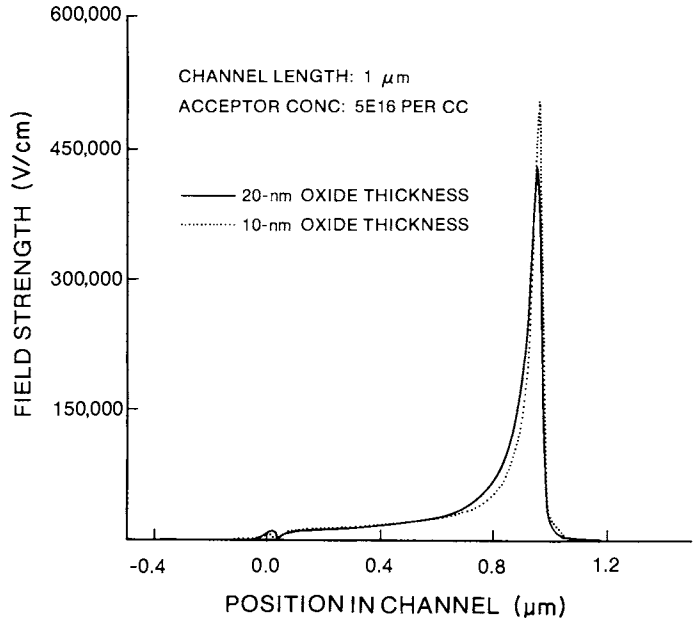


Fig. 5.4. Lateral component of the surface electric field strength within an *n*-channel MOSFET as a function of the gate oxide thickness.

gether and reduce the separation between adjacent contours where these contours drop down into the channel as well as simply between the gate and drain. Figures 5.4(a) and 5.4(b) plot the lateral component of the channel field for two values of the gate oxide thickness.

Minimizing the source–drain junction depth indirectly increases the maximum channel field. Processing methods that reduce junction depth also tend to reduce the lateral extent of the source–drain regions beneath the gate. As a result source–channel and drain–channel $n+-p$ junctions become more abrupt. One method for achieving this abruptness is simply the substitution of arsenic (relatively slow diffuser) for phosphorus (relatively fast diffuser) as the source–drain donor impurity. Since the maximum electric field in a reverse-biased junction is greatest when the doping profile is abrupt, reduction of the lateral extent of the source–drain impurity profile will increase the channel field strength. Similarly, increasing the channel acceptor concentration also increases the reverse-biased diode field strength. Surprisingly, through, the dependence of field strength on channel acceptor concentration is not as strong as one might suspect from one-dimensional analogies. Figures 5.5(a) and 5.5(b) plot the lateral field component for two values of the channel acceptor concentration.

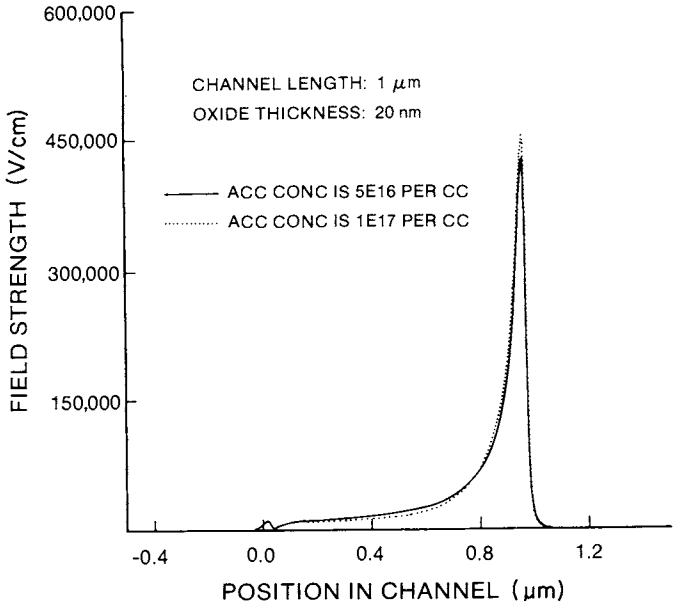


Fig. 5.5. Lateral component of the surface electric field strength within an *n*-channel MOSFET as a function of the channel impurity concentration.

The march toward CMOS miniaturization and the reluctance to reduce power supply voltage force us to grapple with the hot electron problem. One may imagine two strategies: reduction of electron emission into the gate oxide or "hardening" of the gate oxide and interface against degradation due to hot electron emission. The latter possibility does not appear to be ridiculous since there is great experience and success in producing gate oxide films with reduced hole trapping for ionizing radiation environments [35,36]. Unfortunately, even though some ideas have appeared meritorious [37], no concrete approaches for dielectric hardening against hot electron emission have materialized.

All successful, reproducible processing methods for minimizing hot electron degradation are predicated on reduction of the lateral component of the electric field strength in the MOSFET channel. Furthermore, all of these methods deal with the fabrication of the source–drain regions. In terms of processing complexity and effectiveness, the first choice is the replacement of arsenic by phosphorus for the source–drain donor impurity [38]. The additional lateral diffusion of phosphorus "grades" the drain–channel junction and thus reduces the electric field strength. The disadvantage of this approach is precisely this lateral (as well as vertical) diffusion. At some point, this lateral encroachment will shorten the effec-

tive MOSFET channel length enough to precipitate source–drain punch-through. Thus, phosphorus is not suitable for MOSFETs with channel length shorter than some critical value. Experience suggests that channel lengths of 1.25 μm and less require arsenic as the principal source–drain donor impurity. While one may reasonably note that the increased depth and lateral extent of phosphorus source–drain regions will burden the MOSFET with additional capacitance, increased drive current due to the reduced effective channel length will likely compensate for the increased capacitance.

While phosphorus alone may be a poor choice for the source–drain donor impurity in short-channel NMOS FETs, a combination of arsenic and phosphorus will provide some improvement in hot electron reliability [38,39,40] without excessive lateral encroachment within the MOSFET channel. Arsenic is the dominant impurity and phosphorus is added in the hope that it will add a "tail" to the lateral donor impurity profile in order to reduce the abruptness of the drain–channel junction. This "double-diffused" method also enjoys the advantage of negligible additional complexity for process integration. Implementation of the double-diffused source–drain process requires a great deal of process and device calibration. Too much phosphorus or excessive diffusion will yield punch-through difficulties, while too little phosphorus will provide inadequate protection from hot electron problems.

The most widely accepted processing remedy for hot electron reliability is the LDD structure [41,42]. Figures 5.6(a)–5.6(c) sketch this concept. In Fig. 5.6(a), we draw the conventional MOSFET structure with a light donor ion implant masked by the gate. The implantation dose of this LDD implant is one to two orders of magnitude less than that of the source–drain region. After this implant, a conformal deposition of silicon dioxide follows. Such a deposition leaves a thicker film on the edge of the gate-to-active area step than that on flat areas. A highly anisotropic reactive ion etching step then clears the deposited oxide from the flat areas and leaves remnants at all step edges. We now reach the sketch of Fig. 5.6(b). This sidewall oxide spacer process has thus generated a self-aligned protuberance of silicon dioxide on the edges of all gates. The final step performs the high-dose source–drain ion implantation as depicted in Fig. 5.6(c).

The LDD region provides a gradual transition between the highly doped source–drain region and the p-type channel. Figure 5.7 plots the lateral component of the electric field with and without the LDD region. Furthermore, by varying the LDD implant dose and spacer oxide thickness, one has direct control over the LDD donor impurity concentration and LDD length. Wordeman *et al.* have actually employed the LDD concept to effectively satisfy CE scaling demands for shallow junction depth [43]. As

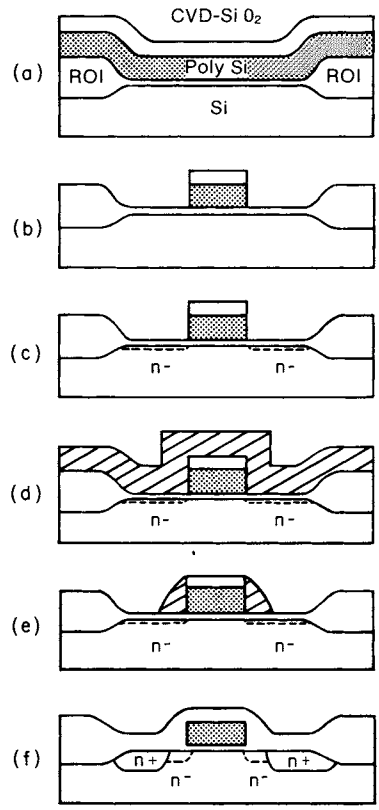


Fig. 5.6. This is a pictorial description of the LDD fabrication process. The steps are: (a) formation of recessed oxide insulator and polycrystalline silicon electrode, (b) reactive ion etch of the stack of films, (c) spacer ion implantation, (d) spacer oxide deposition, (e) reactive ion etch of spacer oxide, and (f) source/drain ion implantation. (From Tsang *et al.* [41].) © 1982 IEEE.

one might suspect from Fig. 5.7, the LDD structure yields significant gains in hot electron reliability compared to the conventional MOSFET as shown in the Fig. 5.8 data of threshold voltage shift versus stress time [44]. Another variant of the LDD, known as the buried spacer, simply specifies a high-energy LDD implant in order to place the *n*-type LDD region beneath the silicon–silicon dioxide interface with the idea that the highest energy electrons are too far removed from this interface to produce any damage [45,46]. Figure 5.9 portrays benefits of this variant.

But the LDD concept exacts a large cost. The reactive ion etching must be precise since there is no natural etch stop. Overetching directly degrades the beneficial aspects of the spacer. Furthermore, the added oxide deposi-

tion must be conformal, uniform, and reproducible. Current drive and transconductance of the LDD MOSFET are less than those of the conventional MOSFET since we have effectively added series resistors (the LDD regions) to the source and drain regions. On the other hand, the reduced Miller capacitance of the LDD reclaims some of this performance degradation [41,47]. An additional benefit of the LDD is its role as an insulator between the gate and source/drain. As discussed in Chapter 5, self-aligned metallization of the gate and source/drain allows increased interconnection and improved circuit performance. The oxide spacer prevents shorting of the gate and source/drain. Thus, if the LDD structure is required for the self-aligned metallization, it is certainly reasonable to choose the LDD approach to hot electron reliability than, for example, the double-diffused source/drain method.

Having discussed the physical mechanism of CHE degradation and the prevalent processing options for minimizing this problem, we must next delve into the definition, measurement, and monitoring of device and circuit lifetime due to CHE stress. We consider device lifetime and circuit lifetime as separate issues and proceed first with the former. CHE degradation will always act to decrease MOSFET current drive. Since catastrophic problems, such as open lines or excessive subthreshold leakage, do not result from hot electron stress, the device designer must only worry about the performance loss inherent to reduced source/drain current. In this

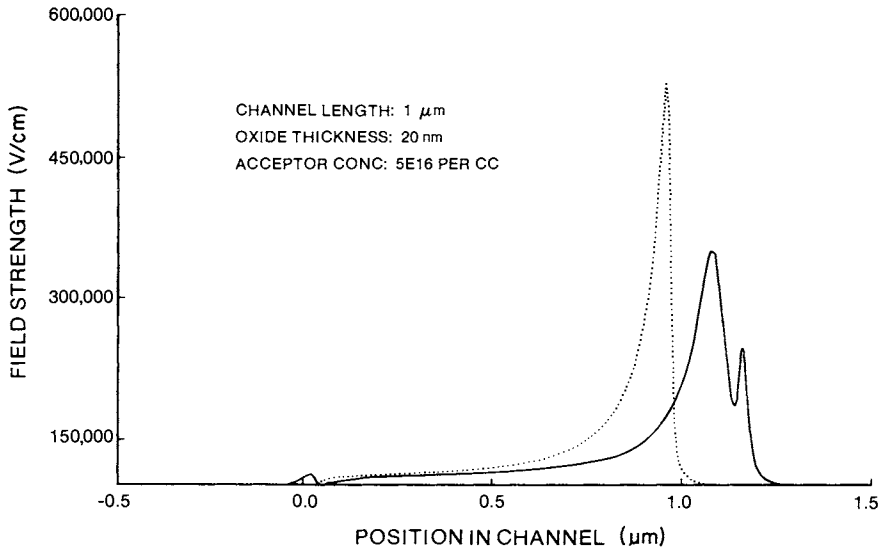


Fig. 5.7. Lateral component of the surface electric field strength within an *n*-channel MOSFET with and without the LDD extension.

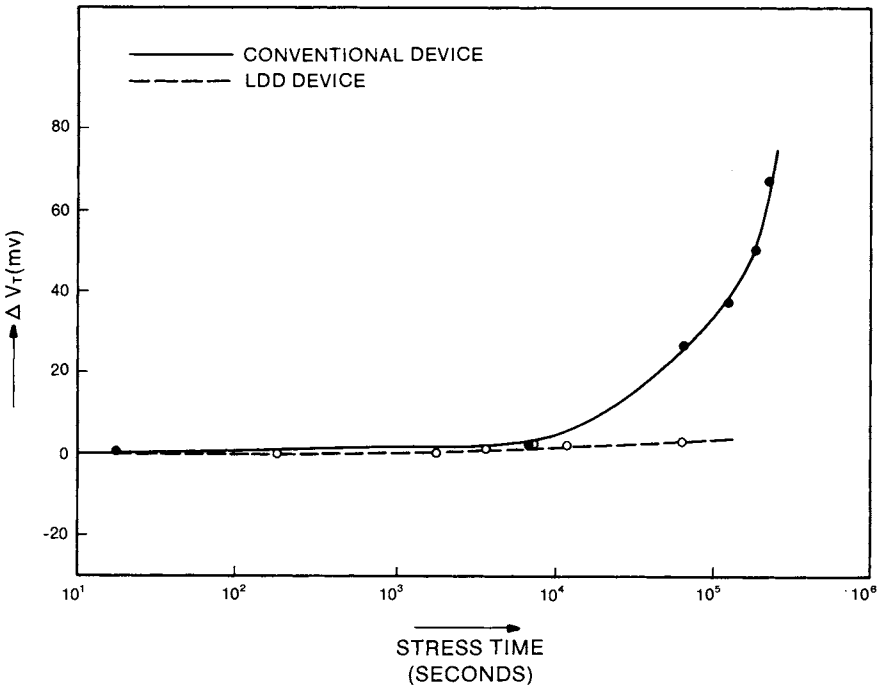


Fig. 5.8. Measured degradation in MOSFET threshold voltage under hot electron stress with and without the LDD extension from Baglee and Duvvury [44]. © 1984 IEEE.

respect, definition of a device lifetime is a complicated task since the imposition of a black-and-white pass/fail criterion will always be, to some degree, arbitrary.

Ning *et al.* approached this issue sensibly [1]. These authors described an NMOS process with a minimum gate length of 1.3 μm $\pm$ 0.3 μm. They reasoned that the shortest allowable gate length of 1.0 μm would exhibit the worst CHE reliability but would also possess the highest initial transconductance. Ning *et al.* "allowed" this short-channel MOSFET to absorb its initial transconductance advantage by CHE degradation. Thus, the device lifetime was specified by the requirement that the 1.0-μm MOSFET could lose no more than 30% of its initial transconductance. Furthermore, to compensate for the fact that a given MOSFET is biased above threshold only a fraction of the time in real circuit operation, the continuous, single-device stress time for a MOSFET is multiplied by an appropriate factor for translation to an effective circuit lifetime. An estimation of a 3% duty cycle (fraction of time in which the MOSFET is "on") allows a 10-yr (10^5 hr) circuit operation requirement to be assessed by continuous MOSFET stress in 3000 hr (about 4 months).

This concept of stressing a single MOSFET continuously as a means of accelerating the effective circuit stress is of great importance since it is clearly impractical to stress an entire circuit for any period of time approaching the typical field lifetime of 10 yr. But the continuous MOSFET stress poses additional questions. The specification of an allowable transconductance loss or threshold voltage shift trivializes the complicated dependence of circuit performance on individual MOSFET properties. For example, the concern with circuit "speed" usually arises when a MOSFET is switched on (gate above threshold) in order to discharge the initially high-voltage, floating drain terminal. The time required to discharge the drain (i.e., bring the drain to the source potential) is a function of the source/drain current at all source–drain potential differences. Since the criteria of transconductance loss and threshold voltage shift are specified only for the (worst case) linear region, these criteria may bear little resemblance to circuit performance issues. For example, delay time in a CMOS inverter can be virtually unaffected by CHE stress while the constituent *n*-channel MOSFET shows "unacceptable" linear transconductance loss [48,49]. Another difficulty of the single, continuous MOSFET stress is the choice of stress biases. The worst-case drain bias is the maximum (power

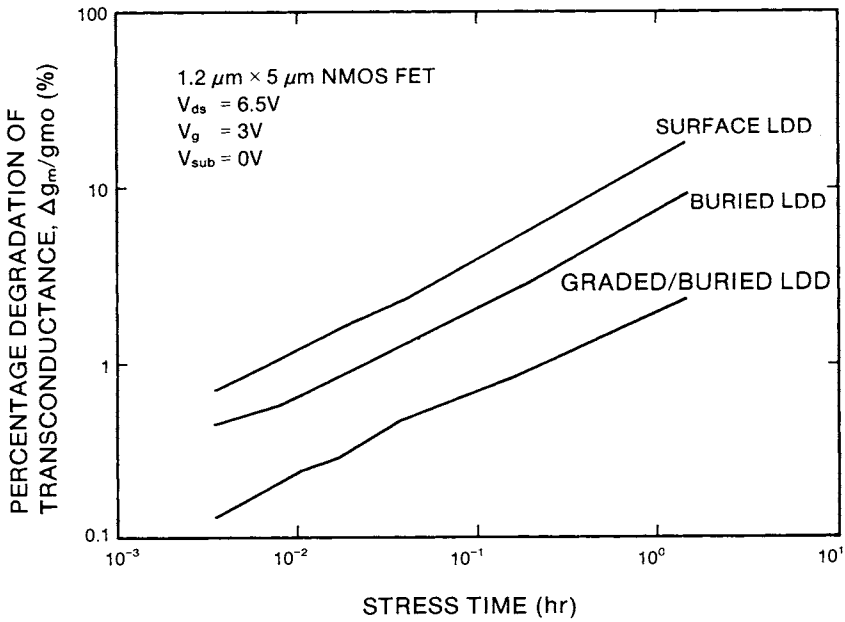


Fig. 5.9. Measured degradation in MOSFET transconductance under hot electron stress for the conventional and "buried" LDD structures. (Reprinted from Wei *et al.* [46].) © 1986 IEEE.

supply) bias. The worst-case gate bias is significantly lower than the maximum bias [50] since, as discussed previously, the lateral channel field component decreases with increasing gate bias. While it is certainly possible to find worst-case gate and drain biases, the substitution of these biases for the more complex circuit waveforms should grossly exaggerate the CHE reliability problem. One might consider such an exaggeration to be harmless, and even desirable, if the MOSFET passes this continuous reliability test. We disagree. Presumably the process engineer has sacrificed process manufacturability (with the LDD process) and MOSFET performance to achieve this unnecessary level of CHE reliability.

In any event, the most prevalent monitor of CHE lifetime specifies a continuous stress of several thousand hours in which one follows the degradation of a single, linear region parameter, such as maximum transconductance. Subject to the criticisms of the relevance of such a simple device characterization to circuit performance, this continuous stress is the best option available to the device engineer for assessing CHE reliability. We must note, however, the disconcerting observation that pulsing the gate bias yields more hot electron damage, for equal FET on time, than continuous stress [29,30]. This experimental result challenges the implicit assumption that we may accelerate the circuit stress by maintaining a continuous stress for a much shorter time. The resolution of this conflict may be that electrons in the channel while the gate is biased above threshold "see" an increasing lateral component of electric field as the gate bias decreases. Thus, the gate turn-off actually heats the electrons more than the constant (high) gate bias so that each turn-off adds more damage to the MOSFET.

The continuous stress requires several thousand hours and thus is inconvenient. One would greatly prefer a much shorter assessment of CHE lifetime. One research report noted a strong correlation between CHE lifetime and substrate current [50] and developed an empirical, quantitative relationship between these two parameters. Such a relationship would be ideal since one may easily measure the substrate current at wafer level. Unfortunately, we did not find experimental agreement with this study in our laboratory. Furthermore, we do not believe that substrate current will correctly predict qualitatively the trend in CHE lifetime as one varies, intentionally or unintentionally, the MOSFET channel length [50]. A vastly superior qualitative assessment is given by the ratio of substrate current to source (or drain) current [51]. This ratio measures the avalanche multiplication factor and, given the exponential dependence of this factor on the maximum channel electric field, the substrate–source current ratio is an excellent monitor of CHE reliability. Therefore, we recommend the incorporation of this current ratio measurement along with all other automated, wafer level test element measurements. These data, combined with

the experience gained from the continuous stress measurements of several thousand hours, give the process engineer quick feedback if CHE reliability begins to drift. Still, the continuous stress measurements are necessary in order to establish definitively the device lifetime.

As we shall discuss with electromigration, it is sometimes possible to accelerate a reliability test based on temperature dependence. With CHE reliability, the lifetime decreases as the temperature decreases. This apparently paradoxical result arises from the observation that an electron in a semiconductor experiences fewer collisions as the lattice temperature decreases. With fewer collisions to dissipate the electron energy gained by the electric field, the mean electron energy increases as the temperature decreases. Thus, while lowering the temperature will indeed accelerate the CHE reliability test, the acceleration factor is unknown.

Another interesting manifestation of hot electrons in the *n*-channel MOSFET is photon emission [52]. The physical mechanism of this emission, equivalent to that in a reverse-biased diode [53], is the radiative intraband and interband demotion of conduction electrons. Designation of bremsstrahlung [52] as the cause, presumably related to electron scattering, appears to us to be unsubstantiated. Clearly, this photon emission would provide an inconvenient monitor of CHE reliability and is likely of little practical importance.

Several recent reports attempt to model hot electron reliability [54]. This problem is challenging since the physics, even without the numerical computation, of high-field semiconductor transport is quite complicated. The conventional formulation of the semiconductor device equations [55] specifies conduction electron and valence hole concentration with no information whatsoever on electron or hole energies. A complete modeling attempt should then reformulate the device equations to include this energy information [14,15]. A simpler approach is to take the results of conventional numerical device simulators and, in some approximate yet plausible manner, infer the mean electron energy as a function of position. Such an attempt has succeeded with the aid of an adjustable parameter [38]. For practical application, we find that numerical simulation of the lateral field component is adequate for aiding device design.

Given that the ultimate minimum design rule is roughly $0.2-0.3\ \mu m$ [56,57], we predict that channel hot electron reliability will degrade from that of the present $1.25\text{-}\mu m$ design rule technology. This prediction stems from the observation that the power supply cannot scale down with the design rule (for room temperature operation), and thus we cannot even approximate constant field scaling. In fact, CHE degradation may be significantly enhanced. But the device engineers will find that, as we discussed, circuit operation will not suffer nearly as much as the observed

damage to single MOSFETs under continuous stress. Thus, we predict that CHE reliability will not limit CMOS miniaturization.

II. ELECTROMIGRATION

The metallic interconnections of devices suffer from reliability problems as do the devices themselves. Patterned metallization serves to connect, ideally with zero resistance, the integrated devices with each other and circuit input and output terminals. The interconnection function implies current conduction within the metallization and this current flow degrades the integrity of the metal. Even if one reduced physical dimensions and applied biases in CMOS miniaturization under ideal scaling rules [32,33], one would still observe a multiplicative increase in the conductor current density by the scaling factor. Typical deviations from ideal scaling will certainly lead to a more rapid increase in this current density and thus exacerbate any existing inherent degradation of metallization by current flow.

The mechanism by which current flow degrades metallization is interesting. Consider the positively charged metallic ions sitting in lattice positions. In metals, of course, the conduction electron density is quite high. Current within the conductor is carried predominantly by these electrons, and these electrons collide frequently with the lattice ions. Though the metallic ion finds itself in an energy minimum as it resides in its lattice site, the ion may spontaneously gain enough energy from either the thermal energy of the lattice or the electron collisions to vacate the lattice site and occupy an interstitial position. Electron–ion scattering increases greatly when the ion occupies such an interstitial position [58]. This scattering transfers momentum from the electrons to the ion and thus exerts a force on the ion in the direction of electron flow. Thus, there develops a net transport of positively charged ions in the direction of the most positive potential. Certainly there exists an opposing force due to the conductor electric field on the positive ion. This Coulomb force is generally small as one might infer from the observation that an ideal conductor would exhibit zero resistance and thus zero electric field while the "electron wind" force would persist. Electron screening further reduces the effective Coulomb force. Ironically, then, one will find net transport of the metallic ions in the same direction as the electron flow in metals in which conduction electrons dominate current flow. The term "electromigration" denotes this current-induced metal ion transport.

Electromigration leads to integrated circuit failure. An ion leaving a

lattice site leaves behind a vacancy. A neighboring ion "upwind" of the vacancy may jump to the vacated lattice site with far greater probability than a neighboring ion "downwind" of the vacancy. The result is a net transport of ions in the direction of electron flow and an equivalent transport of vacancies in the opposite direction. The excess ions tend to coalesce to form hillocks and extrusions [59,60], as do the vacancies. The vacancy coalescence simply leads to macroscopic gaps in the metal film, and if uninterrupted, this gap will grow to the entire cross section of the film and thus "open" the metal line. The net effect of electromigration, then, is to redistribute the metallization into regions of excess metal (hillocks and extrusions) and depleted metal (voids). The excess metal may also lead to circuit failure [61–63]. Extrusions will grow and bridge to another metallization line. This unintended short circuit renders proper circuit function impossible. Clearly, the short-circuit lifetime will depend on the closest separation of neighboring metallization lines as well as the physical dimensions and current density of the primary line. The extrusion length does not scale! On the other hand, one may also note that the critical size a macroscopic void must attain for the open-circuit failure mode decreases as circuit dimensions decrease. Thus, one should expect electromigration to degrade more quickly than indicated merely by increased conductor current density.

Early research in electromigration studied macroscopic void motion directly. Huntington and Grone [58] measured the displacement as a function of time of razor scratches on gold wires. The scratches (voids) travel toward the cathode (negative terminal) indicating metallic ion transport toward the anode. Furthermore, the scratch velocity was proportional to the current density and to the exponential of a negative activation energy divided by kT (thermal energy). This activation energy closely approximated the known activation energy of gold self-diffusion. Thus, the observation that increasing temperature increases the electromigration phenomenon is consistent with the picture of random thermal excitation of lattice ions into interstitial positions followed by transport due to momentum exchange from conduction electrons. One immediately infers that metals with inherently greater immunity to such thermal excitation of the lattice ions (i.e., higher melting temperature) will exhibit correspondingly less electromigration.

The electrical engineering community soon learned that this fascinating experimental result from a solid-state physics laboratory translated to a reliability problem in the integrated circuit structure [59,60]. Beyond qualitative elucidation of the physical mechanism, the Huntington–Grone work provided valuable background on the current density and temperature dependence of electromigration. Based on these observations, one is

able to estimate the conductor line mean time to failure (MTTF) for open-circuit failure as [59,60]:

$$\text{MTTF} = A\, J^{-n} \exp(qE/kT). \tag{5.1}$$

In Eq. (5.1), J is the current density, E is the activation energy which appears to be equivalent to that of metallic self-diffusion, and A is a proportionality constant.

We list the J exponent as $-n$ since there is some disagreement on the correct value of this exponent. Various theories claim values of 1 [64] or 2 [59,60], while experimental results show greater diversity [65]. The experimental difficulty in isolating the exponent n arises from complications such as joule heating of the wire at high-current density. Clearly, the increased temperature will lead to greater electromigration. The experimentalist must proceed cautiously in the separation of current density and temperature. Furthermore, as voids and hillocks develop, the current density will vary locally, and the positive feedback of increased current density at voids and reduced current density at hillocks will accelerate electromigration failure. In this stage, the electromigration is driven by a different current density than is "applied" and one could be easily misled in a study of current density dependence of MTTF. A recent careful examination of this current density dependence measured an exponent of 1.53 ± 0.02 [66].

Experimental data suggests that the proportionality constant A of Eq. (5.1) is proportional to the conductor cross-sectional area [59]. Note that this is in addition to the current density dependence (which increases with reduced cross-sectional area at constant current). This additional area dependence denotes the observation that a void must attain the dimensions of the cross-sectional area in order to precipitate open circuit failure.

Presumably, by the same reasoning, the analogous MTTF for short-circuit failure would feature a proportionality constant that varies directly as the separation to the nearest conductor. This presumption is partially verified by an experiment of Towner [61]. The Towner study investigated the MTTF of short-circuit failure to an overlying conductor separated by a deposited dielectric from the conductive line under stress. The MTTF increased linearly with dielectric thickness, but the linear extrapolation did not intersect the MTTF-dielectric thickness origin. Thus, the pure linear dependence of MTTF on conductor separation (dielectric thickness in this case) that one expects from simple considerations is only approximately replicated by experiment.

Finally, the activation energy E of Eq. (5.1) can also vary. As we hinted previously, refractory (high melting point) metals will have a greater activation energy than metals with lower melting points. Even for the same

metal, though, the activation energy can vary. For example, in small grain metal films, ion transport proceeds predominantly at grain boundaries since the migration energy is lower at these boundaries. Large grain films possess fewer grain boundaries and ion transport is impeded [60]. Surface ion migration is also possible at lower energy cost than bulk transport in the absence of metal passivation. Passivating the metal inhibits the surface migration component and thus increases the observed activation energy [60,67,68].

The intent of formulating the MTTF expression of Eq. (5.1) is not to forecast the electromigration lifetime for arbitrary design rule, current density, and metallization system. Rather, one exploits the expected increase of electromigration with current density and temperature by specifying experimental stress conditions that are more severe than expected operation. The stress then accelerates the expected degradation and it is possible to simulate years of normal operating life in seconds under the high-current/high-temperature stress. The key to this critically important issue of screening wafer lots for electromigration lifetimes of, for example, ten years, by accelerated stressing of test elements lies in accurate estimation of the acceleration factor. This estimation is complicated by joule heating and other factors as noted previously for the determination of MTTF on current density. Recent studies have demonstrated reasonable and successful methods for accelerated electromigration testing [69]. Of great importance is the requirement that the test element for electromigration assessment be equivalent to the circuit metallization in terms of metal pitch (for short-circuit failure), wafer topography (for metallization steps), and passivation. We would also add that stress tests specify a constant current within the conductor as opposed to a constant voltage drop. Even though one applies voltages in real circuit operation, the adoption of a constant current stress is appropriate since the conductor will always appear in series with more resistive elements, such as MOSFETs. Since the resistance of these other elements determines the circuit current level, one expects candidate conductors with widely varying resistivities to handle equivalent current levels.

Since the current density carried by the conductive lines increases as design rules shrink, the electromigration MTTF decreases. The fabrication process must therefore evolve such that this trend in electromigration is minimized or negated. In fact, the Chapter 3 discussion of metallization alludes frequently to electromigration constraints. In LSI (large scale integration), aluminum is the dominant choice for metallization because of its high conductivity, low-resistance ohmic contacts to silicon, manufacturability (deposition and etching), and bondability. With all these advantages, aluminum compares poorly with most other metals with respect to elec-

tromigration. The low aluminum melting point (660°C [70]) suggests this electromigration susceptibility. Thus, this reliability hazard dictates a departure from conventional aluminum metallization. We now discuss several alternatives to aluminum. One must recognize that all alternatives are inferior in some aspect of manufacturability or performance.

The addition of copper to aluminum, typically up to about 5% at. wt., depresses electromigration [71–74]. The copper additive reduces "electron-driven" aluminum diffusion within grain boundaries. The relatively small copper content in this aluminum–copper alloy essentially maintains many of the processing advantages of aluminum. But disadvantages do exist. This alloy tends to corrode and is also difficult to etch [75] by plasma (dry) methods. The latter etching difficulty is nonexistent if one may pattern the metal by a lift-off technique [76], but this is often not the case. A recent study suggested adding 0.2% titanium to an aluminum–silicon (1.2%) alloy [75]. The titanium also acts to reduce the aluminum electromigration with no penalty in corrosion or etching. This titanium addition appears to be the best method for modifying aluminum for extended MTTF.

Placing a barrier layer, such as a titanium–tungsten (Ti-W) alloy [61] or a chromium film [77], beneath an aluminum film is another possibility. Such a barrier layer may also be exploited for the formation of low-resistance ohmic contacts to shallow silicon diffused regions. There is no electromigration in the barrier layer since the activation energy for diffusion of titanium or tungsten is quite large. Thus, the composite film will likely never open. If the aluminum portion is completely voided in one region, current will flow in the underlying Ti–W. In this case, the dominant failure mode is short circuits due to electromigration-induced extrusions bridging to nearby metallization runs [61]. The Ti–W barrier does nothing to prevent the short-circuit failure since there will still be excessive motion of the aluminum ions. In fact, the presence of the barrier layer appears to enhance the extrusion growth [61]. The same complaint may be brought to bear on a concept in which an aluminum intermetallic compound is sandwiched between two aluminum-based alloys such as aluminum–copper [63]. In this case, the middle (aluminum intermetallic) layer acts as a barrier to void motion and thus precludes complete opening of the metal structure. but there is no reduction in hillock and extrusion formation. There may also be a small gain in prudently choosing the grain size and grain orientation of aluminum or one of its alloys. Electromigration MTTF increases as grain size increases [59]. A recent study argues that the (111) orientation of the aluminum–copper alloy is superior for electromigration resistance [71,78,79].

Going further, one may reject aluminum altogether. Fabricating inte-

grated circuit metallization from molybdenum, tungsten, titanium, or tantalum will virtually eliminate metal electromigration. Disadvantages lie in the increased resistivity and manufacturability. Nevertheless, the future trend appears to favor this approach, most notably with molybdenum, since multilevel metal schemes require a refractory metal in any event for the first metal level [80]. Multilevel metal greatly facilitates circuit design and, having solved refractory metal manufacturability concerns, one may designate the first metal layer as that which carries the highest current densities.

In the previous discussion of the activation energy of Eq. (5.1), we noted that the existence of a passivation layer over the metal inhibits surface atom migration and thus retards electromigration. Several studies substantiate this claim [60,67]. Beyond the suppression of surface atom migration, the passivation layer also separates the conductive line from overlying and neighboring structures. Short-circuit failure can occur only if the metal extrusions can succeed in breaking free of the confinement of the passivation layer. An inventive and laudable study by Severn *et al.* detected the acoustic emission of metal extrusions breaking through (cracking) a glass passivation layer [81]. An additional and intriguing result of this acoustic emission measurement is the signal emitted coincident with void growth in the metal. Thus, the acoustic emission monitors the open-circuit failure mode as well as the short-circuit mode. A plausible explanation suggested by Severn *et al.* for the void growth detection is the formation of microcracks in the glass due to the disappearing mechanical support in the voided metal region.

Simple test structures are advantageous for studying engineering aspects of electromigration, such as dependence on current density, temperature, and conductor material. Defining metal patterns directly on a flat, oxidized silicon wafer minimizes both the expense of the experiment and complicating factors that might frustrate clear interpretation of results. As remarked previously, though, the actual MTTF of a particular circuit design will depend cirtically on metal step coverage and pattern geometry. The conductor thickness over a topographical step will be less than that over a flat region. Since the same current flows in both places, the current density is greatest where the conductor cross section (i.e., thickness) is smallest. Thus, one expects these weak points to fail first. Similarly, 90° metallization corners and special structures [82] will produce regions of higher current density, which thus complicates the electromigration picture.

In addition to this straightforward idea that local regions of high current density are weakest in terms of electromigration is the concept of stress gradient [59]. A positive (negative) stress gradient exists when the current density increases (decreases) over a short distance along the conductor.

Independent of the magnitude of the current density, the stress gradient is deleterious. With zero stress gradient, the current density is uniform and the electromigration of metal ions will also tend to be uniform. As we noted, at some point the voids or hillocks will coalesce to open or short the conductor run. Regions of stress gradient provide natural coalescence points. For example, there will be a net loss of metal ions (i.e., void accumulation) in a region of positive gradient. Thus, a degraded conductor cross section is undesirable because of increased current density and the appearance of stress gradient regions. In fact, there is an ironic lesson here. A well-meaning circuit designer might wish to widen conductive runs in regions where available die space permits this luxury while maintaining the minimum design rule in high density regions in order to reduce current density in a portion of the integrated circuit metallization. This sounds like a good idea. But, unfortunately, such a design would produce stress gradients where the differing conductor widths meet and would likely exacerbate, not improve, electromigration.

Contending with electromigration poses conflicts with process manufacturability. From the electromigration viewpoint, one would prefer as thick a conductor line as possible in order to minimize the current density. But thicker films are correspondingly more difficult to etch. More importantly, thicker films produce poorer (more pronounced) wafer topography for subsequent lithography and depositions. Also, the sidewall oxide spacer described in the preceding section on channel hot electron reliability is advantageous for electromigration in that the metal layer will exhibit improved step coverage over this spacer. But, again, the sidewall spacer process adds complexity. Abandoning aluminum for a refractory metal is a major undertaking and requires a great deal of process development. We have found, though, that such a refractory metal process is viable. With refractory metal capability, electromigration reliability will not impede CMOS miniaturization in the foreseeable future.

Another reliability issue is contact electromigration [83–86]. To interconnect *n*-channel and *p*-channel MOSFETs, one requires ohmic contacts between the metallization and ($n+$ and $p+$) silicon. If we define a positive contact as a metal–semiconductor contact in which electrons flow from the semiconductor to the metal, then it is clear that current flow results in a metal ion flux divergence. That is, the electron wind will elicit a metal ion flux away from the interface into the metal. While metal ions leave the interface, there are no compensating metal ions entering this region because of the metal ion concentration discontinuity of the metal–semiconductor interface. Thus, one effect of electromigration at the contact region will be to deplete metal ions from positive contacts and the conductive line will eventually open. Chern *et al.* have determined the

activation energy (0.5 eV) for this process with aluminum–silicon metallization and conclude that grain boundary aluminum diffusion governs this open circuit failure mechanism [85]. Surprisingly, these authors also find that the positive contact must be located near a large (and more positively biased) bonding pad in order to suffer this electromigration failure. They claim that the bonding pad supplies vacancies, but this explanation is not clear to us. This failure mode is indistinguishable from conventional electromigration open failure with the exception that it is instigated by the interface discontinuity of the metal ion concentration. Thus, substitution of a refractory metal for aluminum or the addition of appropriate impurities to aluminum will ameliorate this contact electromigration problem.

Yet another contact electromigration failure mode may effectively short the silicon $p-n$ junction to which the contact is made [84]. In this case, silicon in the positive contact migrates into the metal contact under the influence of the electron wind. We suspect that the silicon leaving the semiconductor substrate resides initially at the contact interface, and is thus weakly bound, since the large activation energy of bulk silicon ions should preclude low-temperature electromigration. The departure of interface silicon ions leaves another layer of weakly bound ions that are susceptible to electromigration. The end result is a depletion of silicon in the positive contact in the form of etch pits. The shape of the etch pits is determined by crystallographic planes and thus supports the idea that the most weakly bound ions are preferentially removed. Metal ions tend to fill in the etch pits. This "back-fill" is crucial to the continued growth of the etch pit since the absence of metal in the etch pit would cease current flow and thus impede continued etch pit growth. With aluminum metallization, failure occurs in a contact to $n+$ silicon when the reverse ($n+-p$) diode leakage increases to an unacceptable level. Such a leakage increase is not observed in contacts to $p+$ silicon, even though the etch pits form and grow, since aluminum forms a Schottky diode to the underlying n-type silicon [84]. For this failure mechanism, insertion of a thin barrier metal (such as tantalum silicide [87,88]) between the aluminum and silicon will greatly reduce leakage failure. Again, one may choose an alternate metallization that acts as an inherent barrier to silicon migration.

III. OXIDE WEAR-OUT

Perhaps the most insidious failure mechanism of MOS devices is that of gate oxide wear-out. The thermal silicon dioxide insulator, the dominant gate dielectric of silicon semiconductor technology, may instantaneously rupture and form a conductive path between gate and substrate. This

conductive path is permanent and precludes proper circuit operation. Unlike electromigration, one cannot monitor, visually or otherwise, a continuous degradation in the integrity of the film. Rather, the insulator breakdown is quite abrupt. This behavior contributes to the difficulty in understanding oxide breakdown.

All experiments clearly show that the probability of oxide breakdown increases greatly with increased electric field. Figure 5.10 sketches the MOS structure for an *n*-channel device. To a first approximation, the electric field is directed from gate to substrate (positive gate bias) with magnitude given by the potential difference between gate and substrate surface divided by the oxide thickness. With the single observation that oxide wear-out (i.e., eventual and catastrophic oxide breakdown) increases with increasing field strength, we may discuss the role of wear-out in CMOS miniaturization. Ideal scaling [32,33], as discussed previously, dictates the reduction of all linear device dimensions (channel length, gate oxide thickness, spacer length, source–drain junction depth, and extrinsic debye length) and all applied voltages (drain, gate, and substrate) by the same multiplicative factor. In this ideal, constant field, scaling the oxide electric field would remain unchanged and the subject of oxide wear-out would be uncoupled from miniaturization. But, as we have noted several times, various additional constraints and considerations prevent the full application of ideal scaling. Practical realizations of device reduction always tend to increase the oxide field.

Results of oxide breakdown experiments segregate roughly into three categories [89,90]. Upon first application of nominal gate bias some MOS structures fail immediately. These "time zero" failures most likely arise from gross defects within the oxide film. Another group of oxide samples

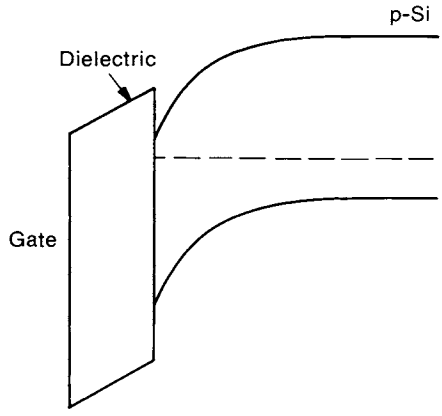


Fig. 5.10. Metal–oxide–semiconductor structure.

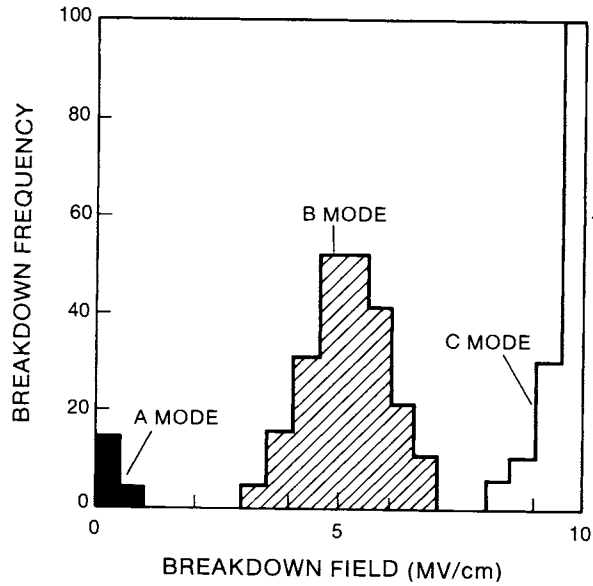


Fig. 5.11. Oxide breakdown data of Yamabe and Taniguchi [89]. © 1985 IEEE.

will fail only at fairly high field strength (greater than 7 MV/cm) and may function well for many hours prior to abrupt breakdown. (For comparison, present 256-K DRAMs may employ a 25-nm gate oxide with a 5-V power supply so that the maximum oxide field is 2 MV/cm.) This high-field failure mechanism is known as "intrinsic" failure and will be discussed later. The third group of oxide samples is intermediate to the time zero and intrinsic populations. These devices may abruptly rupture with applied fields in the approximate range of 2–6 MV/cm after many hours of operation. Subtle defects within the gate oxide are apparently responsible for this behavior since the intrinsic population assures us that defect-free silicon dioxide should be able to withstand these intermediate electric fields. Figure 5.11 reprints experimental data of Yamabe and Taniguchi [89] to illustrate the three distinct populations of oxide dielectrics. Oxide insulators within this wear-out group greatly antagonize the CMOS process engineer since submicron technology is driving the industry into this electric field domain. After describing the present understanding of oxide wear-out, we discuss possible processing improvements and experimental screening procedures.

The precise mechanism of oxide breakdown, either for the intrinsic or wear-out populations, is unknown since there is very little that one may observe directly. Let us first recount the generally accepted experimental observations. If one applies equivalent electric field and temperature to a

large group of similar MOS capacitors, one measures a distribution of failure times. That is, all devices do not fail at the same instant even though all controllable parameters are identical. Thus, uncontrolled and experimentally unquantifiable parameters, such as structural defects in the oxide, chemical contamination, and (high-frequency) oxide thickness variations, determine the failure time of any particular MOS capacitor. For characterization, one measures the time interval in which some initial fraction (e.g., 10% or 50%) of the capacitors rupture. Several studies have shown that the failure rate is log normal in time [91–95].

In terms of controllable parameters, it is clear that increasing either the electric field or the ambient temperature will reduce the mean time to failure of the gate oxide insulator. Reported values of the field acceleration and temperature acceleration factors have varied widely. For example, decreases in the mean time to oxide failure from two decades to seven decades per MV/cm of oxide electric field have appeared in the literature [91,96–98]. Several studies have claimed effective activation energies for the oxide breakdown temperature dependence in the range of 0.3–2.0 eV [91,96–98]. McPherson and Baglee have resolved these apparent discrepancies through experimental and theoretical arguments [99,100]. Essentially, these acceleration factors are both field and temperature dependent.

The breakdown process is highly localized in that a conductive short appears at a specific location within the MOS capacitor. An interesting and useful aspect of oxide rupture is the complete vaporization of *thin* gate materials over the conductive short. With these thin gates, gate material vaporization acts as a self-healing mechanism since, with no gate material over the rupture location, the capacitor will not be shorted. In fact, several studies have employed this property of thin gate electrodes to observe many self-healing breakdown events on one capacitor [90,101].

Deliberate introduction of ionic contamination such as sodium reduces the oxide mean time to failure [102,103] as do substrate contamination and defects [89]. Intrinsic breakdown is always present, however, and an intermediate-field, wear-out population may persist in a fabrication facility in the apparent absence of ionic contamination and substrate imperfections. The breakdown distribution improves (shifts to higher electric field) as oxide thickness decreases [89,104]. This behavior is certainly favorable for CMOS device miniaturization.

One would expect intrinsic breakdown to yield more easily to theoretical understanding than intermediate field wear-out failure since the former process involves ideal, defect-free silicon dioxide films. Prior to intrinsic breakdown, one observes current flow through the insulator between the gate and substrate. Figure 5.12 sketches the energy band picture of this situation with the gate biased positively. Fowler–Nordheim tunneling

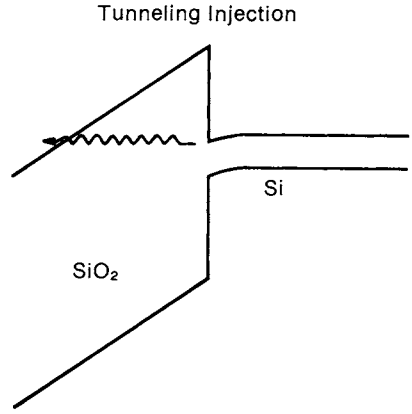


Fig. 5.12. Energy band diagram sketch of the MOS system depicting electron tunneling. (Reprinted from Nissan-Cohen *et al.* [110].) © 1983 AIP.

[105–107], in which electrons in the substrate (anode) tunnel through the triangular energy barrier into the oxide, followed by electron drift to the gate (cathode), dominates the oxide conduction process at typical field and temperature values. The tunneling current has the form

$$I = AE^2 \exp(-B/E), \tag{5.2}$$

where E is the electric field strength at the cathode. The parameters A and B of Eq. (5.2) depend primarily on the maximum barrier height (Fig. 5.12). There is weak temperature dependence in A and field dependence in B [107], however, even though these quantities are often termed constants. The oxide current is clearly exponential in the electric field and experimental studies have found that the oxide current is well described by this equation.

As one increases the positive bias on a MOS capacitor, one finds increasing oxide current up to the point of oxide rupture. Most theories of breakdown assign an important role to this oxide current. The localized nature of breakdown dictates a positive feedback mechanism in which preliminary degradation increases the local electron injection. Enhanced local injection causes more degradation, leads to greater injection, and thus results in local current runaway. One mechanism that does not exhibit this positive feedback requirement is electron trapping. Localized electron trapping, due to random defects, for example, will decrease the local Fowler–Nordheim electron injection since the negative space charge of the trapped electrons will reduce the local electric field at the cathode. The prevailing theory [108,109] asserts that electron tunneling into the oxide in a high electric field leads to hole generation by impact ionization (within

the oxide) and net hole trapping. Thus, paradoxically, electron injection leads to positive oxide charge. The apparent implausibility of this argument decreases when one realizes that hole trapping within silicon dioxide is far more likely than electron trapping. This differential trapping behavior explains negative threshold shifts (positive trapped charge) in MOS devices exposed to ionizing radiation [34,35]. Furthermore, Klein and Solomon [108] proposed a model in which hole generation and trapping are opposed by electron recombination with trapped holes. This impact-recombination (IR) model apparently simulates well the cathode field enhancement due to the accumulation of positive oxide charge. Nissan-Cohen *et al.* [110] subsequently verified this model prediction by demonstrating experimentally that the first-order kinetic equation description of hole trapping and detrapping (recombination) is accurate. We should add that, as the authors note, one may not infer from these experiments that impact ionization, while perhaps most likely, is uniquely identified as the mechanism for hole generation. Rather, Nissan-Cohen *et al.* showed that the hole generation requires both high electric field and high electron current density within the oxide.

Returning to the intrinsic breakdown mechanism, the impact ionization model proposed by DiStefano and Shatzkes [109] and Klein and Solomon [108] specifies net positive charge accumulation due to high-field impact ionization followed by hole trapping. Positive charge within the oxide increases the cathode field so that electron injection [by Fowler–Nordheim tunneling of Eq. (5.2)] increases. The rate of positive charge accumulation will then also increase, and this positive feedback process produces a local and essentially divergent current density that destroys the oxide structure and, as noted previously, can even vaporize the gate material. Several research studies employ "charge to breakdown" [111–113] as a measure of dielectric strength. The implication of this term is that a given MOS capacitor will fail after a critical amount of electron injection into the oxide. Thus, charge to breakdown is the time integral of the oxide current density from the initiation of bias to breakdown. We claim that such a characterization is erroneous when one does not maintain a constant electric field since the impact ionization process is an exceedingly sensitive function of the electric field. For example, one often hears claims of superior breakdown performance of nitrided oxide films over pure oxide films [114]. But the nitrided oxides tend to have reduced injection barriers because of the accumulation of nitrogen at the silicon–dielectric interface. Thus, one applies a smaller electric field to elicit equivalent current density. So the improved charge to breakdown simply results from the lower electric field.

Some experimental evidence is difficult to reconcile with this impact

ionization model of intrinsic breakdown. For example, as we noted, the oxide breakdown process exhibits temperature dependence with an effective, electric-field-dependent activation energy [99,100]. The mean time to failure decreases with increasing ambient temperature. But, Fowler–Nordheim tunneling tends to be insensitive to temperature and, if anything, impact ionization decreases as temperature increases. So one must explain the temperature dependence of oxide breakdown. Another inconsistency of the impact ionization model relates to the oxide thickness dependence of breakdown. One would expect, and indeed finds [104], that breakdown strength increases as oxide thickness decreases since electrons drifting in a thin oxide have less probability of gaining enough energy (about 9 eV) to ionize a valence electron. But, given the impact ionization model, one should expect virtually no impact ionization when the applied bias is less than about 12 V (the 9-V oxide band gap plus the 3-V barrier not "seen" by the tunneling electrons). Yet 10-nm oxides with 10-V bias will rupture [104].

An alternate model that qualitatively addresses the two issues of temperature and oxide thickness dependence specifies resonant, as opposed to Fowler–Nordheim, tunneling for electron injection into the gate oxide [115]. Resonant tunneling requires electron traps within the oxide band gap that act as intermediaries for electron tunneling from the cathode to oxide conduction band. An oxide defect state with a given energy level provides maximum tunneling at an optimum cathode field. This tunneling is enhanced by increased temperature since more cathode electrons will acquire the appropriate energy. One no longer requires the 9 V for the impact ionization. The positive feedback in this model is predicated on the claim that electron trapping within the defect state will raise the energy of this state (as it reduces the cathode field) and that it is possible that raising this energy level will bring it closer to the resonance condition. Though speculative, this model is attractive since it addresses the issue of thin oxide breakdown as well as the temperature dependence. We are not aware of convincing experimental verification for the importance of resonant tunneling in the breakdown process. As we noted earlier, measurements of oxide current appear to be well described by the Fowler–Nordheim tunneling process so that one would require a "unified" model in which both injection processes coexist.

The bulk of this discussion has centered on intrinsic breakdown. While intrinsic (high-field) failures are easier to study than defect-related (intermediate-field) failures, the latter group is of far greater importance since CMOS miniaturization now specifies oxide electric fields entering this intermediate range. One well-recognized cause of intermediate field breakdown is mobile ion (generally sodium) contamination [102,103]. With an

applied field in the oxide dielectric, the positive ionic contamination eventually drifts toward the cathode. The proximity of this ionic charge to the cathode increases the cathode electric field and hence increases the electron injection. This field enhancement continues until the oxide ruptures. In this case, the ambient temperature appears as an important factor in determining ionic mobility within the oxide so that the increase in breakdown rate with temperature is easily understood.

However, one often observes intermediate field failures in the absence of detectable ionic contamination. Because of the sporadic nature of these failures and the clear evidence that "good" oxides tend to break down at fields exceeding 7 MV/cm, one generally ascribes intermediate field failures to oxide defects and fabrication process contaminants. Shatzkes *et al.* [116–118] proposed that oxide defects may effectively lower the barrier for Fowler–Nordheim tunnel injection of electrons into the oxide conduction band. The breakdown mechanism at intermediate fields, then, is identical to that (intrinsic) at high fields. The breakdown distribution is simply shifted to lower fields because of the defect barrier lowering. Figure 5.13 reprints experimental data from Shatzkes *et al.* [116] in which oxide current is measured as a function of applied field. This data decomposes easily into the sum of two Fowler–Nordheim characteristics [Eq. (5.2)] such that the low-field current is dominated by a tunnel injection component with barrier height, in this case, of about 2.4 eV compared to the normal oxide barrier height of 3.1 eV. Thus, this particular capacitor exhibits higher leakage than expected of a defect-free oxide film. Shatzkes *et al.* found that capacitors with this excessive leakage also suffered the intermediate field breakdown with far greater probability than the "ideal" current capacitors. Thus, the present model of intermediate field breakdown holds that oxide defects and contaminants reduce the barrier for cathode electron injection into the oxide so that the intrinsic breakdown mechanism proceeds at lower field strengths. Experiments have shown that process-induced defects, such as substrate imperfections, do lead to increased intermediate field breakdown [89]. As oxide thickness decreases, the defect density apparently decreases since breakdown field increases for this intermediate breakdown mode [104]. As we noted with the similar oxide thickness dependence for the intrinsic case, thinner oxide should lead to reduced impact ionization due simply to the limited spatial extent of the oxide conduction band in which the electrons possess enough energy to ionize a valence electron.

Determination of oxide breakdown voltage requires the application of increasing bias to a MOS capacitor. The bias increases in either a continuous ramp or in a stepwise manner until the oxide ruptures. A voltage step stress technique generated the data of Fig. 5.11 [89]. While oxide break-

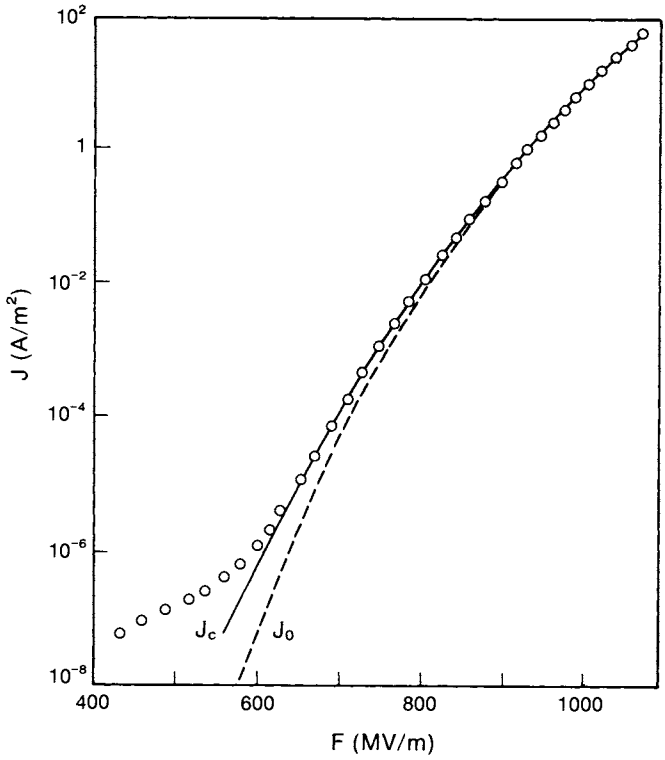


Fig. 5.13. Oxide $I-V$ data. (Reprinted from Shatzkes *et al.* [116].)

down voltage is valuable for evaluation of oxide quality and comparison with theories of breakdown, this breakdown voltage is not directly relevant for long-term device reliability. Clearly, the device engineer specifies a power supply less than the observed breakdown voltage. In fact, the difficult aspect of oxide breakdown relative to long-term reliability is that gate dielectrics may withstand a particular bias for hundreds and thousands of hours prior to irreversible rupture.

Thus, "real world" application of the MOS structure in integrated circuits dictates investigation of the time to failure of the oxide dielectric under given conditions of gate bias and temperature where the gate bias is invariably less than the easily measured breakdown voltage. Such time-dependent dielectric breakdown (TDDB) measurements are expensive in the sense that they may require thousands of hours. This experiment is simplified by the empirical observation that the failure rate (fraction of MOS capacitors rupturing per unit time) approximates a log normal distribution [91–95]. With $f(t)$ as the time-dependent failure rate and μ and σ the mean

and dispersion of the breakdown histogram, respectively, we write

$$f(t) = \frac{1}{\sigma t\sqrt{2\pi}} \exp\left[\frac{-1}{2}\left(\frac{\log(t/\mu)}{\sigma}\right)^2\right] \qquad 0 < t < \infty, \tag{5.3}$$

where "log" denotes the natural logarithm. Experimental TDDB results determine the mean (μ) and dispersion (σ) of Eq. (5.3). Noting that the integral $F(t)$ of $f(t)$ yields the cumulative fraction of failures, that is,

$$F(t) = \int_0^t d\tau f(\tau),$$

one may define [91] the instantaneous failure rate $h(t)$ from Eq. (5.3) as

$$h(t) = \frac{f(t)}{1 - F(t)}. \tag{5.4}$$

Though perhaps not obvious, the instantaneous failure rate $h(t)$ of Eq. (5.4) decreases monotonically in time. That is, the device failure rate, measured as the fraction of the instantaneous population rupturing per unit time, decreases as the stress time increases. Figure 5.14 reprints typical experimental data expressing instantaneous failure rate versus time [91].

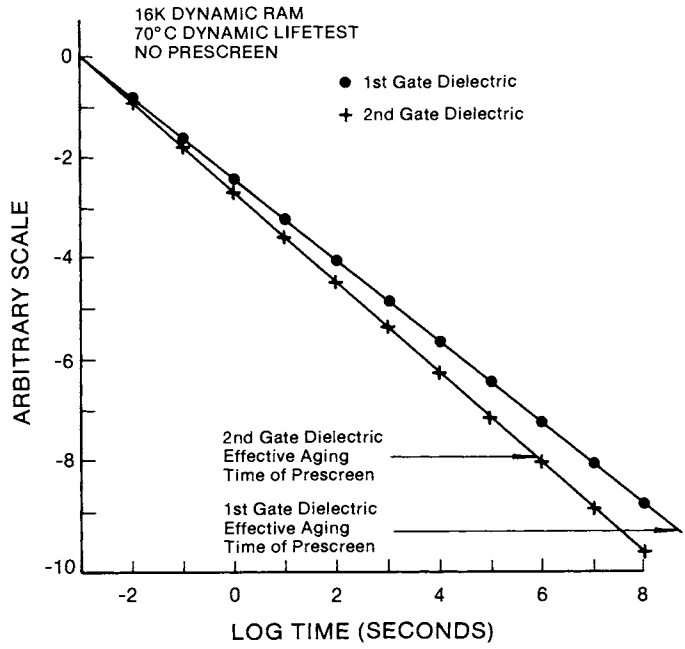


Fig. 5.14. Instantaneous failure rate data. (Reprinted from Crook [91].) © 1979 IEEE.

This observation of predictable decreasing failure rate with time forms the logical basis for aging screens. If the reliability engineer specifies a maximum allowable rate of failure of constituent MOS capacitors of 100 failures in time (FIT), the population may be stressed, or aged, until the instantaneous failure rate falls below this level. (One FIT is equivalent to one device failure in one billion hours [119].) There is also an area dependence for MOS capacitor failure that one must gauge when comparing test devices to actual circuits. Semiempirical yield laws exist for this area conversion [91].

Of course, the problem of assuring acceptable reliability for semiconductor integrated circuits is not completely resolved for at least two reasons. First, the aging time required to reduce the instantaneous failure rate to a prescribed level may be unreasonably long. Second, the remaining population at the end of this aging may be uneconomically small. With regard to the first point, one may accelerate the aging process with the application of elevated temperature and elevated electric field. While the temperature dependence, as embodied in an effective Arrhenius activation energy, and the electric field dependence are not trivial, McPherson and Baglee have proposed a universal, and surprisingly accurate, model for the effect of these controllable parameters on TDDB [99,100]. McPherson and Baglee demonstrated how widely varying activation energies (0.3 eV to 2.0 eV) and electric field acceleration factors (2 decades per MV/cm to 7 decades per MV/cm) [91,96–98,120] could be systematically and logically correlated to stress field, stress temperature, and definition of lifetime. Thus, the McPherson–Baglee model provides an excellent guide for quantitative estimation of acceleration factors for aging. Figure 5.15 presents an excellent example of this aging, or "burn in," method [91]. This figure shows instantaneous failure rate versus time for two identical groups of capacitors. In one group (filled circles), a continuous stress of 2 MV/cm is applied, while in the other group (open circles), one imparts a stress of 2.5 MV/cm for the first second and then drops the field to 2 MV/cm. The temperature is unchanged at 25°C. As one expects, the failure rate is higher within the first second of stress for the high-field (2.5 MV/cm) group. Upon dropping the field on this high-field group, however, the failure rate falls well below that of the continuous low-field (2 MV/cm) group. The failure rate drops to a value attained by the continuous low-field group at a much later time. Application of high field, then, is much more efficient for reduction of instantaneous failure rate than straightforward stress under normal operating conditions.

The second aspect of low yield following the required aging screen is more fundamental. With this happenstance, it is likely that the system and circuit demands are too great for the particular oxide technology. If the

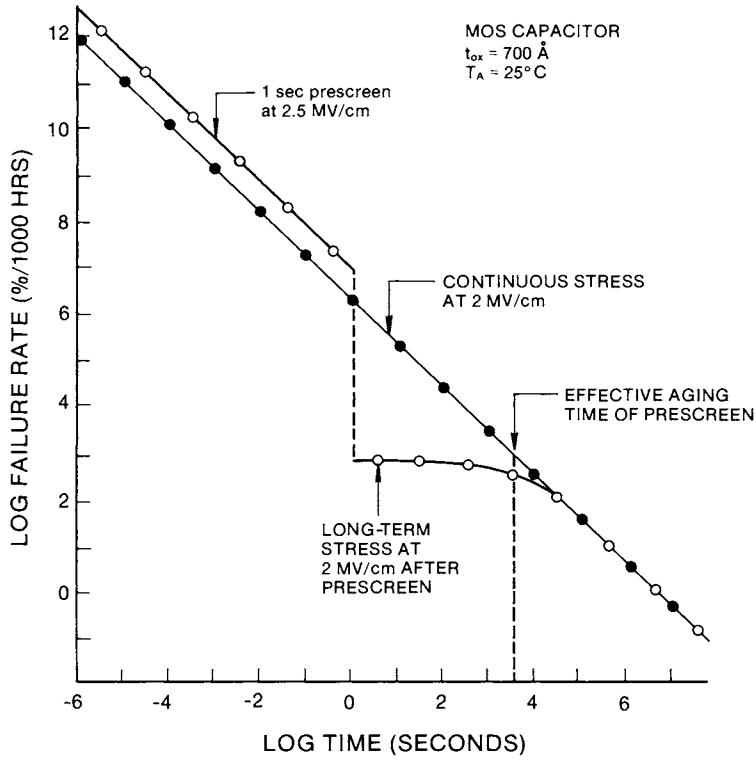


Fig. 5.15. Instantaneous failure rate data showing the effect of aging stress at elevated field. (Reprinted from Crook [91].) © 1979 IEEE.

bulk of the ruptures occur in the defect (intermediate-field) range, one must attempt to improve oxide quality by careful cleaning techniques [121–123], contamination control and special growth methods [124–126]. The remaining unpalatable alternative is electric field reduction with accompanying performance loss.

As operating fields approach the range of 4–5 MV/cm, a new problem with the aging screen arises. With these fields, defect-induced breakdown must be virtually eliminated by careful processing techniques. Beyond this requirement, one has difficulty in specifying an appropriate aging screen since application of stress fields in excess of about 6 MV/cm will begin to damage even the defect-free oxide films. That is, the act of eliminating oxide films with latent defects also undermines the integrity of initially high-quality films. Meyer and Crook have devised an alternative, nonaging screen for dielectric integrity that does not degrade the device [127]. This screen is based on the proposition of Shatzkes *et al.* that breakdown-prone,

defective oxides may be identified by excessive, yet still low, oxide leakage currents [116]. In fact, these authors noted that "leaky" capacitors still exhibit Fowler–Nordheim behavior so that the mechanism of breakdown in defective and intrinsic oxides is similar with the exception of a lower injection barrier in defective films. So potentially defective devices are discarded based on sensitive oxide conduction measurements below the conventional Fowler–Nordheim field level of about 6 MV/cm. This current measurement does not degrade the device under test. The disadvantage of this method lies in the general difficulty of measuring small currents (10 pA/cm^2) in a convenient manner. Meyer and Crook applied the low-current criterion with great success to a DRAM device in which indirect low-current measurements are feasible [127].

We noted previously that measurement of breakdown voltage by ramp or voltage step techniques is not directly relevant to practical concerns since one needs to know how long a given capacitor will survive at lower field strengths. These breakdown measurements are attractive since they require very little time. Since it is intuitively clear that breakdown voltage should be positively correlated with mean time to failure at low bias, several groups have formulated methods for extracting mean lifetime from breakdown measurements. Shatzkes *et al.* described a graphical technique in which one ramps the gate bias and measures the cumulative failure rate as a function of instantaneous bias and ramp rate [128]. The result is a plot of lifetime versus bias which one may informally extrapolate to lower biases typical of circuit operation. Based on an empirical model of the impact ionization and hole trapping process in oxide, Chen *et al.* constructed an expression for mean breakdown time as a function of breakdown field and ramp rate, (low) stress field, and oxide thickness [112,113]. To our knowledge, the complete substitution of inexpensive breakdown voltage measurements for time-consuming "burn-in" of many hours at elevated electric field and temperature is not common because of the lack of confidence in the interpretation of the breakdown technique.

There are complicating factors in the specification of alternative reliability screens. First, it is critically important that test capacitors intended to replicate the gate dielectric of a particular product be fabricated in an identical manner to that of the product. Various fabrication steps may degrade or enhance oxide quality [125,126]. Second, structural features, such as the transition from gate oxide (active area) to field oxide, and the proximity of a self-aligned, heavily doped region, such as the source/drain region of a MOSFET, may have a profound impact on dielectric integrity. Such attributes should be preserved in the test device. These warnings favor performing long-term reliability studies directly on the end product. Beyond losing the advantage of direct current measurement for a nonaging

screen, applying an elevated bias to the end product may be impossible. For example, the maximum bias may be limited by $p-n$ junction avalanche breakdown. To resolve this dilemma, Domangue *et al.* proposed an outstanding method with which to gauge their DRAM product [93]. Within each process lot, these authors fabricated some dies with an alternate metallization mask so that bias application and failure detection were facilitated. Thus, they were able to monitor a completely representative structure while retaining the ability to characterize individual defective oxides.

Finally, we remark that fabrication process differences between good, defect-free oxides and poor oxides are somewhat vague. As always, "clean" processing with minimal contamination and handling, maximum gettering, and high-quality starting material is mandatory. But, beyond the early identification of mobile ionic contamination as an oxide lifetime detractor [102,103], very little specific and reproducible results exist. Several groups have claimed improved processes for dielectric integrity [124–126]. Nitrided oxides, thermal oxide films with nitrogen introduced by high-temperature annealing or ion implantation, have appeared as possible substitutes for conventional silicon dioxide based partially on the claim of improved breakdown strength [114,129–134]. These claims are somewhat misleading, though. Typical evidence for improvement is the greater integral of injected current density at breakdown (i.e., "breakdown charge") of nitrided oxides [114]. But the proponents tend to omit the observation that the breakdown field is not improved. The nitrided oxides simply pass more current than do thermal oxides because of the lowered injection barrier. Thus, practical superiority of nitrided oxides is, in this respect, lacking. Other advantages, such as inhibition of impurity diffusion and oxidation resistance, remain.

REFERENCES

1. T. H. Ning, P. W. Cook, R. H. Dennard, C. M. Osburn, S. E. Schuster, and H.-N. Yu, 1 μm MOSFET VLSI technology: part IV - hot electron design constraints, *IEEE Trans. Electron Devices* **ED-26**, 346 (1979).
2. C. Hu, Hot electron effects in MOSFETs, *IEDM Tech. Dig.*, 176 (1983).
3. T. H. Ning, C. M. Osburn, and H.-N Yu, Emission probability of hot electrons from silicon into silicon dioxide, *J. Appl. Phys.* **48**, 286 (1977).
4. E. Takeda, H. Kume, Y. Nakagome, and S. Asai, An as-p (n+–n–) double diffused drain MOSFET for VLSIs, *Symp. VLSI Tech. Dig.*, 40 (1982).
5. Y. Nakagome, E. Takeda, H. Kume, and S. Asai, New observation of hot-carrier injection phenomena, *14th Conf. Solid State Dev. Dig.*, 63 (1982).
6. M. Noyori, Y. Nakata, S. Odanaka, and J. Yasui, Reduction of V_T shift due to

avalanche hot carrier injection using graded drain structures for sub-micron n-channel MOSFET, *Int. Rel. Phys. Symp. Proc.*, 205 (1984).

7. R. Resnick and D. Halliday, "Physics for Students of Science and Engineering." Wiley, New York, 1960.
8. N. W. Ashcroft and N. D. Mermin, "Solid State Physics." Holt, New York, 1976.
9. J. L. Moll, "Physics of Semiconductors." McGraw-Hill, New York, 1964.
10. P. Vogl *in* "Physics of Nonlinear Transport in Semiconductors," D. K. Ferry, J. R. Barker, and C. Jacoboni, eds. Plenum, New York, 1980.
11. C. Jacoboni, C. Canali, G. Ottaviani, and A. A. Quaranta, A review of some charge transport properties of silicon, *Solid State Elec.* **20,** 77 (1977).
12. R. Williams, Photoemission of electrons from silicon into silicon dioxide, *Phys. Rev.* **140,** A569 (1965); A. M. Goodman, Photoemission of holes from silicon into silicon dioxide, *Phys. Rev.* **152,** 780 (1966).
13. B. Eitan and D. Frohman-Bentchkowsky, Hot electron injection into the oxide in n-channel MOS devices, *IEEE Trans. Electron Devices* **ED-28,** 328 (1981).
14. G. D. Mahan, Hot electrons in one dimension, *J. Appl. Phys.* **58,** 2242 (1985).
15. E. H. Nicollian, C. N. Berglund, P. F. Schmidt, and J. M. Andrews, Electrochemical charging of thermal SiO_2 films by injected electron currents, *J. Appl. Phys.* **42,** 5654 (1971).
16. T. H. Ning, C. M. Osburn, and H.-N. Yu, Threshold instability in IGFETs due to emission of leakage electrons from silicon substrate into silicon dioxide, *Appl. Phys. Lett.* **29,** 198 (1976). Also, T. H. Ning and H. N. Yu, Optically induced injection of hot electrons into SiO_2, *J. Appl. Phys.* **45,** 5373 (1974).
17. T. H. Ning, C. M. Osburn, and H.-N. Yu, Effect of Electron Trapping on IGFET characteristics, *J. Elec. Mater.* **6,** 65 (1977).
18. D. Schmitt and G. Dorda, Interface states in MOSFETs due to hot electron injection determined by the charge pumping technique, *Elec. Lett.* **17,** 761 (1981).
19. E. Takeda, A. Shimizu, and T. Hagiwara, Role of hot hole injection in hot carrier effects and the small degraded channel region in MOSFETs, *IEEE Elec. Dev. Lett.* **EDL-4,** 329 (1983).
20. N. Shiono and C. Hashimoto, Hot electron limited operating voltages for 0.8 μm MOSFETs, *IEEE Trans. Electron Devices* **ED-29,** 1630 (1982).
21. J. M. Pimbley and G. Gildenblat, Effect of hot electron stress on low frequency MOSFET noise, *IEEE Elec. Dev. Lett.* **EDL-5,** 345 (1984).
22. H. Gesch, J. P. Leburton, and G. Dorda, Generation of interface states by hot hole injection in MOSFETs, *IEEE Trans. Electron Devices* **ED-29,** 913 (1982).
    P. Heremans, H. E. Maes, and N. S. Saks, Evaluation of hot carrier degradation of n-channel MOSFETs with the charge pumping technique, *IEEE Elec. Dev. Lett.* **EDL-7,** 428 (1986).
23. Y. Nissan-Cohen, J. Shappir, and D. Frohman-Bentchkowsky, Characterization of simultaneous bulk and interface high-field trapping effects in SiO_2, *IEDM Tech. Dig.*, 182 (1983).
    A. Badihi, B. Eitan, Y. Nissan-Cohen, and J. Shappir, Current induced trap generation in SiO_2, *Appl. Phys. Lett.* **40,** 396 (1982).
    C. S. Jenq, T. R. Ranganath, C. H. Huang, H. Stanley Jones, and T. T. L. Chang, High-field generation of electron traps and charge trapping in ultra-thin SiO_2, *IEDM Tech. Dig.*, 388 (1981).
24. K.-L. Chen, S. A. Saller, I. A. Groves, and D. B. Scott, Reliability effects on MOS transistors due to hot carrier injection, *IEEE Trans. Electron Devices* **ED-32,** 386 (1985).

25. C. Duvvury, D. Redwine, H. Kitagawa, R. Haas, Y. Chuang, C. Beydler, and A. Hyslop, Impact of hot carriers on DRAM circuits, *Proc. 25th Int. Rel. Phys. Symp.*, 201, San Diego, California, April 1987.
26. E. Takeda and N. Suzuki, An empirical model for device degradation due to hot carrier injection, *IEEE Elec. Dev. Lett.* **EDL-4,** 111 (1983).
27. P. E. Cottrell, R. R. Troutman, and T. H. Ning, Hot electron emission in n-channel IGFETs, *IEEE Trans. Elec. Dev.* **ED-26,** 520 (1979).
28. N. S. Saks, P. L. Heremans, L. Van Den Hove, H. E. Maes, R. F. De Keersmaecker, and G. J. Declerck, Observation of hot hole injection in NMOS transistors using a modified floating-gate technique, *IEEE Trans. Elec. Dev.* **ED-33,** 1529 (1986).
29. W. Weber, C. Werner, and G. Dorda, Degradation of NMOS transistors after pulsed stress, *IEEE Elec. Dev. Lett.* **EDL-5,** 518 (1984).
30. K. R. Hofmann, C. Werner, W. Weber, and G. Dorda, Hot electron and hot hole emission effects in short n-channel MOSFETs, *IEEE Trans. Electron Devices* **ED-32,** 691 (1985).
31. R. B. Fair and R. C. Sun, Threshold voltage instability in MOSFETs due to channel hot-hole emission, *IEEE Trans. Electron Devices* **ED-28,** 83 (1981).
32. R. H. Dennard, F. H. Gaensslen, H.-N. Yu, V. L. Rideout, E. Bassous, and A. LeBlanc, Design of ion-implanted MOSFETs with very small physical dimensions, *IEEE J. Solid State Cir.* **SC-9,** 256 (1974).
33. G. Baccarani, M. R. Wordeman, and R. H. Dennard, Generalized scaling theory and its application to a ¼ micrometer MOSFET design, *IEEE Trans. Electron Devices* **ED-31,** 452 (1984).
34. P. K. Chatterjee, W. R. Hunter, T. C. Holloway, and Y. T. Lin, The impact of scaling laws on the choice of n-channel or p-channel for MOS VLSI, *IEEE Elec. Dev. Lett.* **EDL-1,** 220 (1980).
35. K. G. Aubuchon, Radiation hardening of P-MOS devices by optimization of the thermal SiO_2 gate insulator, *IEEE Trans. Nucl. Sci.* **NS-18**(6), 117 (1971).
36. E. H. Snow, A. S. Grove, and D. J. Fitzgerald, *Proc. IEEE* **55,** 1168 (1967).
37. S. K. Lai, J. Lee, and V. K. Dham, Electrical properties of nitrided-oxide systems for use in gate dielectrics and EEPROM, *IEDM Tech. Dig.*, 190 (1983).
38. E. Takeda, H. Kume, T. Toyabe, and S. Asai, Sub-micrometer MOSFET structure for minimizing hot carrier generation, *IEEE Trans. Electron Devices* **ED-29,** 611 (1982).
39. E. Takeda, H. Kume, Y. Nakagome, T. Makino, A. Shimuzu, and S. Asai, An as-p (n+-n−) double diffused drain MOSFET for VLSIs, *IEEE Trans. Electron Devices* **ED-30,** 652 (1983).
40. K. Balasubramanyam, M. J. Hargrove, H. I. Hanafi, M. S. Lin, D. Hoyniak, J. LaRue, and D. R. Thomas, Characterization of as-p double diffused drain structure, *IEDM Tech. Dig.*, 782 (1984).
41. P. J. Tsang, S. Ogura, W. W. Walker, J. F. Shepard, and D. L. Critchlow, Fabrication of high performance LDDFETs with oxide sidewall spacer technology, *IEEE Trans. Electron Devices* **ED-29,** 590 (1982).
42. S. Ogura, P. J. Tsang, W. W. Walker, D. L. Critchlow, and J. F. Shepard, Design and characteristics of the lightly-doped drain-source (LDD) insulated gate field-effect transistor, *IEEE Trans. Electron Devices* **ED-27,** 1359 (1980).
43. M. R. Wordeman, A. M. Schweighart, R. H. Dennard, G. Sai-Halasz, and W. W. Molzen, A fully solid submicrometer NMOS technology using direct-write e-beam lithography, *IEEE Trans. Electron Devices* **ED-32**(11), 2214 (1985).
44. D. A. Baglee and C. Duvvury, Reduced hot electron effects in MOSFETs with an optimized LDD structure, *IEEE Elec. Dev. Lett.* **ED-5,** 389 (1984).

45. Y. Tsunashima, T. Wada, K. Yamada, T. Moriya, M. Nakamura, R. Dang, K. Taniguchi, M. Kashiwagi, and H. Tango, Metal-coated lightly-doped drain (MLD) MOSFETs for sub-micron VLSIs, Proc. 5th Symp. VLSI Tech., 114, Kobe, Japan, May 1985.
46. C.-Y. Wei, J. M. Pimbley, and Y. Nissan-Cohen, Buried and graded/buried LDD structures for improved hot electron reliability, *IEEE Elec. Dev. Lett.* **EDL-7,** 380 (1986).
47. H. Ishiuchi, Y. Matsumoto, S. Sawada, and O. Ozawa, Measurement of intrinsic capacitance of lightly doped drain (LDD) MOSFETs, *IEEE Trans. Electron Devices* **ED-32,** 2238 (1985).
48. S. L. Von Bruns and R. L. Anderson, Hot electron induced interface state generation in n-channel MOSFETs at 77K, *IEEE Trans. Electron Devices* **ED-34,** 75 (1987).
49. J. M. Pimbley, A measurement method for the increase of digital switching time due to hot electron stress, *IEEE Elec. Dev. Lett.* **EDL-6,** 366 (1985).
50. J. M. Pimbley and G. Gildenblat, Comments on "Structure enhanced MOSFET degradation due to hot electron injection," *IEEE Elec. Dev. Lett.* **EDL-5,** 256 (1984). See also reply of F.-C. Hsu, *IEEE Elec. Dev. Lett.* **EDL-5,** 258 (1984).
51. R. R. Troutman, Low level avalanche multiplication in IGFETs, *IEEE Trans. Electron Devices* **ED-23,** 419 (1976).
52. S. Tam and C. Hu, Hot electron induced photon and photo-carrier generation in silicon MOSFETs, *IEEE Trans. Electron Devices* **ED-31,** 1264 (1984).
53. A. G. Chynoweth and K. G. McKay, Photon emission from avalanche breakdown in silicon, *Phys. Rev.* **102,** 369 (1956).
54. K. R. Hofmann, W. Weber, C. Werner, and G. Dorda, Hot carrier degradation mechanism in n-MOSFETs, *IEDM Tech. Dig.,* 104 (1984).
55. S. Selberherr, "Analysis and Simulation of Semiconductor Devices." Springer-Verlag, Berlin and New York, 1985.
56. J. R. Pfiester, J. D. Shott, and J. D. Meindl, Performance limits of CMOS ULSI, *IEEE Trans. Electron Devices* **ED-32**(2), 333 (1985).
57. J. D. Meindl, Ultra-large scale integration, *IEEE Trans. Electron Devices* **ED-31,** 1555 (1984).
58. H. R. Huntington and A. R. Grone, Current-induced marker motion in gold wires, *J. Phys. Chem. Solids* **20,** 76 (1961).
59. J. R. Black, Electromigration—a brief survey and some recent results, *IEEE Trans. Electron Devices* **ED-16,** 338 (1969).
60. J. R. Black, Electromigration failure modes in aluminum metallization for semiconductor devices, *Proc. IEEE* **ED-57,** 1587 (1969).
61. J. M. Towner, Electromigration-induced short circuit failure, *Proc. 23rd Int. Rel. Phys. Symp.,* 81, Orlando, Florida, March 1985.
62. J. Lloyd and P. Smith, *J. Vac. Sci. Technol.* **1,** 455 (1983).
63. S. S. Iyer and C.-Y. Ting, Electromigration study of the Al–Cu/Ti/Al–Cu system, *Proc. 22nd Int. Rel. Phys. Symp.,* 273. Las Vegas, Nevada, April 1984.
64. F. M. D'Heurle, Electromigration and failure in electronics: an introduction, *Proc. IEEE* **59,** 1409 (1971).
65. P. B. Ghate, Electromigration-induced failures in VLSI interconnects, *Proc. 20th Int. Rel. Phys. Symp.,* 292, San Diego, California, March 1982.
66. H. A. Schafft, T. C. Grant, A. N. Saxena, and C.-Y. Kao, Electromigration and the current density dependence, *Proc. 23rd Int. Rel. Phys. Symp.,* 93, Orlando, Florida, March 1985.

67. S. M. Spitzer and S. Schwartz, The effects of dielectric overcoating on electromigration in aluminum interconnections, *IEEE Trans. Electron Devices* **ED-16,** 348 (1969).
68. L. D. Yau, C. Hong, and D. L. Crook, Passivation material and thickness effects on the MTTF of Al–Si metallization, *Proc. 23rd Int. Rel. Phys. Symp.,* 115, Orlando, Florida, March 1985.
69. B. J. Root and T. Turner, Wafer level electromigration tests for production monitoring, *Proc. 23rd Int. Rel. Phys. Symp.,* 100, Orlando, Florida, March 1985.
70. R. C. Weast, ed., "Handbook of Chemistry and Physics." CRC Press, Cleveland, Ohio, 1975.
71. P. Merchant and T. Cass, Comparative electromigration tests of Al–Cu alloys, *Proc. 22nd Int. Rel. Phys. Symp,* 259, Las Vegas, Nevada, April 1984.
72. E. Levine and J. Kitcher, Electromigration induced damage and structure change in Cr–Al/Cu and Al/Cu interconnection lines, *Proc. 22nd Int. Rel. Phys. Symp.,* 242, Las Vegas, Nevada, April 1984.
73. D. S. Gardner, T. L. Michalka, T. W. Barbee, Jr., K. C. Saraswat, J. P. McVittie, and J. D. Meindl, Aluminum alloys with titanium, tungsten and copper for multilayer interconnections, *Proc. 1st Intl. IEEE VLSI Multilevel Interconnection Conf.,* 68, New Orleans, Louisiana, June 1984.
74. B. N. Agarwala, G. Digiacomo and R. R. Joseph, Electromigration damage in aluminum–copper films, *Thin Solid Films* **34,** 165 (1976).
75. F. Fischer and F. Neppl, Sputtered Ti-doped Al–Si for enhanced interconnect reliability, *Proc. 22nd Int. Rel. Phys. Symp.,* 190, Las Vegas, Nevada, April 1984.
76. M. Hatzakis, B. J. Canavello, and J. M. Shaw, Single-step optical lift-off process, *IBM J. Res. Dev.* **24,** 452 (1980).
77. J. R. Lloyd and J. A. Knight, The relationship between electromigration-induced short-circuit and open-circuit failure times in multi-layer VLSI technologies, *Proc. 22nd Int. Rel. Phys. Symp.,* 48, Las Vegas, Nevada, April 1984.
78. M. J. Attardo and R. Rosenberg, Electromigration damage in aluminum film conductors, *J. Appl. Phys.* **41,** 2381 (1970).
79. S. Vaidya and A. R. Sinha, Effect of texture and grain structure on electromigration in Al—0.5% Cu Thin Films, *Thin Solid Films* **75,** 253 (1981).
80. D. M. Brown, M. Ghezzo, and J. M. Pimbley, Trends in advanced process technology —submicrometer CMOS device design and process requirements, *Proc. IEEE* **74,** 1678 (1986).
81. E. T. Severn, H. H. Huston, and J. R. Lloyd, Acoustic emission study of electromigration damage in Al–Cu thin film conductor stripes, *Proc. 22nd Int. Rel. Phys. Symp.,* 256, Las Vegas, Nevada, April 1984.
82. J. M. Pimbley and D. M. Brown, Current crowding in high density VLSI metallization structures, *IEEE Trans. Elec. Dev.* **ED-33,** 1399 (1986).
83. R. E. Jones, Jr., and L. D. Smith, Contact spiking and electromigration passivation cracking observed for titanium layered aluminum metallization, *Proc. 2nd Intl. IEEE VLSI Multilevel Interconnection Conf.,* 194, Santa Clara, California, June 1985.
84. J. G.-J. Chern, W. G. Oldham, and N. Cheung, Contact electromigration-induced leakage failure in aluminum–silicon to silicon contacts., *IEEE Trans. Electron Devices* **ED-32,** 1341 (1985).
85. J. G.-J. Chern, W. G. Oldham, and N. Cheung, Electromigration in Al/Si contacts—induced open-circuit failure, *IEEE Trans. Electron Devices* **ED-33,** 1256 (1986).
86. G. S. Prokop and R. R. Joseph, Electromigration failure at aluminum–silicon contacts, *J. Appl. Phys.* **43,** 2595 (1972).
87. F. Neppl, F. Fischer, and U. Schwabe, $TaSi_x$ as a barrier between Al-based metalliza-

tion and n^+- and p^+-Si for reliable VLSI contacts, *Proc. 22nd Int. Rel. Phys. Symp.*, 185, Las Vegas, Nevada, April 1984.
88. W. Hasse, J. Schulte, J. Graul, and H. Schulte, Electromigration phenomena in $TaSi_2/n^+$ poly-Si layers, *Proc. 2nd Intl. IEEE VLSI Multilevel Interconnection Conf.*, 211, Santa Clara, California, June 1985.
89. K. Yamabe and K. Taniguchi, Time-dependent dielectric breakdown of thin thermally grown SiO_2 films, *IEEE Trans. Electron Devices* **ED-32,** 423 (1985).
90. P. Solomon, Breakdown in silicon oxide—a review, *J. Vac. Sci. Technol.* **14,** 1122 (1977).
91. D. L. Crook, Method of determining reliability screens for time dependent dielectric breakdown, *Proc. 17th Int. Rel. Phys. Symp,* 1, San Francisco, California, April 1979.
92. E. S. Anolick and C. Y. Chen, Application of step stress to time-dependent breakdown, *Proc. 19th Int. Rel. Phys. Symp.*, 23, Orlando, Florida, April 1981.
93. E. Domangue, R. Rivera, and C. Shepard, Reliability prediction using large MOS capacitors, *Proc. 22nd Int. Rel. Phys. Symp.*, 140, Las Vegas, Nevada, April 1984.
94. R. A. Metzler, Theoretical justification for the log normal distribution of time-dependent breakdown in MOS oxides, *Proc. 17th Int. Rel. Phys. Symp.*, 238, San Francisco, California, April 1979.
95. D. Wendell, D. Segers, and B. Wang, Predicting oxide failure rates using the matrix of a 64K DRAM chip, *Proc. 22nd Int. Rel. Phys. Symp.*, 113, Las Vegas, Nevada, April 1984.
96. E. S. Anolick and G. R. Nelson, *Proc. 17th Int. Rel. Phys. Symp.*, 8, San Francisco, California, April 1977.
97. Y. Hokari, T. Baba, and N. Kawamura, *IEDM Tech. Dig.*, **46,** San Francisco (1982).
98. D. A Baglee, Characteristics and reliability of 100 angstrom oxides, *Proc. 22nd Int. Rel. Phys. Symp.*, 152, Las Vegas, Nevada, April 1984.
99. J. W. McPherson and D. A. Baglee, Acceleration Factors for thin gate oxide stressing, *Proc. 23rd Int. Rel. Phys. Symp.*, 1, Orlando, Florida, March 1985.
100. J. W. McPherson and D. A. Baglee, Acceleration factors for thin gate oxide stressing, *J. Electrochem. Soc.* **132,** 1903 (1985).
101. P. Solomon, N. Klein, and M. Albert, A statistical model for step and ramp voltage breakdown tests in thin insulators, *Thin Solid Films* **35,** 321 (1976).
102. S. I. Raider, Time-dependent breakdown of silicon dioxide films, *Appl. Phys. Lett.* **23,** 34 (1973).
103. C. M. Osburn and S. I. Raider, The effect of mobile sodium ions on field enhancement dielectric breakdown in SiO_2 films on silicon, *J. Electrochem. Soc.* **120,** 1369 (1973).
104. T. Kusaka, Y. Ohji, and K. Mukai, Time-dependent dielectric breakdown of ultra-thin silicon oxide, *IEEE Elec. Dev. Lett.* **EDL-8,** 61 (1987).
105. L. W. Nordheim, The effect of the image force on the emission and reflection of electrons by metals, *Proc. Roy. Soc. (London)* **A121,** 626 (1928).
106. R. H. Fowler and L. W. Nordheim, Electron emission in intense electric fields, *Proc. Roy. Soc. (London)* **A119,** 173 (1928).
107. E. L. Murphy and R. H. Good, Jr., Thermionic emission, field emission and the transition region, *Phys. Rev.* **102,** 1464 (1956).
108. N. Klein and P. Solomon, Current runaway in insulators affected by impact ionization and recombination, *J. Appl. Phys.* **47,** 4364 (1976).
109. T. H. DiStefano and M. Shatzkes, Impact ionization model for dielectric instability and breakdown, *Appl. Phys. Lett.* **25,** 685 (1974).
110. Y. Nissan-Cohen, J. Shappir, and D. Frohman-Bentchkowsky, High field current induced positive charge transients in SiO_2, *J. Appl. Phys.* **54,** 5793 (1983).

111. S. Holland, I. C. Chen, T.-P. Ma, and C. Hu, On physical models for gate oxide breakdown, *IEEE Elec. Dev. Lett.* **EDL-5,** 302 (1984).
112. I. C. Chen, S. Holland, and C. Hu, Hole trapping and breakdown in thin SiO_2, *IEEE Elec. Dev. Lett.* **EDL-7,** 164 (1986).
113. I. C. Chen, S. Holland, and C. Hu, A quantitative physical model for time-dependent breakdown in SiO_2, *Proc. 23rd Int. Rel. Phys. Symp.,* 24, Orlando, Florida, March 1985.
114. S. Haddad and M.-S. Liang, Improvement of thin gate oxide integrity using through-silicon gate nitrogen ion implantation, *IEEE Elec. Dev. Lett.* **EDL-8,** 58 (1987).
H.-H. Tsai, L.-C. Wu, C.-Y. Wu, and C. Hu, The effects of thermal nitridation conditions on the reliability of thin nitrided oxide films, *IEEE Elec. Dev. Lett.* **EDL-8,** 143 (1987).
115. B. Ricco, M. Ya. Azbel, and M. H. Brodsky, Novel mechanism for tunneling and breakdown of thin SiO_2 films, *Phys. Rev. Lett.* **51,** 1795 (1983).
116. M. Shatzkes, M. Av-Ron, and R. A. Gdula, Defect-related breakdown and conduction in SiO_2, *IBM J. Res. Dev.* **24,** 469 (1980).
117. M. Shatzkes and M. Av-Ron, Statistics of breakdown, *IBM J. Res. Dev.* **25,** 167 (1981).
118. M. Shatzkes and M. Av-Ron, Determination of breakdown rates and defect densities in SiO_2, *Thin Solid Films* **91,** 217 (1982).
119. M. H. Woods, MOS VLSI reliability and yield trends, *Proc. IEEE* **74,** 1715 (1986).
120. A. Berman, Time-zero dielectric reliability test by a ramp method, *Proc. 19th Int. Rel. Phys. Symp.,* 204, Orlando, Florida, April 1981.
121. W. Kern, Detection and characterization of localized defects in dielectric films, *RCA Rev.* **34,** 655 (1973).
122. N. J. Chou and J. M. Eldridge, Effects of material and processing parameters on the dielectric strength of thermally grown SiO_2 films, *J. Electrochem. Soc.* **117,** 1287 (1970).
123. C. M. Osburn and D. W. Ormond, Dielectric breakdown in silicon dioxide films on silicon, *J. Electrochem. Soc.* **119,** 591 (1972).
124. A. Bhattacharyya, C. Vorst, and A. H. Carim, A two-step oxidation process to improve the electrical breakdown properties of thin oxides, *J. Electrochem. Soc.* **32,** 1900 (1985).
125. S. K. Lai, D. R. Young, J. A. Calise, and F. J. Feigl, Reduction of electron trapping in silicon dioxide by high temperature nitrogen anneal, *J. Appl. Phys.* **52,** 5691 (1981).
126. J. M. Green, C. M. Osburn, and T. O. Sedgwick, The influence of silicon heat treatments on the minority carrier concentration and the dielectric breakdown in MOS structures, *J. Electron. Mater.* **3,** 579 (1974).
127. W. K. Meyer and D. L. Crook, A non-aging screen to prevent wearout of ultra-thin dielectrics, *Int. Rel. Phys. Symp. Proc.,* 6 (1985).
128. M. Shatzkes, M. Av-Ron, and K. V. Srikrishnan, Determination of reliability from ramped voltage breakdown experiments: application to dual dielectric MIM capacitors, *Int. Rel. Phys. Symp. Proc.,* 138 (1984).
129. T. Ito, T. Nakamura, and H. Ishikawa, Advantages of thermal nitride and nitroxide gate films in VLSI process, *IEEE Trans. Electron Devices* **ED-29,** 498 (1982).
130. T. Ito, T. Nozaki, and H. Ishikawa, Direct thermal nitridation of silicon dioxide films in anhydrous ammonia gas, *J. Electrochem. Soc.* **127,** 2053 (1980).
131. S.-T. Chang, N. M. Johnson, and S. A. Lyon, Capture and tunnel emission of electrons by deep levels in ultra-thin nitrided oxides on silicon, *Appl. Phys. Lett.* **44,** 316 (1984).

132. T. Ito, H. Arakawa, T. Nozaki, and H. Ishikawa, Retardation of destructive breakdown of SiO_2 films annealed in ammonia gas, *J. Electrochem. Soc.* **127,** 2248 (1980).
133. M. M. Moslehi and K. C. Saraswat, Thermal nitridation of Si and SiO_2 for VLSI, *IEEE Trans. Electron Devices* **ED-32,** 106 (1985).
134. J. A. Nemetz and R. E. Tressler, Thermal nitridation of silicon and silicon dioxide for thin gate insulators, part I, *Solid State Tech.* **26**(2), 79 (1983).
J. A. Nemetz and R. E. Tressler, Thermal nitridation of silicon and silicon dioxide for thin gate insulators, part II, *Solid State Tech.* **26,**(9), 209 (1983).

Chapter 6

Yield

Key to the success of any semiconductor technology is the ability to yield circuits. Without yield the technology cannot survive if for no other than purely economic reasons. If a technology other than the one in question can produce the desired end product for a lower cost, it will by definition be the survivor. As a result of this yield, considerations begin with the very first process considerations and influence all phases of process development and utilization. In this chapter, we will consider yield and yield enhancement concepts in detail. The chapter will begin with a review of basic yield concepts from definitions to yield model development and applications. Sources of yield loss will then be considered. With this background, the chapter will then go on to discuss concepts for yield enhancement methodology. Topics in this area will address selection of a yield vehicle as well as in-process and postprocess characterization of defects and yield loss mechanisms.

I. YIELD BASICS

A. Importance of Yield

Yield can be defined as the ratio of the number of working circuits to the total number of die locations introduced into fabrication. This encompasses a broad range of process technology, from a blank silicon wafer to a fully packaged device ready for system insertion. As a result, it is convenient to separate yield into the three components of line yield, Y_l, probe yield, Y_p, and assembly yield, Y_a. Line yield is simply the ratio of wafers

completed to wafers started. Yield loss mechanisms in this category are the result of gross misprocessing, such as wafer breakage or severe overetch, that either physically destroys the wafer or makes it pointless to continue processing from a yield standpoint. Probe yield is the result of yield loss mechanisms introduced during the processing (lithography defects, particulates, scratches, etc.) that cause some number of chips to fail. Assembly yield is the yield associated with dicing and packaging the final product.

Of these yield components, probe yield is the component of significance to the yield enhancement effort. Both line and assembly yields are relatively high and stable in mature facilities with respect to product and process. Probe yield, on the other hand, is very sensitive to product (chip size, packing density, layout) and process (feature size, complexity, throughput) and is thus amenable to yield enhancement efforts. Usually, line yield falls in the 80% to high 90% range, depending on the process complexity and the maturity of the line. For a modern process on a mature line, this number is typically in the low-to-mid-90 percentile. This is quite amazing when one considers that a modern VLSI CMOS process can have on the order of 150 process steps implying an individual step yield of 99.93% or better. Assembly yield runs typically on the order of 85% and seems to be relatively independent of the product. The reason for this is that basically the same number and types of steps are required for all chip assemblies. Improvements in packaging technology tend to show up primarily in reliability improvements and to a much lesser degree in yield improvements. Probe yield, as previously mentioned, is highly susceptible to unit step operations either from contamination or mishandling. Typically, products are introduced at a 10% yield level and through yield enhancement efforts are increased to the order of a 50% yield level over a one-and-one-half to two-year period. Then a new product is introduced and the yield enhancement game starts all over again.

Since probe yield is the component of interest for yield enhancement activities, we will use the term yield as synonymous with probe yield for the remainder of this chapter. Total yield, Y_T, will be used to designate the product of the component yields:

$$Y_T = Y_l Y_p Y_a. \tag{6.1}$$

Yield reflects the number of good die per wafer and thus chip cost and system cost are both inversely proportional to yield. Therefore, yield is of utmost importance to both circuit designer and process engineer. With regard to the designer, the technology chosen determines the basic cost per wafer leaving yield as the only leverage on chip cost. This results in a yield versus chip partitioning of the design as shown in Fig. 6.1. Chip cost decreases with the number of chips used to implement a given design

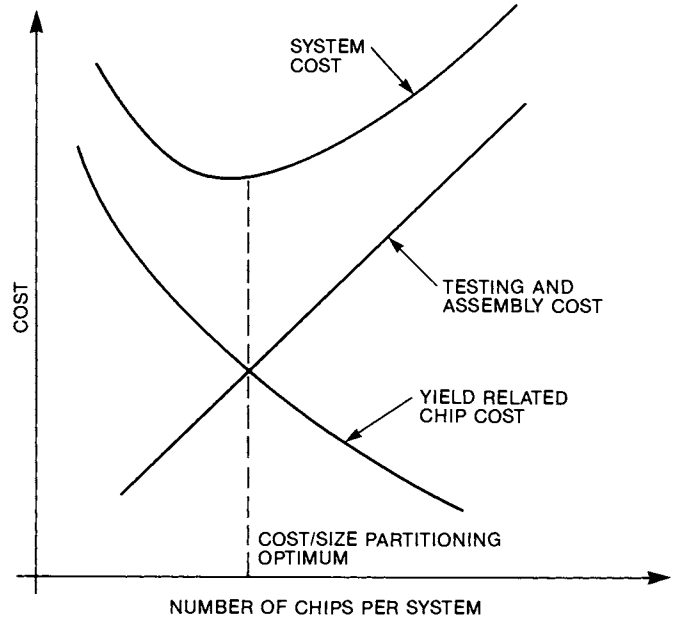


Fig. 6.1. The effects of chip cost and test cost on system partitioning.

because the yield increases as chip size decreases. However, the testing and assembly cost increases essentially linearly with the number of chips. Therefore, based on test requirements, assembly, and yield, one can determine an optimum configuration for minimum system cost. In general, the optimum cost will be achieved by partitioning the system into more than one chip. In spite of this, there will always be a drive to integrate more and more circuit elements on a single chip because the cost of testing and assembly is significantly greater than the cost of the chip. As a result, there are usually gains to be made by increasing chip complexity. For the same reasons that lot cost is relatively constant for a given technology and thus die cost is completely dependent on yield, it is of utmost importance to the process engineer. Improvement in yield, through yield enhancement and process engineering as the process goes down the learning curve, dictates competitive pricing.

The importance of the impact of yield on chip cost can be illustrated by considering the learning curve in closer detail. As pointed out by Cunningham [1], history has shown that a proportional change in manufacturing volume results in a proportional reduction in cost for a range of technologies from Model T Ford production in 1910 to integrated circuits in 1977. Furthermore, when cost was plotted logarithmically with volume,

some 15 different cases considered demonstrated constant slope learning curves ranging from 60% to 90%, as shown in Fig. 6.2. In general people-dominated industries, such as integrated circuit production, show faster learning (i.e., lower slopes) than totally automated activities, such as catalytic cracking in the petroleum industry. Typically, integrated circuits have shown dramatic learning curves with slopes in the low- to mid-70% range, resulting in constant and sometimes dramatic price reductions. The main reason for this has been the yield improvements that have determined the slope of the learning curve for the process area. The learning curve can be expressed mathematically by the differential equation

$$dC/C = -m(dV/V), \tag{6.2}$$

where C is the unit cost, V is the manufacturing volume, and m is a constant of proportionality. The negative sign results from the fact that cost decreases with increasing production volume. Integrating Eq. (6.2)

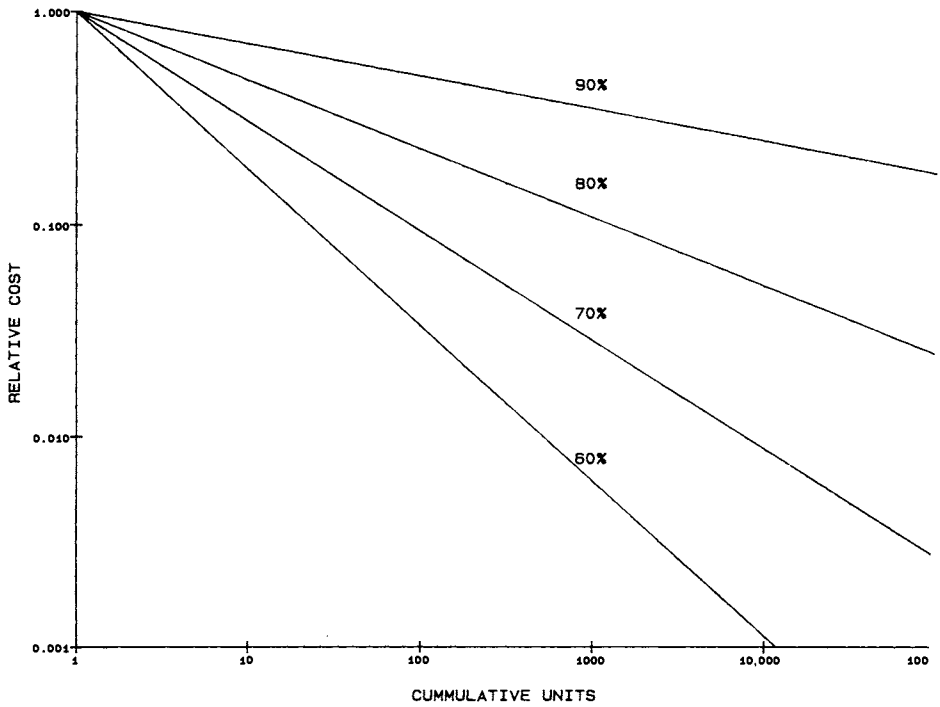


Fig. 6.2. Generalized learning curve showing cost reduction with cumulative units produced as a function of slope.

over the number of volume doublings, n, as the cost is reduced from C1 to C2 gives

$$\ln \frac{C_2}{C_1} = -mn(\ln 2). \tag{6.3}$$

The slope of the learning curve is defined as the percentage decrease in cost for one volume doubling. Mathematically, from Eq. (6.3), this is just

$$s = e^{-m(\ln 2)} = 1/2^m, \tag{6.4}$$

with the percentage expressed numerically (i.e., $s = 0.6$ represents 60%). When analyzing learning curves, the number of volume doublings over a given manufacturing volume can be calculated:

$$n = \frac{\ln(V_2/V_1)}{\ln 2}, \tag{6.5}$$

and this number used in Eq. (6.3) to calculate m and then the slope determined by Eq. (6.4). Now, the fact of the matter is that traditionally a unit volume doubling has been used for the definition of the learning curve slope. Obviously, one could use any arbitrary volume multiple, a, and integrate Eq. (6.2) from V to Va^n and define an equally valid learning curve slope. However, tradition prevails and one must realize that a learning curve slope expressed as a percentage carries with it implicitly a doubling of the unit volume.

The impact of yield learning on chip cost for the process engineer can now be illustrated with an example. Consider two facilities, A and B, that are selling 1-megabit DRAMS fabricated in 1.25-μm CMOS VLSI technology on 150-mm wafers with a chip size of 58 mm^2. Both facilities are 10,000 square feet and can produce 8500 wafer outs per month. For this analysis, we will assume that within one month of production initial design bugs have been worked out and both A and B are producing the DRAMS at a total yield of 10%, including an 85% assembly yield. Now, up to this point, everything has been equal and both facilities are headed for full-scale production. The difference is that facility A has a faster turnaround yield enhancement effort which allows them to follow a 65% learning curve whereas facility B follows a 75% learning curve. Based on this volume, Eq. (6.5) shows that both facilities achieve 3.58 volume doublings in a year's time. The parameter m for both facilities can be calculated from Eq. (6.4), using their learning curve slope and then the yield calculated from Eq. (6.3) based on the reciprocal relationship between yield and cost. The results of such calculation are summarized in Fig. 6.3. This figure shows that facility A climbs to 50% yield over the first year of production because of its more

efficient yield enhancement program. Facility B, on the other hand, achieves only 30% yield in the same time frame. With this yield difference, facility A can continually cut prices until facility B is forced to sell at a loss and has to drop out of the business. Now, it is worth noting that it is not just volume production alone that forces cost down a learning curve. All too often people loose sight of this fact and concentrate only on increasing throughput. Granted, fast throughput is important, but of equal, if not more, importance to determination of the learning curve slope is a rapid turnaround interactive feedback through process engineering for yield improvements.

B. Definitions

The preceding example has shown that yield has to be continually monitored and improved for a facility to be successful. This implies some sort of yield analysis and improvement program throughout the process life. The remainder of this chapter will be devoted to developing a better understanding of yield concepts, yield loss mechanisms, and the elements of a viable yield enhancement strategy. To assist us in this task, we will first give consideration to some basic definitions. We assume circular wafers of

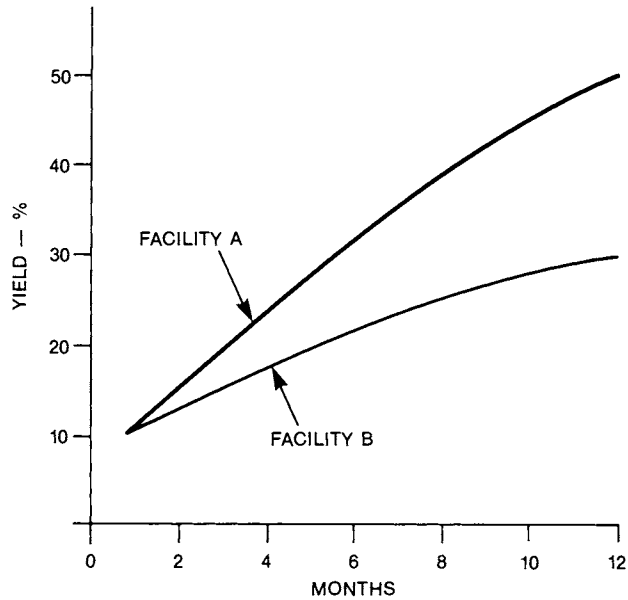


Fig. 6.3. Yield improvement resulting from facility A on a 65% learning curve compared with facility B on a 75% learning curve.

radius R, each containing N rectangular chips of length a and width b and hence of area A:

$$A = ab. \tag{6.6}$$

The chips are rectangular so that they can be laid out in an array separated by scribe lanes for ease of sawing the wafer into individual die. Clearly, one wants to maximize the number of chips on a given wafer within the constraints of wafer area, chip area, and scribe lane dimensions. Elaborate computer programs have been developed to optimize the number of chips per wafer; however, a simple formula for calculating N, from Cunningham and Jaffe [2] will be used here:

$$N = \pi \frac{(R - \sqrt{A})^2}{A}. \tag{6.7}$$

This formulation simply accounts for edge die loss by reducing the wafer radius by a dimension proportional to the linear edge of the chip. Then the number of chips per wafer is obtained by dividing the resulting wafer area by the area of a single chip. In general it is not the total area, A, of the chip that is sensitive to defects but some smaller critical area, A_c, containing high-density patterned regions. Hence, for yield analysis, we are more interested in A_c than A. We will discuss the relationship between A and A_c in more detail later in the chapter. If there are M killer defects on the wafer, the average number of killer defects per chip, d, is given by

$$d = M/N. \tag{6.8}$$

Based on the critical area of the chip, we can therefore define an average defect density, D_o, as

$$D_o = M/NA_c. \tag{6.9}$$

Both the average number of defects per chip and the average defect density are important numbers that are often used in yield analysis. Combining Eqs. (6.8) and (6.9), we see that the average number of defects per chip can be alternatively expressed as

$$d = D_o A_c. \tag{6.10}$$

Histograms of the results of fabrication are often used in yield analysis [3]. Consider a collection of X wafers containing $N(k)$ chips with k defects per chip. This collection of wafers could represent a fabrication lot, splits within a lot, or several lots over a period of time for the facility. The results could thus be plotted as a histogram showing the number of defective die with k defects per chip as illustrated in Fig. 6.4. The total number of chips in the collection of wafers is given by

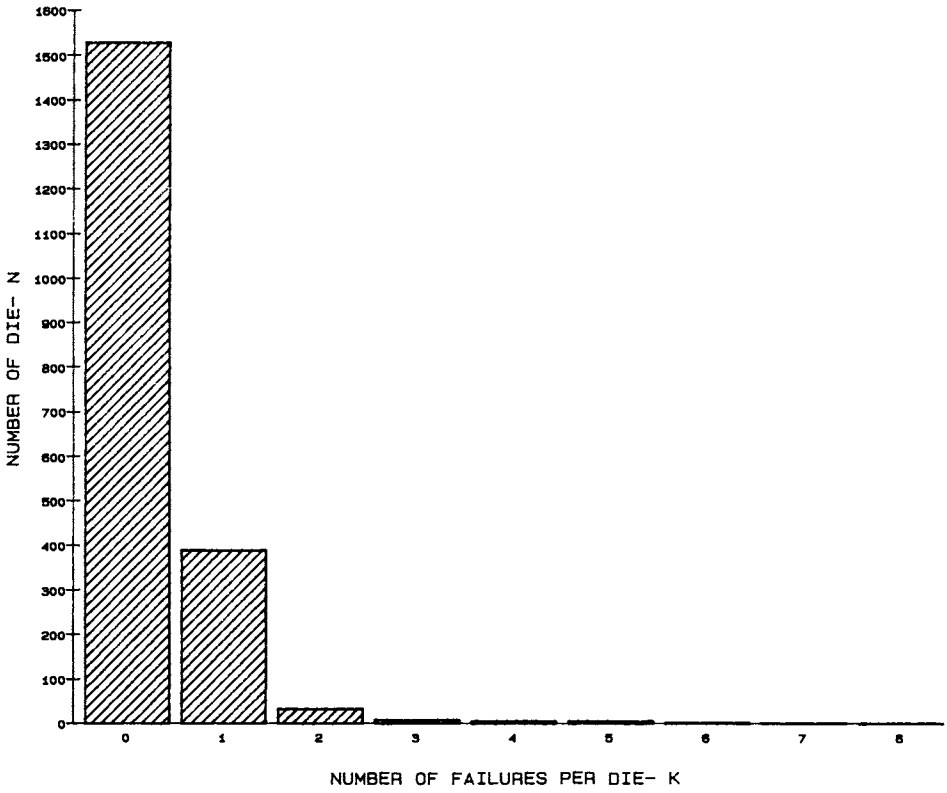


Fig. 6.4. Yield histogram showing the distribution of chips with k failures per chip.

$$N_T = \sum_{k=0} N(k). \tag{6.11}$$

Now, the average process yield is just the number of chips with zero defects divided by the total number of chips:

$$Y_{AVE} = \frac{N(0)}{\sum_{k=0} N(k)}. \tag{6.12}$$

The total number of defects is just the sum of the number of defects per chip times the number of chips with that many defects summed over all defective chips as given by

$$M_T = \sum_{k=1} kN(k). \tag{6.13}$$

Hence, the average number of defects per chip is just the total number of

defects divided by the total number of die that can be obtained from Eqs. (6.11) and (6.13):

$$d = \frac{\sum_{k=1} kN(k)}{\sum_{k=0} N(k)}. \tag{6.14}$$

A numerical example will illustrate the use of these equations. Consider the distribution of defects summarized in Table 6.1 and shown previously by the histogram of Fig. 6.4. These represent actual yield data over a three-month period from a 4-K SRAM used as a yield monitor on a VLSI 1.25-μm CMOS line. Using Eqs. (6.11) and (6.13), we see that the total number of chips and the total number of defects are 1978 and 566, respectively. The number of chips with zero defects from Table 6.1 is 1529, resulting in an average yield, from Eq. (6.12) of 77% over this time period. Using Eq. (6.14), the average number of defects per chip is 0.29, which is significantly less than one, as would be expected from a good yielding VLSI process. Another important point clearly shown by Table 6.1 and Fig. 6.4 is the relatively low number of chips with multiple defects. This again is a necessary condition for a well-yielding VLSI process.

The preceding example has demonstrated the usefulness of histograms in the analysis of yield. There is one problem, however. The yield calculated in this example is an *a posteriori* yield, that is, it is the yield calculated from experimental data for a chip of specific geometry and design. What is really needed is the ability to predict the yield for a design *a priori* before committing a design to fabrication. In other words, we need a yield model that takes experimentally obtained defect data from an operating line and projects the yield for any arbitrary circuit design.

TABLE 6.1
Number of Chips with k Defects/Chip, Experimental Data

k	$N(k)$	$kN(k)$
0	1529	
1	390	390
2	34	68
3	9	927
4	6	24
5	6	30
6	2	12
7	1	7
8	1	8

C. Yield Models

Yield models span a range of complexity from a very simple exponential model, $\exp(-AD)$, to sophisticated computer programs, such as VLASIC [4]. As we will see shortly, the exponential model known as the Poisson model derives from the simplest of considerations, namely, uniformly distributed random point defects. VLASIC, an acronym for VLSI layout simulation for integrated circuits, is a computer program developed at Carnegie Mellon University that uses Monte Carlo techniques to simulate the manufacture of faulty circuits. In conjunction with a process model, VLASIC places catastrophic point defects on a chip layout and analyzes the resulting fault. The information obtained can be used to predict yield, optimize design rules, and even evaluate the effectiveness of redundancy. In this chapter, we will concentrate only on the simple models used for yield analysis. These are the reasons for this approach:

1. In general, complex models are more useful for specific applications rather than global discussion.
2. Simple models show the proper dependencies for yield degradation with chip area and defect density.
3. Simple models, if applied with insight, can be surprisingly accurate and hence extremely effective in the yield enhancement effort.

Before we get too deeply involved with the model, we must first understand the nature of yield loss in integrated circuits. We will initiate this by stating that yield loss is a result of defects. This mandates the global definition of a defect as any imperfection in the wafer either from the starting material or introduced during the processing through physical steps or through the chip design and layout. Within this definition, defects fall into one of three classifications of point defects, line defects, or area defects. Point defects are oxide pinholes, isolated etch pits, particles, or process-related effects of particles, etc., that affect an area much smaller than the chip itself. Line defects are scratches, slip lines, etc., that have a high length to width aspect ratio. Area defects consist of misalignment, stains, cleaning problems, and most parametric design-related failures. In general, area defects affect an area as large or larger than the chip area itself. Typically, defect density is associated with randomly distributed point defects. Line and area defects are considered gross loss mechanisms that would adversely affect the yield of a die of any practical size. We will return to the subject of defect classification in more detail later in this chapter. For the present, we will simply state that the yield model consists of a chip area

independent component as well as a chip area dependent component:

$$Y = Y_G\, Y_R\, (A_c, D_o). \tag{6.15}$$

In this equation, Y_G represents the chip area independent term due to gross yield loss mechanisms. The other component represents the portion of yield governed by random defects and is therefore highly dependent on the critical chip area, A_c, and average defect density, D_o.

One of the first and perhaps most significant questions one encounters in yield is that of distinguishability of defects [3]. Namely, are defects distinguishable from one another in a statistical sense? The answer to this question is of fundamental importance to the development of a yield model. Consider the familiar statistical analogy of placing M balls (defects) in N urns (chips). The M balls are distinguishable if they can be labeled with identifiable numerical subscripts and kept track of over the duration of the experiment. The N urns are used with replacement if more than one ball can be placed into any one urn. Now, in the case of distinguishable defects, M defects can be placed on N chips with replacement in X_1 ways:

$$X_1 = N^M, \tag{6.16}$$

which leads to Maxwell–Boltzmann statistics. Using this placement, the probability that a given chip will contain k defects is given by the binomial distribution:

$$P(k) = \frac{M!}{k!(M-k)!}\,\frac{1}{NM}\,(N-1)^{M-k}. \tag{6.17}$$

Hence, the yield is given by the probability that a chip will contain zero defects, which is just

$$Y_{MB} = \left(1 - \frac{1}{N}\right)^M. \tag{6.18}$$

In the limit of large N and M such that the ratio of M/N remains finite, Eq. (6.18) reduces to the simple exponential yield

$$Y_{MB} = e^{-M/N} = e^{-A_c D_o}. \tag{6.19}$$

For the case of indistinguishable defects, M defects can be placed on N chips in X_2 ways:

$$X_2 = \frac{(N+M-1)!}{M!(N-1)!}, \tag{6.20}$$

which leads to Bose–Einstein statistics. In this case, the probability distri-

bution for a chip containing k defects is much more complicated as given by

$$P(k) = \frac{P(0)\left[\frac{M-k+1}{N}\right]\left[\frac{M-k+2}{N}\right]}{\left[1+\frac{M-k+1}{N}\right]\left[1+\frac{M-k}{N}\right]} \cdots \frac{\frac{M}{N}}{1+\frac{M-2}{N}}, \quad (6.21)$$

where $P(0)$ is the yield, that is, the probability for zero defects per chip:

$$P(0) = \frac{1-\frac{1}{N}}{1+\frac{M}{N}-\frac{1}{N}}. \quad (6.22)$$

In the limit of large N, Eq. (6.22) becomes

$$Y_{BE} = \frac{1}{1+A_c D_o}. \quad (6.23)$$

Thus, we see that two quite different yield models result in the limit of large N based upon the assumption of distinguishable or indistinguishable defects. The fact of the matter is that both Eqs. (6.19) and (6.23) are popular yield models, and both types of statistics are used. The general points regarding distinguishability of defects as presented in the literature [3] can be summarized as follows.

1. Distinguishability of defects does not depend on their exact wafer location.
2. Distinguishability of defects cannot be determined by electrical measurements alone.
3. Defects can be examined by other means (i.e., visual or SEM inspection) and classified by their physical nature.
4. Defects within a given classification can be considered indistinguishable.

This allows one the latitude to consider defects as either distinguishable or indistinguishable based on personal conviction. These authors' bias is toward distinguishable defects since it seems more a case of visually observable rather than distinguishability in the statistical sense. It would seem that if defects could be classified according to physical nature then they should be classifiable within a given group albeit with some difficulty.

Let us return to the case of the simple exponential yield of Eq. (6.19) to see if we can attach some physical significance to this model. This model was a result of the yield obtained from the binomial distribution under the

conditions of a large number of chips, N, and defects, M, such that M/N remains finite. Now, let us visualize each defect as a tiny light source and recognize that the conditions placed on the development of the exponential yield is essentially that the number of defects is on the same order as the number of chips. Hence, when we stand back and look at the wafer, we see a uniform distribution of light (i.e., defects). Therefore, the exponential model is the one we would expect under the physical conditions of uniform defect density. We also note that under these conditions the binomial distribution is reasonably approximated by the Poisson distribution, and thus the exponential model for yield is referred to as the Poisson yield model. Now, if we maintain the same large number of chips and start reducing the number of defects (i.e., turning off the lights), the light distribution from the wafer starts looking less and less uniform and appears clustered at various points on the wafer surface. In the limit, as M becomes small and N remains large, the yield from Eq. (6.18) rather than being expressed exponentially can be approximated as

$$Y \sim (1 - M/N). \tag{6.24}$$

Recalling that M/N is exactly equal to $A_c D_o$, which we have already noted is small, this equation can be rearranged as

$$Y \sim 1 - A_c D_o \sim \frac{1}{1 + A_c D_o}. \tag{6.25}$$

But wait a minute, this is just the Bose–Einstein model from Eq. (6.23)! Furthermore, these conditions comply with the conditions the Bose–Einstein model was developed under, that is, large N with no restrictions on M. What this means physically is that the wafer defect distribution, just as the lights in our light analogy, is nonuniform and clustered. However, the most interesting point is that even starting from opposite points of view (i.e., distinguishable versus indistinguishable defects) the same yield formula results under appropriate conditions.

Based on the preceding discussion, it is not surprising that in the early days of integrated circuits the Poisson model was applied with reasonable success for yield predictions. In those days, chips contained relatively few electronic functions. Consequently, the chips were small and large in number on the wafer. Also, clean rooms and clean-room practices were not what they are today, and thus defect densities were large by comparison and hence appeared much more uniform. However, as fabrication facilities and techniques improved, defect densities decreased and technology developed larger chips with more components. For these large chips, the Poisson model was shown to be a very pessimistic yield predictor. The discrepancy between measured and predicted yields led researchers to investigate varia-

tions in defect density. It was found that defect density was nonuniform, not only within a given wafer but from wafer to wafer within a lot. A typical example of defect density variation is the radial distribution exhibited by some defects that can be expressed as [5]

$$D(r) = D_c + D_R e^{(r-R)/L}, \tag{6.26}$$

where D_c is the defect density at wafer center, D_R is the defect density at wafer edge, r is the radial coordinate, R is the wafer radius, and L is the characteristic length of defect variation. Generally, this defect distribution is obtained from a composite average evaluated over a wafer lot since per wafer defect densities are too low to establish the demonstrated trend.

Murphy [6] was one of the first to recognize that yield formulas had to be modified to accommodate variations in defect density. He proposed that the simple Poisson model be extended to accommodate variations in defect density through the use of a defect probability density function, $F_{(D)}$, such that the yield is given by

$$Y = \int_0^\infty F_{(D)} e^{-A_c D}\, dD, \tag{6.27}$$

with the restriction on $F_{(D)}$ that

$$\int_0^\infty F_{(D)}\, dD = 1. \tag{6.28}$$

Basically, this equation teaches that within a unit area the yield is given by the Poisson yield with a local defect density D. Variations in D within a wafer and from wafer to wafer are accommodated through the normalized defect density distribution. It is interesting to note, as pointed out by Okabe *et al.* [7], that Eq. (6.27) is of the form of a LaPlace transform. Hence, if one can determine a given defect probability density distribution is applicable, the resulting yield equation can be obtained quite readily through the use of tables of LaPlace transforms. Several of these points can be demonstrated by considering the case of a uniform defect density. Here, there are no variations in defect density across the wafer so $F_{(D)}$ is just the delta function $\delta_{(D_o)}$. Substituting this function for $F_{(D)}$ into Eq. (6.27) gives the Poisson model:

$$Y = \int_0^\infty \delta_{(D_o)}\, e^{-A_c D}\, dD = e^{-A_c D_o}, \tag{6.29}$$

as expected for a uniform defect distribution.

Based on yield analysis experience, Murphy proposed a bell-shaped distribution for $F_{(D)}$ as shown by the dotted curve in Fig. 6.5. In order to

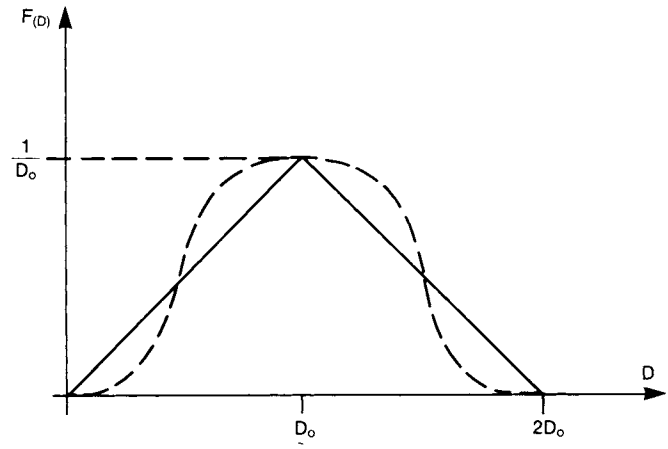


Fig. 6.5. Bell-shaped defect distribution with triangular approximation proposed by Murphy [6].

make the mathematics more tractable, he approximated this distribution with the triangular distribution shown by the solid lines in the figure and represented mathematically by

$$F_{(D)} = \begin{cases} \dfrac{D}{D_o^2} & 0 \le D \le D_o \\ \dfrac{2}{D_o} - \dfrac{D}{D_o^2} & D_o \le D \le 2D_o \\ 0 & D \ge 2D_o. \end{cases} \tag{6.30}$$

Substituting this equation for $F_{(D)}$ into Eq. (6.27) gives the well-known Murphy yield law for random defects:

$$Y = \left[\frac{1 - e^{-A_c D_o}}{A_c D_o}\right]^2. \tag{6.31}$$

Note that in the limit of large $A_c D_0$ the Murphy yield falls off as $(1/A_c D_o)^2$, which is much less pessimistic than the Poisson model. Hence, the model was more in agreement with experimental yield and gained early acceptance.

An even less pessimistic fall-off with large $A_c D_0$ is demonstrated by the yield model derived from Bose–Einstein statistics as given by Eq. (6.23). This in fact was the method Price [8] used to derive this yield law. The Price yield gained early acceptance because not only did it show a less pessimistic yield fall-off for larger area die, but it showed this for the right

reason. Namely, because A_c was increasing faster than D_o was decreasing, hence A_cD_o was on the overall increasing. Undoubtedly, the large acceptance of this yield law was also due in part to its simplicity and the ease with which it could be inverted to obtain defect density from experimental yield data. Price was not the only one to develop this model. Seeds [9] somewhat earlier arrived at this model through quite different reasoning. He made the common sense observation that $F_{(D)}$ should be an exponential function if the probability for a low defect density was high and the probability for a high defect density was low, as would be required to yield large area integrated circuits. Thus, he used the exponential defect density distribution as shown in Fig. 6.6. and given mathematically:

$$F_{(D)} = \frac{1}{D_o} e^{-D/D_o} \tag{6.32}$$

in Eq. (6.27) to obtain the yield model of Eq. (6.23) and repeated here for convenience:

$$Y = \frac{1}{1 + A_c D_o}. \tag{6.33}$$

We will refer to this as the Seeds–Price yield model acknowledging the contributions of both investigators. It is worthwhile mentioning that Okabe *et al.* [7] somewhat later provided further motivation for the exponential defect density distribution based on defect build-up considerations. Previously, we have shown that under limiting conditions the Maxwell–Boltzmann yield reduced to the Bose–Einstein yield. Of interest here is

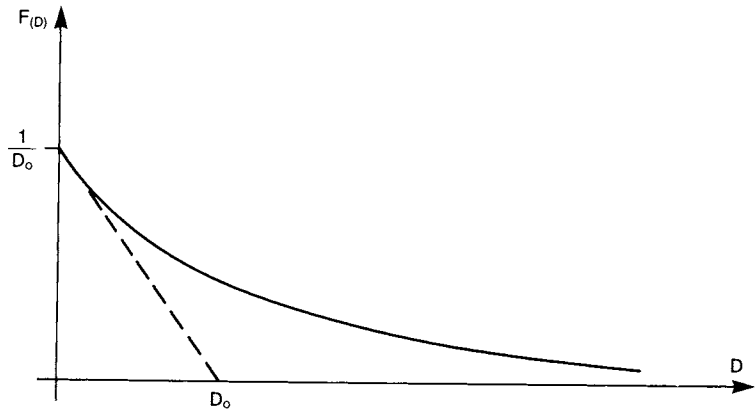


Fig. 6.6. Exponential defect density distribution.

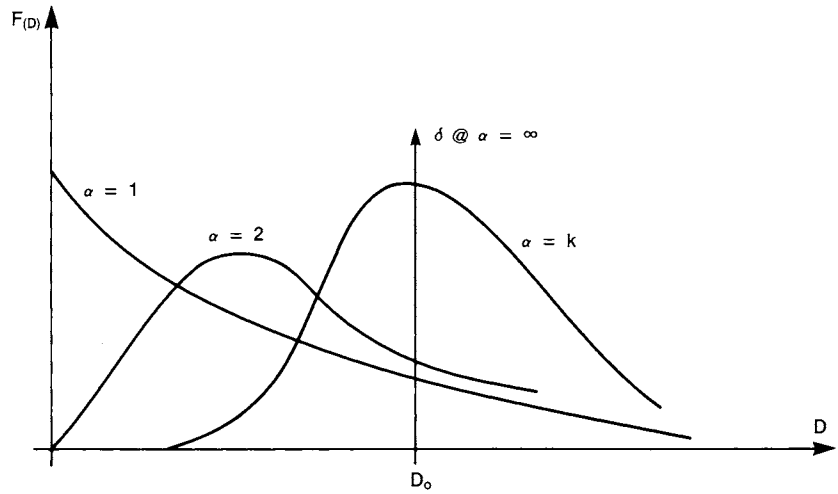


Fig. 6.7. Gamma defect density distribution for several values of α.

that we have shown more global convergence of classical and Bose–Einstein statistics in yield modeling through the use of an exponential defect density probability distribution. This serves only to illustrate that for practical considerations it is probably not so important as to what statistics are used as it is to how well the model describes the yield variation.

For a final model, we consider the one based on a gamma function for the defect distribution as described initially by Okabe *et al.* [7] and brought to popularity by Stapper [10,11]. The gamma distribution for defects is shown in Fig. 6.7 and is expressed mathematically as

$$F_{(D)} = \frac{\alpha^{\alpha}}{\Gamma(\alpha)} \frac{D^{\alpha-1}}{D_o^{\alpha}} e^{-\alpha D/D_o}. \tag{6.34}$$

Using this distribution in Eq. (6.27) results in the commonly referred to Stapper yield model for random defects:

$$Y = \left[\frac{1}{1 + \dfrac{A_c D_o}{\alpha}} \right]^{\alpha} \tag{6.35}$$

This model has gained considerable popularity because of its ability to fit experimental data. This is not surprising when one realizes that this distribution provides two parameters (α and D_o) for experimental fit. All previous models relied on just the single parameter D_o to provide the fit with experimental data. The parameter α is related to the gamma function

defect distribution through the average value and variance as

$$\alpha = \frac{D_o^2}{\text{Var}(D)}. \tag{6.36}$$

The general applicability of this model is further illustrated through consideration of different values for α. Note that $\alpha = 1$ corresponds to the case of the exponential distribution, whereas as $\alpha \to \infty$, the delta function is obtained. For intermediate values of α, various skewed distributions for $F_{(D)}$, as shown in Fig. 6.7 for $\alpha = 2$ and $\alpha = k$, are obtained.

D. Generalized Process Yield Model

We noted in Eq. (6.15) that, in general, yield is modeled by a gross component that is area independent and a random component that is a function of random defect density and chip area. Several models for the random yield component were presented in the previous section. Sometimes the yield for a multistep or full process is modeled simply by using the random yield model of choice in Eq. (6.15). A more useful formulation for yield results from considering Eq. (6.15) to be associated with the ith process step:

$$Y_i = Y_{Gi}\, Y_{Ri}(A_{ci}, D_i). \tag{6.37}$$

Recognizing that this equation represents the probability for zero defects for the ith process step allows us to write the yield for n independent process steps as

$$Y = Y_G \prod_{i=1}^{n} Y_{Ri}(A_{ci}, D_i). \tag{6.38}$$

Since the gross yield is not area dependent and is more operation rather than level dependent, it is often extracted from the multiplicand and lumped as a single constant:

$$Y = Y_G \prod_{i=1}^{n} Y_{Ri}(A_{ci}, D_i). \tag{6.39}$$

Some comments on the general applicability of Eq. (6.39) are in order. The formulation of the model with the area independent gross yield is essential for accurate process modeling. In a yield analysis application, mental note is usually taken of the sources of gross losses; however, the yield is usually separated out from level as indicated in Eq. (6.39). For the random yield component, virtually any of the models discussed in the

previous section could be used and most likely have been at one time or another. We will summarize some of the more common uses of the equations as presented in the literature shortly. For now we will just look at n in a more global manner. In the most general sense, the n components could represent every process step. However, in the context of a modern VLSI CMOS process with 150 or more individual process steps, this would be absurd. Generally, n is taken as some critical number of processing steps. In the early days of LSI and perhaps even into early VLSI, n was taken as the number of critical masking steps. However, as dry etching and thin film deposition have taken their rightful roles in VLSI processes, the number of critical steps, n, represents more a combination of masking, etching, and deposition steps. Typically, for a modern three- or four-conductor level CMOS process, n would be on the order 10–12. Yield models are often applied to a given facility. It must be recognized that the defect densities used in the simple models are process specific, that is, they are not global constants. This means that if a facility is running different processes of perhaps even different minimum feature size then defect densities should certainly be related, but not necessarily the same. However, within a given process the model should be able to appropriately handle issues such as packing density or changes in minimum feature size.

The formulation of Eq. (6.39) allows maximum modeling flexibility through the specification of chip area and defect density for each critical level. This can be an important consideration when one considers the packing density and minimum feature size for different processing levels. Variations in packing density are usually handled by representing $A_c D_o$ in the random yield component with a linear proportionality QAD_o. Several techniques have been employed to evaluate Q:

1. Empirical derivations from test structures of different geometries [12].
2. Mathematical formulations for simple structures that accommodate defect size distributions [13,14].
3. Density estimates from CAD plots or by measuring light transmission through the actual mask.
4. Computer simulations [4].

No one of these techniques is universally accepted or beneficial for that matter. The ease with which any technique can be applied is basically a function of the in-house capabilities of a particular facility. For example, to effectively use test structures, one must either have excess processing capabilities or excess wafer sites to accommodate defect density test structures. There is also a considerable effort required to correlate test structures with

actual circuit yield. Perhaps the easiest to apply are density measurements from either the actual masks or CAD plots. The accuracy and usefulness of such techniques often does not justify the effort. Computer techniques are perhaps the most accurate; however, program complexity is beyond that of simple scientific programming capability, and thus one must have the dedicated personnel available for the job. Because of the complexities and inaccuracies in evaluating Q, people often tend to consider it equal to one for VLSI and beyond and estimate it for other technologies, for example, $Q = 0.8-0.9$ for CMOS SOS depending on island spacing. Obviously there is an accuracy penalty here, but depending on the application of the model, it may be totally justifiable.

The impact on yield due to changes in minimum feature size for a constant process is usually handled through the defect density on a level by level basis. The number of defects are assumed to decrease monotonically with size according to a power law above some critical resolution size X_r:

$$S = 1/X^m. \tag{6.40}$$

Below the critical resolution size the defect linearly extrapolates to zero assuming that be it a lithography or etch step the defect is just not resolved. The actual power m is determined experimentally for a given facility. In the literature, one finds values of m in the range of $2 < m < 3$, [15,16]. Undoubtedly, the lower limit of 2 resulted from the particle size distributions in federal standard 209B [30] from which the class distinctions were defined. The upper limit of three was obtained from actual defect density measurements on test structures [16]. Consider a change in minimum feature size, such as might be due to a design shrink, for example, from minimum feature size X_1 to X_2. Assuming no improvements in defect density, the defect density D_2 at the new minimum feature size would be related to the previous defect density D_1 by the relationship

$$D_2 = D_1 \left[\frac{X_1}{X_2}\right]^m. \tag{6.41}$$

The relative number of defects based on parity at 3-μm size assuming square and cubic laws is illustrated in Fig. 6.8. As can be seen from the figure, the relative number of defects increases rapidly with decreases in minimum feature size and diverges rapidly for different power laws. Hence, it is extremely important to know the defect distribution as a function of size in order to assess the impact on yield for future technologies in a given facility.

We now return to the question of general usage of yield models as presented in the literature. As previously noted, all of the random defect

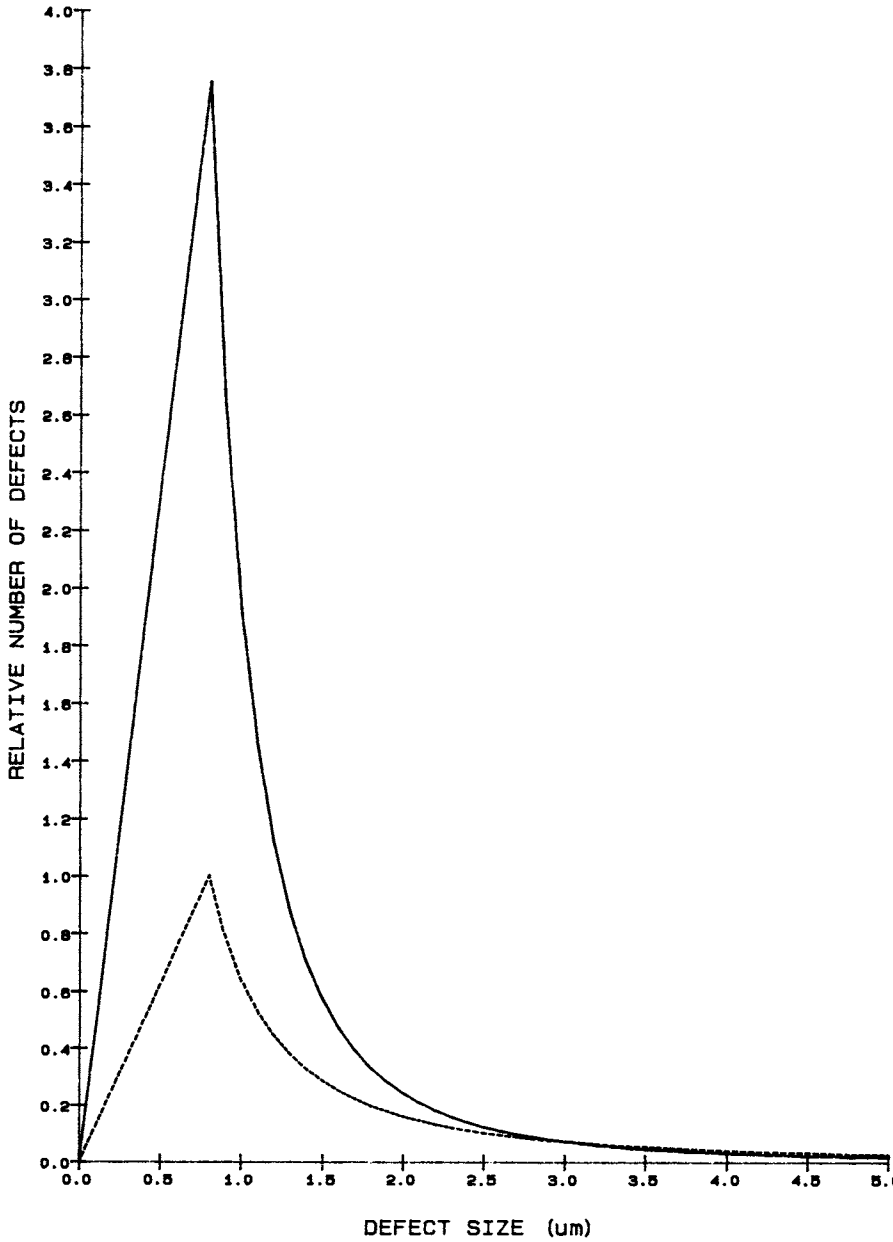


Fig. 6.8. Scaling implications for defects based on square and cubic defect size distributions with parity at 3.0 μm. Cubic law, ——; square law, ------.

models can be generalized to a multistep process, such as through the use of Eq. (6.39). The questions seem to be: Do we interpret all yield loss as random defects or do we include gross defects? Do we associate random defects with a common process defect density or do we break the process down into critical levels? The common usage for the various model representations as reported in the literature is summarized in Table 6.2. Both the Poisson and Murphy yield models are generally applied to an overall process by associating all yield loss to random defects (i.e., $Y_G = 1$) as

TABLE 6.2
Yield Model—Common Use

Poisson–Murphy

$$Y = e^{-A_c D_T} \tag{6.2.a}$$

$$Y = \left[\frac{1 - e^{-A_c D_T}}{A_c D_T}\right]^2 \tag{6.2.b}$$

Seeds–Price

$$Y = \frac{1}{1 + A_c D_T} \tag{6.2.c}$$

$$Y = Y_G \prod_{i=1}^{n} \frac{1}{1 + A_c D_{oi}} \tag{6.2.d}$$

$$Y = \left[\frac{1}{1 + A_c D_{ol}}\right]^n \tag{6.2.e}$$

Stapper

$$Y = Y_G \left[\frac{1}{1 + \frac{1}{\alpha} A_c D_{os}}\right]^{\alpha} \tag{6.2.f}$$

$$Y = Y_G \prod_{i=1}^{n} \left(\frac{1}{1 + \frac{1}{\alpha_i} A_c D_{oi}}\right)^{\alpha_i} \tag{6.2.g}$$

shown by Eqs. (6.2.a) and (6.2.b) in Table 6.2. On rare occasions one finds the Seeds–Price yield equation used in this same manner as indicated by Eq. (6.2.c). The unpopularity of this representation for VLSI is because of its extremely optimistic yield prediction for large area chips. Occasionally the Seeds–Price model is used in the most general sense as indicated by Eq. (6.2.d) with gross defect yield, Y_G, and critical area and defect density per critical level A_{ci} and D_{oi}, respectively. The use of this model representation is usually associated with yield enhancement efforts where it is desirable and beneficial to have knowledge of the actual defect density per level. However, in the open literature, the most common usage of this model is given by Eq. (6.2.c), where it is generalized to an average defect density per critical level, D_{ol}. In this form, the model is quite useful for general process comparisons. The model is convenient in that it does show a dependence on process complexity as well as on area and defect density and because it is easy to invert to obtain D_{ol} from experimental yield data. One must remember that D_{ol} is the effective average defect density per critical level and thus does not represent a real defect density. It can be related to the actual defect density per critical level by the equation

$$D_{ol} = \sum_{i=1}^{n} \frac{D_{oi}}{n}. \tag{6.42}$$

This representation of the yield is usually referred to as the Bose–Einstein model generalized to n process steps or simply as the Bose–Einstein model. The Stapper model is quite analogous to the generalized Seeds–Price or Bose–Einstein models. In fact, as pointed out earlier, these can be interpreted as special cases of the Stapper model with $\alpha = 1$. In the literature, one usually finds the Stapper model applied in the most general sense as indicated in Eq. (6.2.g). The model has been claimed to be extremely accurate for both yield enhancement and yield prediction. This is no doubt due to the fact that there are two fitting parameters (α and D_{oi}) per critical level for the defect distribution. The relaxation of this model to an overall process descriptor as shown by Eq. (6.2.f) also occurs frequently in the literature [10,11,17,28].

Finally, as a comparison of these models, we show them plotted as a function of the average number of defects per chip in Fig. 6.9. Looking at this curve, one can envision the thought process that might have gone on in yield model development. In the early days, defects were numerous and

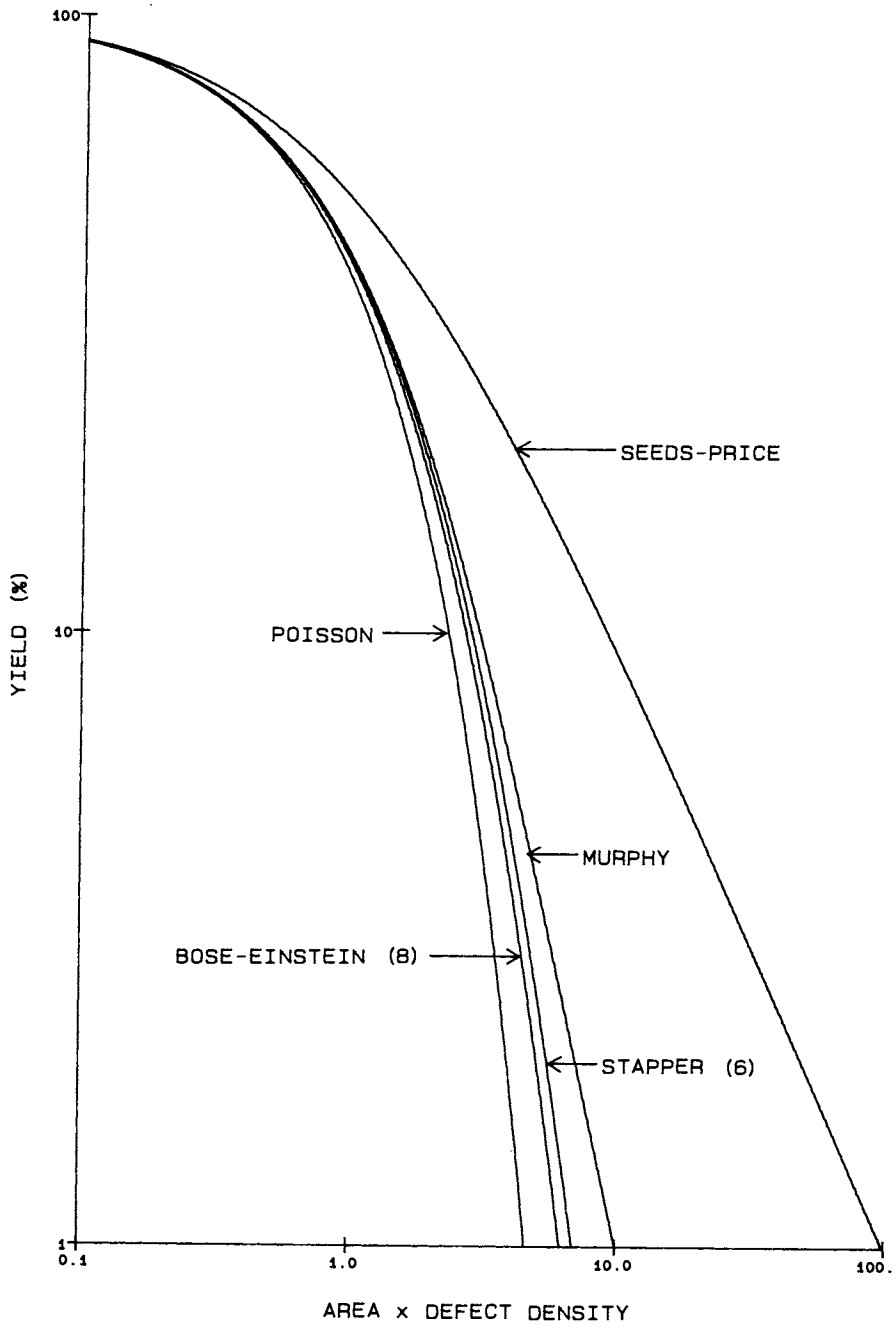


Fig. 6.9. Comparison of random yield models as a function of the average number of defects per chip, $A_c D_o$.

appeared randomly distributed. Also, chip areas were small and hence not only did the Poisson model seem theoretically sound it most likely gave reasonable agreement with experimental yield. As defect density was reduced and chip area increased, the Poisson model was found to be extremely pessimistic and thus interest was directed toward the Seeds–Price and Murphy yield models. As time went on, the Murphy model gained popularity because it more closely matched real circuit yields than the pessimistic Poisson model and the overly optimistic Seeds–Price model. Finally, the development of multistep models based on generalizing the Seeds–Price model or defect distribution fitting with the Stapper model took over because of their accurate representation of real circuit yields. One can see from this figure that the Murphy, Bose–Einstein, and Stapper models remain in close agreement over a rather broad range of parameters. In fact, practical economic reasons dictate yields greater than 10% are mandatory. Hence, if yield is where it should be, any of these three models can be used with reasonable success, and if the yield is extremely high, then any model can be used for yield estimation. If, on the other hand, the yield is lower than 10%, the situation is so grave that all attention is focused on yield enhancement and the issue of model accuracy is a rather moot point.

E. Uses of Yield Models

There are basically two general uses of yield models. The first is in a yield enhancement program where the model can be used as a guide to defect characterization as well as a measure of the process by monitoring defect density. The second is in the prediction of yield to determine specific product viability, evaluate cost for market planning, and determine defect requirements to meet technology trends. For both of these applications, the simple yield models extended to full process description, as discussed in the preceding paragraphs, are extremely useful and surprisingly accurate.

The application of simple yield models to a yield enhancement effort will be dealt with in more detail in subsequent sections of this chapter when defect characterization and yield enhancement methodology are discussed. For the present, we will outline their general applicability. As a guide to defect characterization in a yield enhancement effort, the simple models are most useful when represented by gross yield loss and random loss per critical level as indicated in Eq. (6.2.d) and (6.2.g) in Table 6.2. In this manner, gross yield loss and random defects can be separated and subclassified according to origin based on measurements of a yield monitor or an actual circuit. Such partitioning of defects is illustrated in Fig. 6.10 for a mature 1.25-μm CMOS process based on measurements of a SRAM

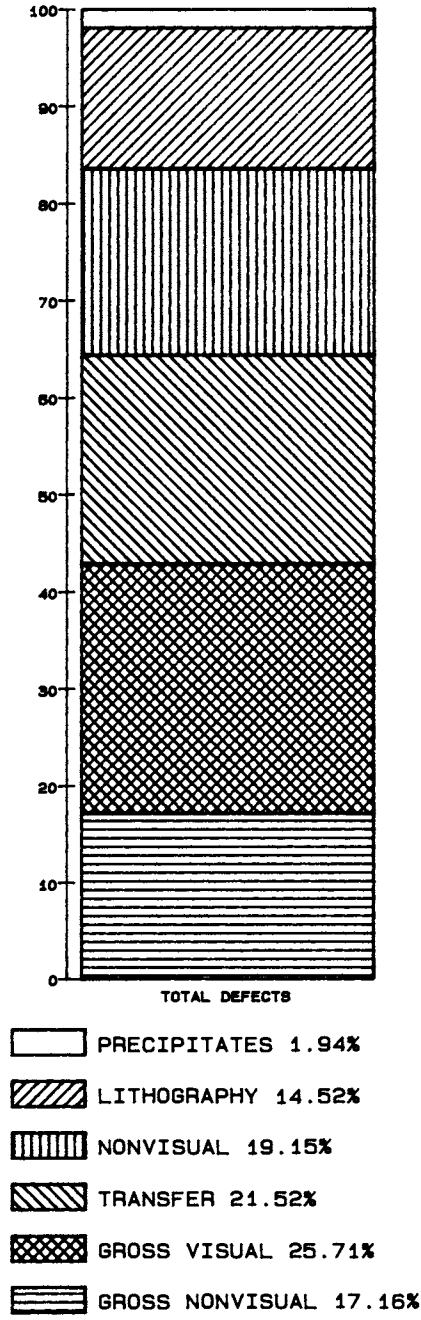


Fig. 6.10. Defect partitioning for gross and random defects for a mature process.

yield monitor. The maturity of the process is illustrated by the roughly equipartitioning of gross (43%) and random (57%) yield losses coupled with the additional knowledge that gross yield averaged 93% for the time interval of these measurements. The actual partitioning of defects into the subclasses indicated in this figure will be discussed later in this chapter. The comparison of the actual yield of the SRAM monitor with the calculated yield using a Bose–Einstein model [Eq. (6.2.d)] is shown in Fig. 6.11. This figure shows the actual lot by lot comparison of experimental yield versus calculated yield based on measured yield loss mechanisms over the time interval of the data shown in Fig. 6.10. As can be seen, the model is within a few percent of the experimental yield except when the yield is extremely low. This results from the fact that when yield is low all defect sources cannot be adequately characterized because some defects are masked by the presence of others. Nevertheless, the figure indicates that yield losses can be sufficiently classified to provide corrective actions to bring the yield back to acceptable standards. As seen by this example, the model provides a valuable aid in the correct classification of defects. If there was not such

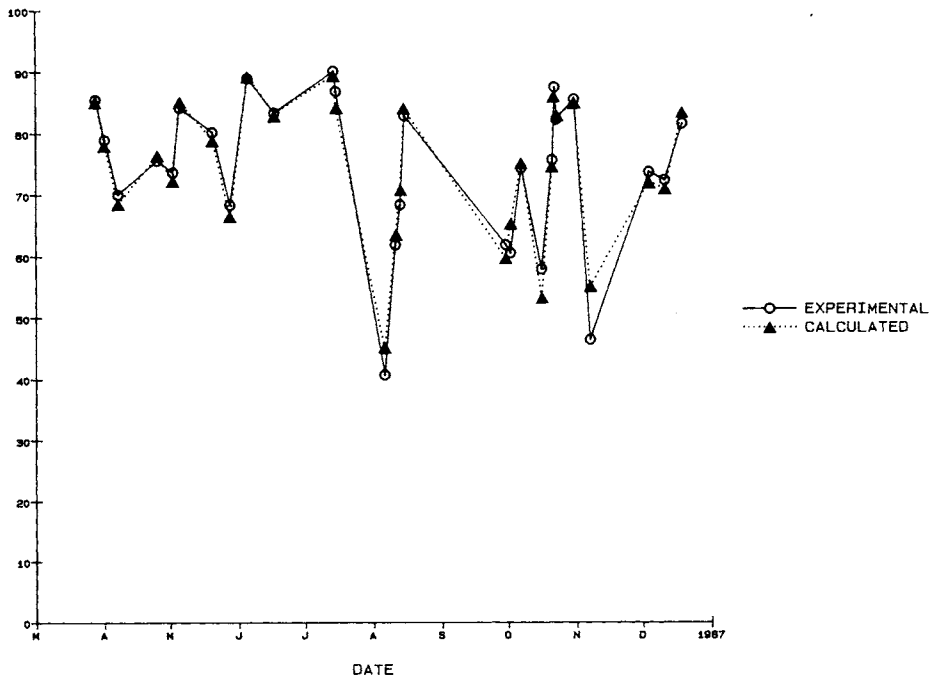


Fig. 6.11. Comparison of experimental and calculated yields for a 1.25-μm CMOS 4-K SRAM circuit using a Bose–Einstein yield model with 8 critical levels.

good correlation of model and experimental yield, then one could believe that important yield loss mechanisms were either not being adequately detected or classified. This would cause one to delve more deeply into the nature of the yield loss to determine the source of the discrepancy.

Another application of simple yield models in yield enhancement activities is as a monitor of the effectiveness of the yield enhancement effort itself. Here, we are not so interested in the exact classification of defect types or value of defect densities as we are in the relative improvement in defect density with time. Hence, just about any of the model representations given in Table 6.2 that classifies defect density by a single number [i.e., full process (6.2.a), (6.2.b), (6.2.c), and (6.2.f) or effective per level (6.2.e)] could be used. This type of tracking of yield enhancement progress for the random component of yield for a 1.25-μm process is shown in Fig. 6.12. This figure shows the experimental random yield component and the effective reduction in defect density per critical level obtained by inverting the Bose–Einstein model [Eq. (6.2.e)]. This representation clearly shows the improvements made by the yield enhancement effort over time with-

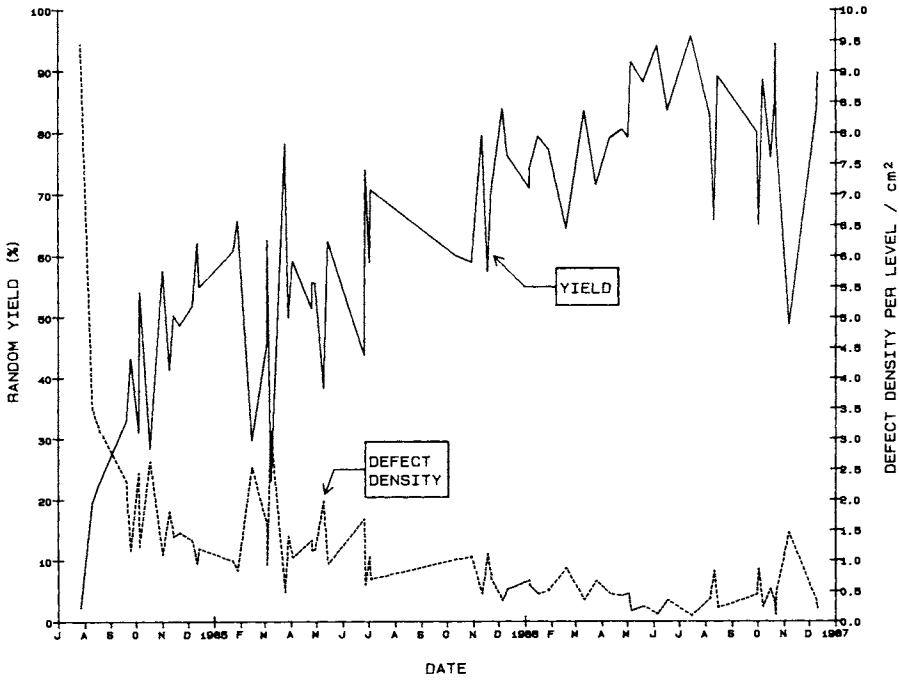


Fig. 6.12. Calculated effective defect density per critical level from experimental yield using a simple Bose–Einstein yield model with 8 critical levels.

out encumbering the viewer with the details of the exact loss mechanism. The Seeds–Price or Bose–Einstein models were applied in these examples; however, most likely, one of the other models could have been used with equal success. The important point that has been made is that both detail process yield representation and a global simple yield representation are useful in a yield enhancement effort.

As previously mentioned, the second area of yield model application is for yield projection. In this area, the applicability of the simple models is strongly dependent on the process and model selected. If a yield monitor is being used to determine defect densities and can be modeled with the type accuracy exhibited in Fig. 6.11., then it has the potential for being quite accurate for product yield predictions at least for that process. Both models given in Table 6.2 by Eqs. (6.2.d) and (6.2.g) model yield as a function of area quite well. Thus, future products, for example, of larger area, can be accurately modeled in an effort to understand product viability in terms of cost and marketability. Small deviations in the process complexity (i.e., additional levels) or small dimensional changes (i.e., linear shrink) can likewise be modeled accurately provided proper defect relationships can be established. For process complexity, this means being able to ascertain reasonable estimates of the defect densities for additional levels of integration through knowledge of equipment, facility, and process sources of defects. For both process complexity and design rule changes, it means being able to accurately estimate the effects of critical area as well as defect size with respect to minimum feature size, such as through the use of Eq. (6.41). The most difficult job for simple model yield projections is in long-range technology evaluations. This, however, may not be germane to just simple models. The complex models may also have a problem in this area because there is not enough hard data to define the model parameters.

Consider the long-range technology trends and projections for both U.S. [15] and Japanese [37] semiconductor technology summarized in Table 6.3. The data in this table show that minimum feature size has been decreasing by 15–20% every year and will continue to do so into the 1990s. Die size has been increasing by about 25 mils per edge every two years and will also continue to do so. Finally, process complexity has been increasing by roughly one critical level per year. The fact that the U.S. data are now nearly four years old and have followed this table quite accurately lends to the credibility of the predicted trends. It should also be noted that the U.S. trends are based on mature technology and the Japanese trends are based on technological introduction, which accounts for the slight time offset of the two data sources. Reviewing the trends shown in this table, it is not surprising that a simple yield model cannot accurately describe such data. The simple models can, however, provide a feel for the defect density

requirements that will be necessary to meet process yields of the future even if they are not exactly precise. The results of such a study are shown in Fig. 6.13. Based on the data shown in Table 6.3, the defect density required at the minimum feature size to achieve yields of 10%, 30%, and 50% have been calculated and are shown by the isoyield curves in the figure. The effective defect densities per critical level were calculated with an inverted Bose–Einstein model [Table 6.2, Eq. (6.2.e)]:

$$D_{ol} = \frac{1}{A}[Y^{-1/n} - 1] \tag{6.43}$$

using the cited yield value, Y, with the process complexity, n, and chip area, A, taken from Table 6.3 for the appropriate year. Now, recall that these specified defect densities are at the minimum feature size for the technology at the cited time. In terms of defect size, this means $\frac{1}{2}$ the minimum feature for lithography and significantly less for interlevel layers when one considers gate oxide thicknesses are on the order of 250 Å or less today. The impact of these requirements is illustrated with respect to an achieved yield of 48% for a 60-mm^2 at 1.5-um minimum feature achieved in the authors' CMOS line in 1986. Note that this yield achievement is in excellent agreement with the trend predictions of Table 6.3. Assuming no improvement in defect density, the effective increase in defect density for cubic and square law dependencies, according to Eq. 6.41, due to technology trend are shown by the filled and open triangles, respectively, in the figure. What this means is that if 10% yield is required for product intro-

TABLE 6.3
Integrated Circuit Technology Trends

Year	Minimum feature (μm)		Die size mil/edge		Critical levels	DRAM
source	1.	2.	1.	2.	1.	2.
1978		3.0		200		64 K
1980						
1982	3.0		250		6	
		2.0		250		256 K
1984	2.0		275		7	
1986	1.5	1.3	300	300	8	1 M
1988	1.25		325		10	
		0.8		350		4 M
1990	1.0		350		12	
1992	0.8	0.5	400	430	(12)	16 M

1. Source: Morgan and Burnett [15].
2. Source: Imai and Hashimoto [37].

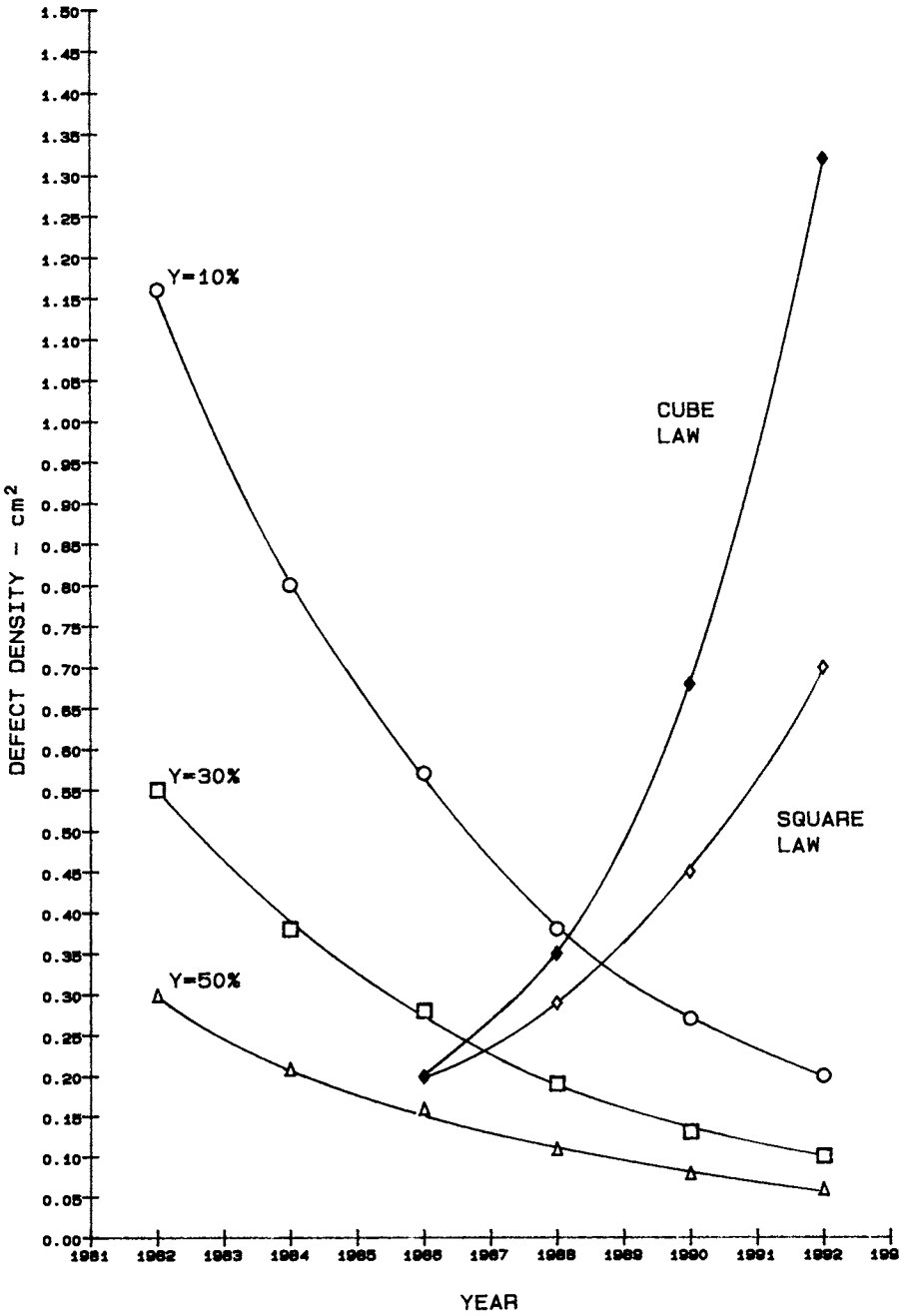


Fig. 6.13. Defect density requirements at $Y = 10\%$, 30%, and 50% based on predicted U.S. technological trends. Also shown are the defect density limits that are rapidly imposed on yield by square and cubic law defect size distributions with advancing technology and without any attendant defect reduction. (From Morgan and Burnett [15].)

duction, in early or mid 1988, depending on the power law of the defect size distribution, the facility could not meet minimum production standards. Furthermore, regardless of the power law, significant decreases in defect density will be required not only to introduce new products but to bring them to maturity (i.e., $Y \cong 50\%$) as shown by Fig. 6.14. Although the numbers may not be exact, this example does serve to illustrate the impact of technological trends on yield and demonstrates the importance of having a strong yield enhancement effort.

The discussions in this section have centered on the use of generalized process Seeds–Price or Bose–Einstein type yield models. This has been a result of what has actually worked well in our yield enhancement efforts. Other models may also be applicable and nothing other than the fact that these models performed excellently in these applications is intended. We feel that the choice of a yield model should be individual and guided by the philosophy that the best yield model is the simplest one that gives reliable results.

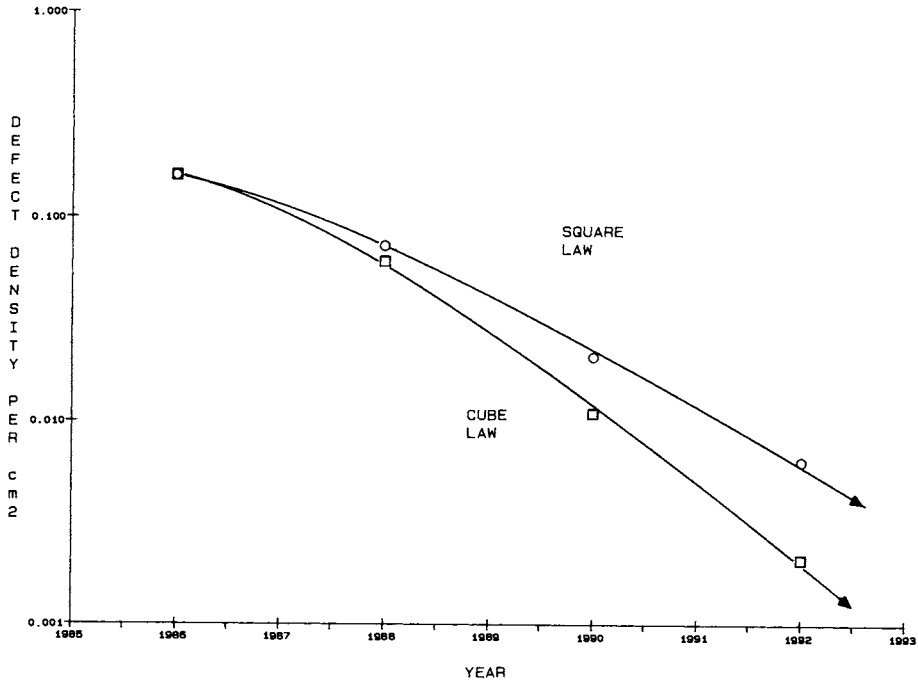


Fig. 6.14. Required defect density reduction for square and cubic size distributions to maintain 1986 yield levels with technological advancement.

II. SOURCES AND CLASSIFICATION OF YIELD LOSS

The preceding sections of this chapter have considered the basic concepts of yield. Hopefully, from these sections, the reader has become aware of the importance of yield and has at least been introduced to the fundamental issues of this important segment of semiconductor technology. Ideally, every circuit and process is designed and every facility is built with the philosophy of achieving 100% yield. Of course, in practice this is impossible and real circuit yields for VLSI range from a few good circuits per wafer to some relatively large number depending on the process, facility, and product itself. The yield enhancement effort is dedicated to achieving yields as close to the ideal as possible. Key to this endeavor is an understanding of the sources and classification of yield loss mechanisms. Throughout this chapter we have alluded to these issues and will consider them in detail in this section.

A. General Yield Classification

Previously, in Eq. (6.39), we generalized the yield model into a gross yield component and random defect limited yield component. This partitioning represents a first-order classification of yield loss mechanisms. The fact that these components could be further subdivided was alluded to when area, line, and point defects were introduced in the model development section of the chapter. Further implications of defect subclassifications were illustrated by the partitioning of gross and random defects in Fig. 6.10. From this figure, we see that both gross and random yield components can at least be split into visual and nonvisual defects. The potential for even further subdivision is indicated by the random visual defect subclasses in that figure.

As noted, the random yield is characterized by the sensitivity of the product to point defects at a specific critical level, n. Of the n critical levels, n_1 are associated with visual defects, that is, defects that can be seen with a conventional light microscope or an SEM. The remaining n_2 critical levels are associated with nonvisual random defects that are observable only by electrical test or delayering followed by delineation and visual inspection. Gross defects are chip area independent in that regardless of size a chip would be rendered nonfunctional by their presence. Because of this, they are often referred to as "nonrandom" [12] or "chip area independent" [18] defects. These defects fall in the general classifications of area or line defects. Again, as with random defects, they can be classified to first order

as visual or nonvisual defects. One difference for the visual classification is that gross defects are always large enough to be identified with a light microscope.

As a result of these classifications the general yield formulation can be rewritten with visual components (subscript V) and nonvisual (subscript N):

$$Y = Y_{GN}\, Y_{GV} \prod_{i=1}^{n_1} Y_{RV}(A_{ci}, D_i) \prod_{j=1}^{n_2} Y_{RN}(A_{cj}, D_j). \tag{6.44}$$

It is pointless to try and develop the model from a general point of view with further subclassifications than this. Processes are dynamic, evolving entities that continually respond to new limits set by technological trends and paced by new unit steps and equipment developed to achieve these limits. As a result, some of the visual and nonvisual defects in today's technologies will be rendered insignificant only to be replaced in tomorrow's technologies by new defects that are nonexistent today. We can only hope to provide a formalism for classifying defects and a knowledge base of defect sources developed on past and present process experience that will assist us in identifying, classifying, and reducing defect sources in future processes.

B. Gross Defects

Gross yield loss mechanisms are generally of the area or line type defects and are defined by the premise that they would adversely affect yield regardless of chip area. Several subclassifications of visual and nonvisual gross defects according to area or line type are summarized in Table 6.4. Reviewing the types of defects in this table may cause some controversy as to classification. For example, some workers might be tempted to categorize scratches as random defects because they can occur randomly on the wafer. Similarly, one might be tempted to classify gross particles, hair, or threads as random defects because of their airborne particulate nature and random occurrence. However, if one reminds oneself of the fundamental operating premise adopted in this chapter, of adverse yield effect regardless of die size, then the classifications of Table 6.4 can be justified. As in many areas of science and engineering, the classification of defects represents somewhat of a grey area. The only justification we can provide for this classification methodology is that experience in actual yield enhancement activities has verified the resulting yield models (e.g., Fig. 6.11).

In general, the visual types of defects summarized in Table 6.4 can be determined with optical microscopy at relatively low power (50–200×).

TABLE 6.4
Gross Defect Classification

Area type
Visual
Misalignment
Step coverage
Focus/development
Spots/stains
Gross particles
Nonvisual
Blanket implant
Substrate
Step coverage
Overetch
Layer thickness variations
Line defects
Visual
Particle streaks
Scratch
Hair
Thread
Nonvisual
Substrate
Implant

However, electrical tests of parametric or yield test structures verified by optical inspection can also be used to identify visual gross defects. An example of this is that metal 1 misalignment shows up as contact string failures in the parametric tests. This, of course, must be verified by inspection to ensure that the failure is misalignment and not some other contact failure mode. Nonvisual defects can usually be isolated by electrical tests of parametric or yield test structures. Sometimes delayering, delineation, and visual or SEM inspections are required to isolate nonvisual defects or confirm electrical test results. The fact that many of the gross defects manifest themselves as parametric deviations has led to the occasional use of the term parametric yield as synonymous with gross yield. We do not adopt this definition because, as can be seen from Table 6.4, gross yield loss mechanisms are much more than just parametric in nature.

Many of the defects given in Table 6.4 are self explanatory; however, some further comments may prove enlightening with regard to sources of these defects. Misalignment is straightforward and usually, unless extremely bad, shows up only in the most critical layers (gate, contacts, metal 1). This can be the result of run-out in the stepper or human error in set up

or adjustment. Alignment marks may become distorted or partially obliterated over the whole wafer or part of the wafer because of deposition thickness variation or other processing problems. Alignment problems may also be caused by wafer warpage. When a silicon wafer is oxidized, the top and bottom oxide layers are in compression and the interior silicon layer is in tension. This is a result of the approximate doubling in volume of the oxide relative to consumed silicon in the thermal oxidation process. If the backside SiO_2 is removed, the top surface will develop a convex bow. Over the course of a process this can lead to dimensional changes upward to 2.5 μm on 125-mm wafers with attendant alignment problems. Step coverage can be bad enough to be visually detectable but most often is nonvisual and is found with use of electrical test structures. Metal serpentines over underlying topology or contact strings are typical step coverage monitors. After electrical detection, delineation and SEM inspection of cross-sectionalized wafers are often used to further evaluate the step coverage. Focus/development problems can be the result of localized wafer warpage, resist thickness variations, or matter on the chuck. They may also be the result of marginal focus conditions or random stepper malfunctions. Improper development can lead to a resist residue between patterned sections or in contact or via holes. All of these can lead to blanket transfer problems over a whole die or several die in a localized area of the wafer. Large area spots and stains are often the results of cleaning problems or residues and manifest themselves as a multitude of problems. They can appear as adhesion problems in subsequent resist or deposited layers or result in poor contact formation because of trapped residues in contact windows. Cleaning residues can also cause large areas of increased oxidation induced stacking faults, which can result in excess leakage and subsequent device and circuit failure. Variations in deposition thickness can cause feature size changes due to over etching of lines in thin layers when the normal clearing etch is used for nominal thickness. At the gate level, this can result in increased threshold voltage, increased leakage, and potential source/drain punch-through problems. Any of these, under the right conditions, can lead to parametric circuit failure. Linewidth or feature size variations can also be the result of improper lithography conditions (i.e., exposure, focus, or development). Large area particles, although less of a problem with today's clean rooms and clean-room procedures, can result in lithography or pattern transfer problems causing gross yield loss in one or more localized die. Blanket implant problems can result from wrong implant conditions, masking errors, or handler problems within the implanter. Such implant problems manifest themselves as parametric problems in threshold voltage, contact resistance, or diffusion resistance and usually result in circuit failure.

Line defects provide similar type errors to the area defects. Particle streaks can be the result of abraded material deposited on the wafer because of transport mechanisms or they can be the result of poor cleaning. Such defects can lead to line type lithography or pattern transfer defects that appear much like scratches in visual inspections. Substrate line defects are usually the result of the propagation of defects introduced by wafer handling. Implant line defects are the result of human hair or thread dropping on the wafer and blocking the implant or the result of a scratch in the implant mask causing erroneous doping. Scratches, either in actual layers or in resist patterns that are transferred to the wafer during etch, have been by far the most common type of line defect. Significant reduction in scratch defects is expected through the increased use of automation in wafer transport, handling, and loading.

C. Random Defects

Random defects, by and large, belong to the general class of point defects and as with gross defects can be visual or nonvisual in nature. Based on the adopted criterion for separating gross and random yield loss mechanisms, the defects must be small relative to the die size. Again, we point out that there is no hard and fast rule such that one can assign a defect size as the transition from random to gross yield loss mechanisms. This comes through experience in yield enhancement activities and requires not only consideration of the defect size but also of the source of the defect. One practical guideline that has served us well is defects up to the dimension of the scribe lane width can certainly be classed as random in nature. Visual random defects are generally the result of particulates from people, equipment, materials, process, handling, etc., which alter lithographic or pattern transfer operations. The manifestation of defects such as missing or excess lithography and pattern transfer are shown schematically on a gate level layout in Fig. 6.15. Note that pattern transfer defects are easily discerned from excess lithography because the pattern is seen in relief in the defect in the transfer case. Also, the edges of excess lithography defects are usually more sharply defined than in transfer defects. Nonvisual defects are oxide/insulator defects, perturbations in local implants or etch conditions, or localized substrate defects. Many of these are caused by the same particle sources as visual defects but occur in underlying layers and thus are not detectable by visual or SEM inspection without deprocessing and delineation.

Since particles play such a major role in random defects, we show a typical distribution of particle sources [19] for a VLSI facility operating in

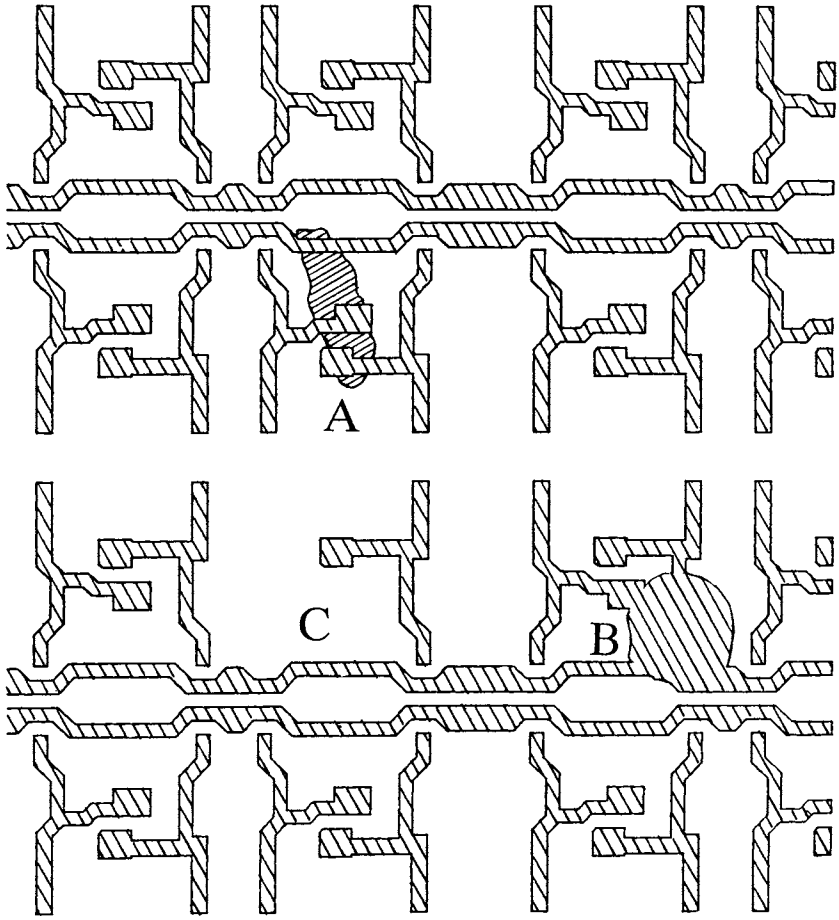


Fig. 6.15. Random defect classifications: A, pattern transfer; B, excess lithography; and C, missing lithography.

the 1.0–2.0-μm regime in Fig. 6.16. As the technology develops beyond this into the submicron range, we may well expect the proportionments in this figure to change. For example, increased automation and reduction of people in direct wafer contact operating with optimized process flows and facilities will reduce the particle sources in the facility/process, people, and handling categories. As a result, equipment particle sources will become more important, and as feature sizes decrease, chemicals and electrostatic effects will become more important.

As shown in Fig. 6.16, equipment is the major source of particulates in current VLSI and is expected to increase in significance as a defect source

in the future. Every piece of equipment in the process facility is a potential source of particulates. Spin dryers contribute particles through drive mechanisms, feedthroughs, door seals, and in loading and handling. They also can contribute significantly to electrostatic charge buildup on the wafer and subsequent particle accumulation. Resist spinners contribute particles through exhaust control, splash back, and edge lip. The resist bottle or package, dispensing pump, and lines can also be sources of particles. Ion implant equipment can contribute particles in mechanical transport, gas ports, and vacuum components. Redeposited resist chips from handling are also a large particulate source in implanters. Deposition equipment, LPCVD, PECVD (Plasma Enhanced Chemical Vapor Deposition) and metal all contribute particles through loading and handling mechanisms, as well as through spalling of sidewall buildups. Metal deposition systems contribute particles through internal moving parts and many CVD reactors contribute quartz tube particles. Dry etching systems contribute particles through autoloaders, pedestals, ionized contamination, sidewall accumulations, internal mechanisms, and backside wafer residues. Automation will certainly become more important in future technologies. Current robotics can be a large source of particulates, and it is basically a buyer beware market.

People and facility/process account for nearly equal second and third

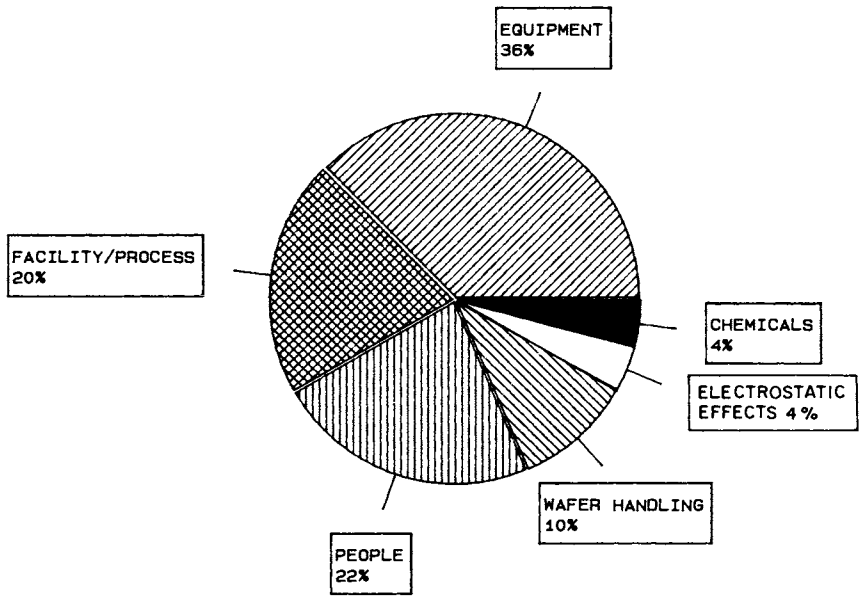


Fig. 6.16. Typical front-end facility particulate sources.

largest sources of particulates. People are the greatest single source of particulate, chemical, and biological contamination in the clean room. Particulates from people consist of skin flakes and oil, hair and hair follicles, dust, lint, bacteria from the breath, and numerous aerosols from cosmetics, deodorant, and hair sprays. The chemical composition of these particulates consists of compounds of metals (including Zn, Fe, Ti), mobile ions (K, Na), mica dust, carbon black, numerous organics, and protein (specifically keratin from shedding of the outer skin layer). These particulates range in size from 20 nm to hundreds of micrometers. The extent of this size distribution accounts for the fact that some of these (hair, large lint, threads, and dust) result in gross yield loss. Excellent reports of contamination introduced by people in the clean room have been given by Moore [20] and by Lowry *et al.* [21]. Facility/process particles are a result of the facility layout and process flow as well as general equipment, materials, and services that are not related to a specific unit step. As an example, deionized, (DI) water and piped gases would be classed in this category, whereas specific wet chemicals for cleaning procedures and gases for etches would come under chemicals. Therefore, in some respects, this represents a catch-all category, which probably accounts for its relatively large proportion. Silicon starting material defects would also be classed under the process category. Important to this category is process flow, throughput, and storage of wafers. Close to 80% of a wafer's life within the processing facility is spent waiting for actual processing; hence, cleanliness in wafer storage is imperative. As pointed out earlier, it is expected that defects in this category will be significantly reduced in the future. Reasons for this will be reduced people to wafer contact, new clean-room garments, and modular facilities of ultrahigh cleanliness or of the Standard Mechanical Interface (SMIF) [22] type isolation concept.

Handling accounts for roughly 10% of the particle sources. This is down considerably from earlier LSI when tweezer handling accounted for not only a large particulate source but also a large fraction of gross yield loss through scratches. Two of the largest sources of handling particulates are due to silicon dust generated in dump transfers [23] and redeposited resist from wafer edges. These defect sources are expected to be reduced in the future through improved automated transfer mechanisms and the development and utilization of resist edge trimming techniques [24].

In today's technology chemicals and electrostatic effects account for roughly 4% each of the particulate sources. It should be noted that electrostatic effects [25] are not really a source of particles but rather provide a mechanism for trapping particles. As feature size (and hence critical particle size) becomes smaller and steps more sharply defined, electric fields will increase in the active circuit area thereby enhancing electrostatic defect

entrapment. This will be aided by the high air velocities in the clean room as well as by equipment-related effects as previously mentioned with the spin dryer. Liquid chemicals, even those of the "ultrapure variety," have many more particles than the air in the clean room. Fortunately, most particles from liquids do not stick to the wafer [26] because

1. electrostatic and other attractive sources are weakened because the particle charge is dissipated through the liquid, and
2. the abrasive force of the liquid is relatively strong.

Unfortunately, the converse is true in air and gases. Nevertheless, the potential for liquid particulate problems cannot be overlooked. Consider the ultralow particulate chemicals (ULPCs) that have quoted particulates greater than 1.0 μm of 1 to 10 per ml [26,27]. Typical bath flow rates in a VLSI facility range from 30 m/min to 150 m/min. Assuming a flow rate of 30 m/min, the surface of a 150-mm wafer is exposed to 5.3×10^5 mL/min of solution. Based on the ULPC particle densities, this means the wafer is exposed to somewhere between 0.5 and 5 million particles per minute. Hence, it does not take a very large particulate trapping probability to create devastating problems in VLSI and beyond.

III. YIELD ENHANCEMENT METHODOLOGY

At the beginning of this chapter we discussed the impact of yield over the process life in terms of the learning curve. There it was shown that modest differences in the slope of the learning curve can basically determine the survival of a product or perhaps even a whole facility. Furthermore, it was shown that the learning curve slope that a given facility follows is determined by its yield enhancement effort. A large part of this is the ability to identify defect sources and reduce or eliminate them in a timely fashion. This requires a diligent effort during lot processing to minimize defects as well as after lot completion in failure analysis. However, this is not all; it begins much earlier at preprocess with solid concepts and methods in both process and product design. In this section, we will discuss the preprocess, in-process, and post-process aspects of yield enhancement methodology.

A. Preprocess

Yield enhancement must start the moment that a new process or product is conceived. The effort invested at preprocess can mean the difference between success and failure, not only for the product but potentially for the

whole facility. For a process, the primary emphasis in preprocess yield enhancement is developing a process that is not parametric yield limited from unit step technology that will be capable of supporting design rule targets at the time of volume manufacture. From the product point of view, the designs must be capable of achieving projected yield goals with a high confidence level. This is true not only for new products in a new process but also for new products in an existing technology as well.

Consider the development flow chart of Fig. 6.17 for an advanced process technology that is driven by system needs, as is usually the case. The process and system design phases progress down parallel paths, which as indicated in the figure are highly interactive. These culminate in a viable product at introductory yield target levels that are subsequently improved over the product life. Early system considerations establish targets for electrical and physical parameters that influence the initial process architecture. As an example, system considerations might require a high-density CMOS design, capable of 5-V latch-up free operation, based on macrocell methodology. The process technology response to these might be a

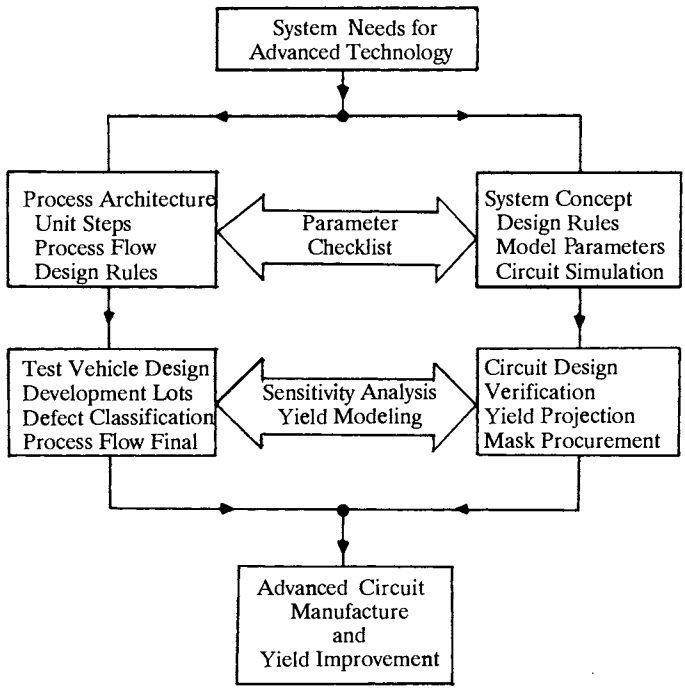


Fig. 6.17. Interactive process/design cycle for an advanced technology that is driven by system needs.

1.25-μm polysilicon gate twin-tub CMOS process with lightly doped drains, retrograde p well, epitaxial substrate, and two-level metal capability. Further system considerations provide more detailed basic rules for physical and electrical parameters, such as gate oxide thickness, desirable dimension control, threshold voltage, and leakage characteristics. The process technology responds with the unit step designs to provide these features, taking into consideration equipment capability, dimensional tolerances, reliance on proven technology, or development aspects with implications to manufacturing acceptance and cost effectiveness. The compatibility of the system requirements and unit step capabilities are evaluated with a parameter checklist. Here, such things as whether or not the electrical and physical parameters are matched to each other and self-consistent are examined. For example, is the threshold voltage versus gate oxide thickness and surface doping matched physically and electrically? Are threshold voltage and subthreshold characteristics matched and self-consistent with device dimensions and design requirements? Finally, is the unit step technology capable of supporting the design rule targets now, or will it be at the time of volume manufacture? The interactive resolution of these results in a preliminary process flow and calculated set of design rules (electrical and dimensional) that are verified through circuit simulation. The process development continues with development lots using a suitable process test vehicle to evaluate process limits, device performance, and potential yield performance. Here, the details of the process integration resulting in a finalized process flow are worked out. Initial defect classification and yield improvements through use of yield structures on the test vehicle are also made. Simultaneously the system design has proceeded down a parallel path through circuit design, design verification, yield projection, and mask procurement. Ideally, the masks and process maturity arrive together to begin circuit manufacture at yield levels as good as, or hopefully better than, introductory product requirements. As before, these are not isolated parallel developments; they interact strongly through a process sensitivity analysis and yield modeling. The process sensitivity analysis ensures that the process design and process tolerances yield electrical parameters within specification windows compliant with the circuit designs. Statistical tolerances of process parameters are used in conjunction with a process sensitivity matrix, derived from process and device simulations and modeling, to calculate components of process-induced variance of electrical parameters. The calculated standard deviations are compared with the specification windows of electrical parameters to predict the parametric process yield, If the parametric yield is too low, either the unit steps of the process can be changed to minimize the variation or the process tolerance can be changed through changes in equipment or

operating conditions, or the design may have to be modified to reduce the sensitivity to the parameter in question. In this manner, a compatible process and product insured of high parametric yield will result. The process and design phases of the development also interact through yield modeling of gross and random defects. Based on preliminary defect classifications and estimates, yield modeling is used to evaluate the impact on chip designs. This provides input for the overall chip partitioning and focuses attention on defect sources that must be improved to meet initial and long-term yield goals.

One other area of importance to preprocess yield enhancement that has not been emphasized up to this point is the development or selection of a yield vehicle. This may be the same as the test vehicle used in the process development or it may be an entirely new design. The basic purpose is to provide a vehicle that can be used for defect classification for yield monitoring and enhancement efforts throughout the process life. Several potential candidates for a yield vehicle are listed in Table 6.5. Fundamental circuits such as SRAMs and DRAMs are ideally suited to defect classification because they can be electrically mapped for failures. Once mapped the failed site can be easily correlated with physical chip location for defect detection. Gate arrays can also be used as well as custom circuits designed for testability. In general gate arrays do not stress the density as much as SRAMs or DRAMs and thus are somewhat more relaxed in defect classification. Custom circuits designed for testability are excellent for proving a design but are not nearly as useful as RAMs or gate arrays for defect isolation because exact physical circuit location is still hard to determine from the failed vectors. All of the circuits have the common problem that

TABLE 6.5
Yield Vehicle Candidates

Functional circuits
SRAM
DRAM
ASIC/Custom
Gate array
Design for testability
Yield monitor structure
TEG
Scribe lane: combined parametric and yield monitor
Chip: combined parametric, yield monitor functional circuit
Full wafer
Drop in

the process must be run to completion before the structure can be tested. Furthermore, defects must be classified by optical or SEM inspections for visual defects, delayering, and delineation followed by inspection for nonvisual defects. Yield monitors are test structures similar in nature to parametric test structures that can be electrically tested for failures. Many excellent articles have been written that discuss the design and use of such test structures [28,29]. Interleaved serpentine and comb structures over topology can be used for detecting interlevel and intralevel shorts as well as continuity of lines. Various spacings of elements can be designed into the structures to allow information on defect size to be obtained. Some problems exist with correlating defect monitor results with defect sources without visual inspections. As a result, these structures are excellent process yield monitors but less effective as yield enhancement vehicles. One advantage they do have over functional circuits is that they can be pulled from the process after any step, given a metallization, patterned, and tested, and hence can be used as in-process monitors as well as postprocess monitors. Test element group (TEG) structures are an excellent candidate for a yield vehicle because they also contain parametric structures that are excellent diagnostic tools for nonvisual defects. The TEGs can be of two configurations, scribe lane or full chip, as indicated in Table 6.5. Scribe lane TEGs contain parametric test structures and yield monitors and are designed to fit in the unused border surrounding a product chip, known as the kerf or scribe lane, that is used for separating the chips from the wafer by sawing [31]. Advantages of this configuration are that every chip has its own set of parametrics, and they are fabricated in space that otherwise would be wasted. TEGs can also be designed as a full chip that can contain parametric test structures, yield structures, and a suitable functional circuit such as a SRAM or DRAM for yield enhancement as well [32]. The chip can be used as a drop-in at specific die locations of product wafers, or it can be used solely on wafers accompanying the product wafers within a lot. Since the TEG is a full chip, it can be designed with enough parametric structures for advanced process development as well as lot tracking. In this respect, it is extremely useful in that one design can be used for process development, parametric tracking, and yield enhancement throughout the process life.

Selection of an appropriate yield vehicle is by and large a function of the type of product produced within a given facility. The TEG structures seem to certainly have advantages for all facilities in that parametric data for lot tracking as well as yield data can be obtained from the same structure. A DRAM or gate array manufacture may well choose a scribe lane TEG and use the product itself for yield enhancement as well. A custom circuit house might choose a TEG chip containing a SRAM or DRAM as a

functional circuit yield monitor. The choice of the type of functional circuit used on the TEG chip as a yield monitor is based on the preference and proficiency of a given facility and its personnel. We have had excellent success with the TEG chip containing a SRAM yield monitor for both process development and yield enhancement activities. The selection of this vehicle was influenced by the custom circuit nature of our facility and the in-house design capability and process compatibility for SRAMs as opposed to DRAMs. In the remainder of this section, the yield enhancement methodology will emphasize this approach based on these authors' experience. However, we remind the reader to look at the methodology and realize that virtually all of the yield vehicles listed in Table 6.5 could be used with equal effectiveness.

B. In-Process

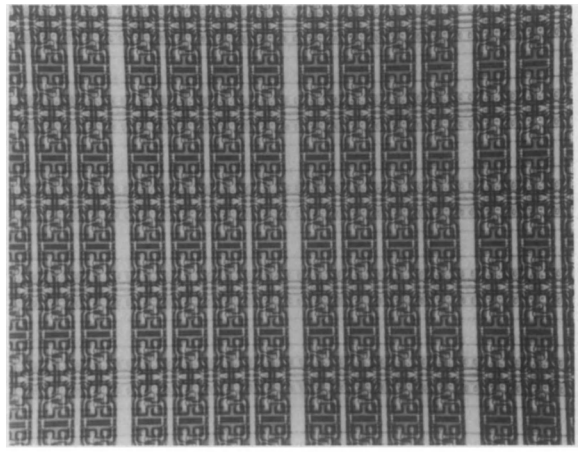
In the early days of integrated circuit manufacture, in-process yield enhancement efforts were essentially responsible for defect characterization and reduction. This was accomplished primarily by in-process inspections of lithographic patterns. As time has gone on and integrated circuit technology has progressed through the LSI and VLSI eras into the ULSI regime, in-process yield enhancement has taken on a much more global meaning. It is no longer only a pattern inspection endeavor but one of process and facility monitoring and control as well. Due to the small dimensions of today's devices and the large number of elements per chip, the emphasis of in-process inspection has shifted from one of primary defect characterization to one of catastrophic yield loss prevention. This does not mean that in-process inspections are not important, but merely that their charter has changed as mandated by the evolution of integrated circuit technology. In this section, we will review the elements of the in-process portion of yield enhancement.

An important part of the in-process yield enhancement effort is facility monitoring. This should include airborne particulate counts throughout the facility. Ideally, this should be done with a fixed, multistation, continuous monitoring system. As a minimum it can be done with a portable system, but it must be done with the facility in operation as well as at rest to ensure design targets on class levels are met and held. Another important aspect of particulate monitoring can be accomplished with strategically placed witness plates throughout the facility. These wafers are stationed for a fixed time and monitored for particulate size distribution and relative increase in number using an optical particle measurement system [33]. Such plates are excellent monitors for evaluating the effects of process flow and facility layout on particulate accumulation. Witness plates are

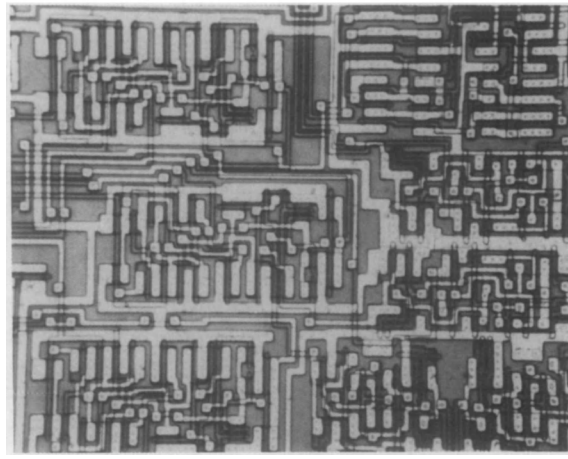
also an excellent means for qualifying new equipment for added particles per wafer pass (pwp) as well as monitoring existing equipment [34]. Various process stations (e.g., cleans, resist, and deposition) should be monitored for particulate levels. Standard qualification procedures should be developed for the various stations. As an example, our VLSI facility has developed a daily qualification procedure for photoresist that is reverified for particulate levels at the start of each shift. Cleans and resist are also monitored on a lot by lot basis with go/no-go limits established for particulate counts. Quality control procedures, monitors, and preventative maintenance schedules have also been established for all major equipment (etch, deposition, furnace, etc.) throughout the facility. Clean-room procedures for personnel (gowning, protocol, etc.) and operations (mask cleaning, maintenance schedules, etc.) must be defined in, monitored by, and strictly enforced by the in-process yield enhancement program. Quality control procedures for incoming chemicals and materials must also be established and carried out.

Process monitors play a large role in the in-process yield enhancement program and are a compliment to, if not a part of, many of the facility monitoring activities. For example, a test wafer may be cleaned and the added particles measured for a go/no-go test of the cleaning baths on a lot by lot basis. Other in-process monitors should include critical dimension, linewidth, and alignment measurements at all critical masking steps. Thin films, either grown or deposited, should be checked for integrity, thickness, and imbedded particulates with in-process monitors. Last but not least, in-process inspections round out the list of in-process monitors.

As pointed out earlier, the role of in-process inspections has shifted with the changes in integrated circuit technology. At one point in time, they provided the primary method for defect detection and classification. Today they are used primarily to detect high defect levels and provide immediate corrective feedback to avoid catastrophic yield losses. To this end they will always play an important role in integrated circuit manufacture. Woe be the facility manager that eliminates in-process inspections because they are not effective for defect classification and suffers a catastrophic yield loss that would have been avoided with their use. Therefore, in-process inspections are geared toward detecting high-level, visual, random defects and gross yield loss mechanisms. They should be done at all critical patterning and pattern transfer steps as a minimum. At our facilities, all patterns, except the most noncritical (i.e., passivation pad etch), are inspected at lithography and after pattern transfer and resist removal. Even wafers that have gone through a nonvisual pattern transfer, such as an ion implant, are inspected after resist removal for excessive particle counts. The effectiveness of the in-process inspection procedure is significantly enhanced with a repetitive circuit pattern, such as SRAM, as illustrated in Fig. 6.18. Also, in



(a)



(b)

Fig. 6.18. A, SRAM; B, 24 × 24 bit VLSI custom multiplier circuits at low magnification. The repetitive nature of the SRAM pattern aids in-process inspections.

order for the inspections to be done rapidly, they must be effective at low magnification. Defects as small as a few microns in size are easily detected at powers as low as 100–200× in the repetitive SRAM pattern shown in this figure. Reject/rework criteria for the in-process inspections are set by the product yield requirements. The number of die inspected is determined with a statistically based sampling plan based on the confidence level of defect detection for a given die size at the reject level. As an example, if the reject criteria were 1 defect per cm^2, then the number of die would be based on the area needed to ensure a 95% confidence level in defect detection at this density. Again, this suggests the usefulness of the SRAM for in-process inspections. Because of the repetitive nature of the pattern, defects are more likely to be found. Furthermore, because the die is always the same size, a fixed number of die can be established for the in-process inspections based on realistic defect density requirements for products being run in the facility. Strict procedural guidelines for defect inspections as illustrated in Table 6.6 must be developed and adhered to. Finally, it must be realized that as yield is improved through in-process and post-process yield en-

TABLE 6.6
In-Process Inspection Instructions

Wafers are inspected with photoresist pattern and again after pattern transfer with photoresist removed. Six critical levels [AA, PO, CT, M1, VIA, and M2] and four separate implant masks (*n* well, *p* well, source, and drain) are inspected.

Use an $x-y$ pattern in inspecting die.

Inspect nine die on each of nine wafers (center, $\pm X$, $\pm Y$, each quadrant/diagram).

Record the number of defects counted on each wafer on the defect chart according to type 1–7 as defined:

1. Missing/defective
2. Particulates
3. Residue
4. Redeposited resist
5. Mask defects
6. Scratches
7. Alignment

Carry out inspection at 100× dark field except for contacts/vias, which are bright field. Additional magnification is used as required to verify defect type.

Count multiple defects in a local area that have a common cause as one defect (i.e., multiple metal opens due to local photoresist lifting).

Count any defect that is larger than one-half the minimum space or line on the level being inspected. Count visible oxide tears and holes.

Count only defects on one mask level at a time (i.e., when inspecting a metal level, do not count poly defects).

Calculate defect density based on the given area/site:

X = area = total sites × area/site = ____.

Y = defects (total) = ____.

Defect density = Y/X = ____.

hancement activities the accuracy of in-process inspections will suffer [35]. With fewer detected defects in a fixed sample size, the uncertainty in defect detection from a purely statistical point of view increases. Furthermore, the inspectors begin to experience a sort of "psychological burnout" from inspecting more and more die and finding fewer and fewer defects. At this point, in order to maintain effectiveness of the in-process inspection, the chip being inspected must either be increased in size or redesigned with an increased layout density.

C. Postprocess

Postprocess yield enhancement provides failure analysis feedback to process engineering regarding the level and sources of defects so that they may be reduced or eliminated. This is accomplished by identifying and classifying yield loss mechanisms by a host of techniques including electrical and physical location of defects, optical and SEM inspections, delayering, delineation, implication from parametric data, and various analytical techniques. In this section, we will discuss the aspects of postprocess yield enhancement activities.

Rapid detection and classification of defect sources is imperative for an effective yield enhancement activity. This implies some means of electrically detecting defects with subsequent classification of the majority of these through visual confirmation or parametric test results. A capability for delayering (i.e., deprocessing) wafers to determine persistent defect sources not readily apparent from parametric test results or surface inspections is also desirable [36]. We have found the TEG chip with a SRAM to be an excellent vehicle for this purpose, and we will discuss the methodology within this framework [32]. Electrical failures in the SRAM can be easily detected by bit mapping and correlated with physical die location as indicated in Fig. 6.19. The bit map superimposed on the TEG chip in this figure shows column 5 stuck at "1" (i.e., double intensity). A stuck at "0" failure would appear as a blank (i.e., zero intensity). The tester is configured to show a one to one position correspondence with the physical circuit as shown by the arrow connecting the base of column 5 on the bit map to the base of column 5 on the physical circuit. Given the physical location of the defect, one can visually inspect the wafer in the indicated region. Of course, if the defect is nonvisual, it will not be found, however, knowing the type of fault (i.e., column stuck at "1") and its location puts bounds on the search preventing a large waste of time. Furthermore, it has been our experience that most defects are visual in nature and thus can be found by inspection of the wafer as indicated in Fig. 6.20. Here, visual

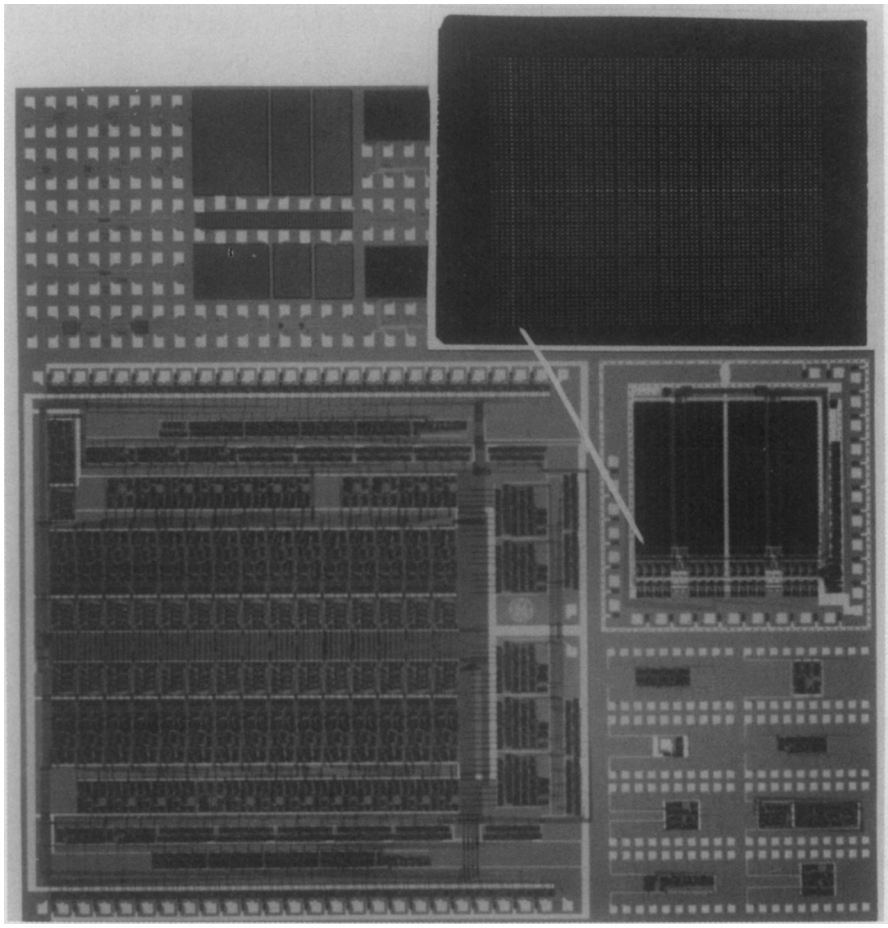


Fig. 6.19. SRAM-based TEG yield vehicle with superimposed bit map showing 1 : 1 correlation between electrical bit map failure and physical circuit layout.

inspection of column 5 has revealed a metal transfer short of submicron dimensions. Based on circuit considerations this defect causes a short between the "not bit" line and ground as indicated on the schematic shown in Fig. 6.20. This results in the bit line failing high and confirms the bit map error. In this example, we have shown the characterization of a relatively simple failure. Many of the defects that were classified in our 1.25-μm CMOS yield enhancement work were of this type. However, it should be pointed out that knowledge of the layout and operation of the

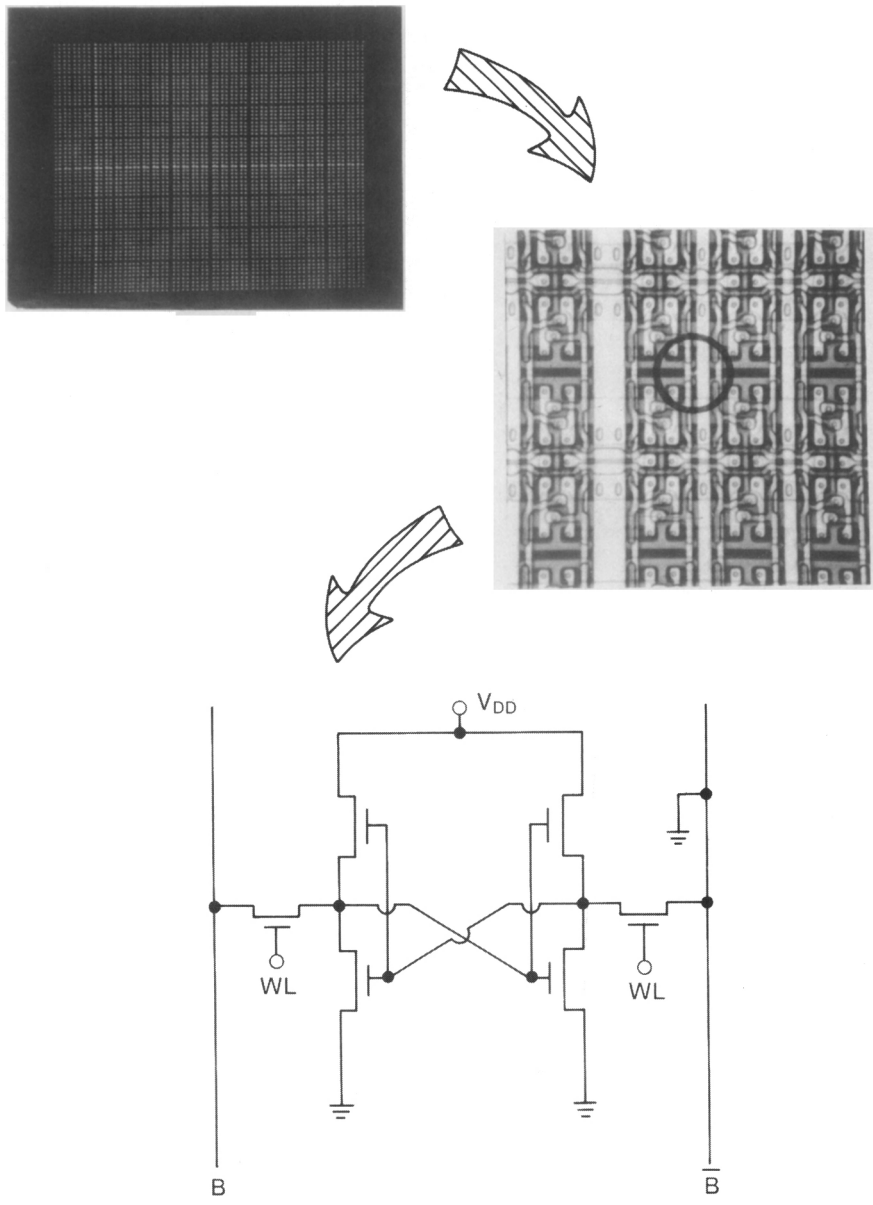


Fig. 6.20. Sequence of error bit mapping, defect detection, and circuit performance verification for complete postprocess failure analysis.

SRAM leads one to search the correct area of the circuit for much more complicated failure maps and detect, classify, and confirm defect sources. Note that this is classified as a transfer defect because the pattern is seen in relief in the defect as was illustrated previously in Fig. 6.15. Also, we stress that the defect classification is not considered complete unless the observed defect can verify the bit map failure through circuit considerations as illustrated in Fig. 6.20.

The most easily detected failures are obviously gross and random visual defects. Visual gross defects such as scratches map such that they appear as scratches on the bit map and are thus generally easily confirmed. Sometimes the SRAM gives a failure signature of all bits high or all bits low as a result of gross visual defects. An example of this type of defect is shown in Fig. 6.21. Here, massive active area shorts have resulted from a focus/resolution problem at active area patterning and electrically manifest themselves as all bits high on the bit map and are easily confirmed by visual inspection. Several examples of visual random defects that illustrate the ease of identifying and classifying them without delayering, even though they can occur at sublevels, are shown in Figs. 6.22 and 6.23. Note that many of the defects shown in these figures are extremely small (i.e. micrometer to submicrometer). The fact that they can be easily detected is a result of the electrical fingerprinting of failure position through use of the SRAM. Occasionally, visual defects cannot be completely characterized by optical inspection alone and require other surface analysis techniques such as SEM and X-ray analysis. As an example, consider the particle defect shown in Fig. 6.24, which can be seen to have caused a lithography problem even at the low power of 200× magnification. However, from visual inspection alone, we have no idea what the particle is and thus we cannot fully characterize it. SEM inspection at 3000× magnification clearly shows the physical nature of the particle to be a metal chip. X-ray analysis with the SEM shows the particle consists of Ti–W. The metal layer used in this circuit was a thin deposition of Ti–W followed by a much thicker layer of Mo done sequentially in the same pumpdown. The dotted lines on the X-ray spectra indicate the Mo; however, the Ti and W signals are too strong for the normal thin underlying Ti–W layer. Thus, we can identify the particle as Ti–W that was spalled from the deposition chamber in the later stages of the Mo deposition.

Nonvisual defects are more difficult to detect than visual defects. Gross nonvisual defects can often be detected and classified with the aid of parametric test data. An example of this type of defect classification is illustrated in Fig. 6.25. Two cases, each involving a single wafer from two different lots in which SRAM measurements showed large standby current, are indicated. In the first case, the large standby current was also accompa-

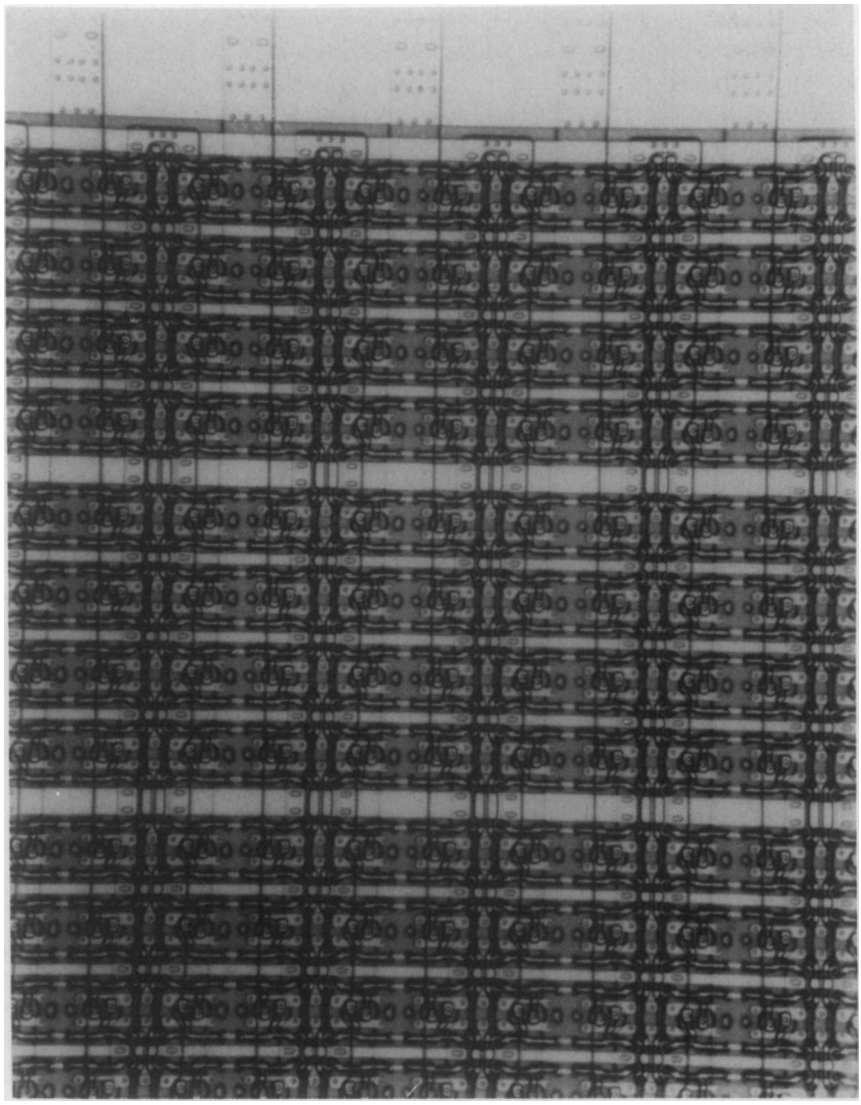


Fig. 6.21. Focus/resolution gross yield loss at active area patterning resulting in an "all bits high" failure for the SRAM yield monitor.

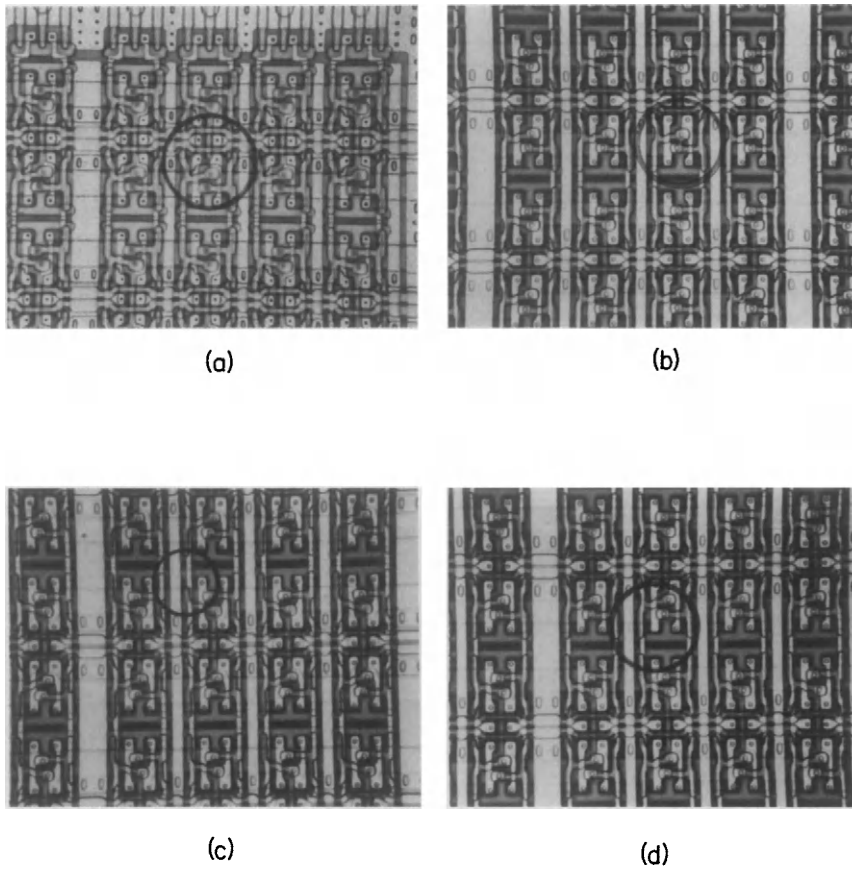


Fig. 6.22. Examples of visual random defects: A, excess active area; B, excess polysilicon (transfer); C, missing polysilicon; and D, missing contact for a 1.25-μm CMOS process.

nied by the RAM failing with all bits high. Visual inspection did not reveal any cause for this failure. Inspection of the parametric test data showed high NMOS leakage, low *n*-channel threshold voltage, and low *n*-channel punch-through voltage. Diagnostic interpretation of the parametric data suggested a missed NMOS threshold implant as the failure mechanism that was subsequently confirmed by simulation. In the second case, a large SRAM standby current was also observed, but in this case the RAM itself was functional. Examination of the parametric data indicated high PMOS leakage, high PMOS threshold voltage, and low PMOS punch-through voltage. Diagnostic interpretation of the parametric data suggested that the

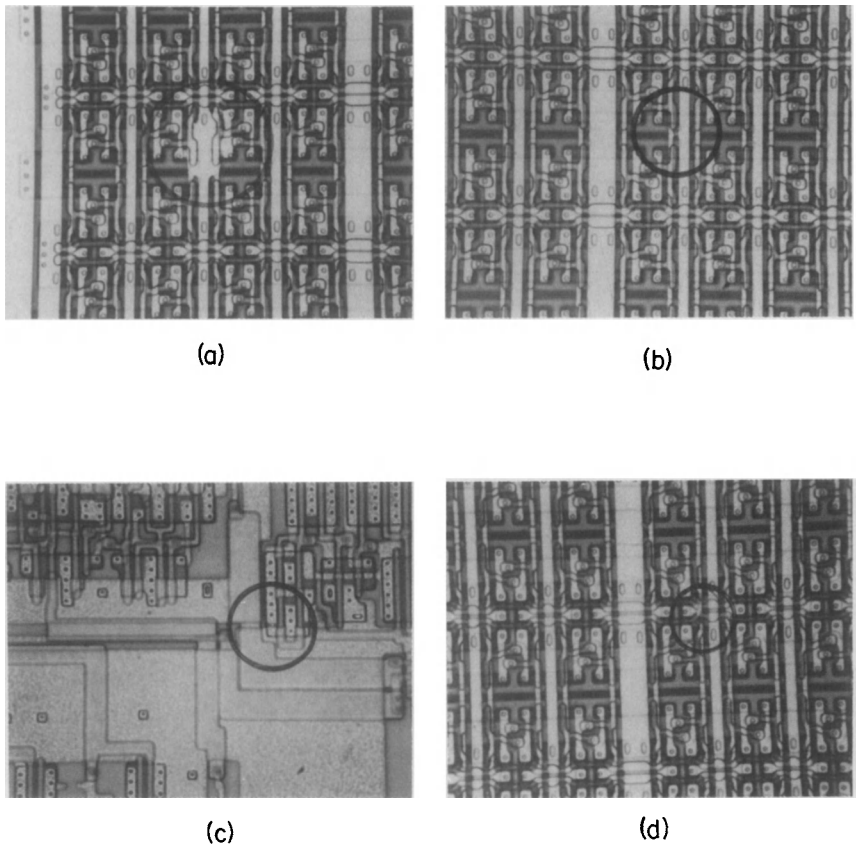


Fig. 6.23. Examples of visual random defects at metal 1: A, excess metal lithography; B,C, excess metal pattern transfer; and D, missing metal lithography defect classifications.

PMOS punch-through implant was missed, and again, this failure mechanism was confirmed by simulation. Now, the fact that the RAM did not work in case 1 and did work in case 2 but showed a large standby current in both cases can be explained from the parametric data and consideration of the circuit half-cell shown in Fig. 6.25. In case 1, the NMOS transistor leaked and this allowed the charge to be continually dumped on the bit line giving an all bits high failure. In case 2, the PMOS transistors that were leaky were blocked from the bit lines by the good NMOS transistors and thus the RAM functioned although out of specification due to the large standby current. Random nonvisual defects are the most difficult to clas-

sify. Sometimes their cause can be inferred from parametric data (high oxide defect density, interlevel shorts, more leaky devices than usual, etc.); however, definite confirmation of nonvisual SRAM failure is extremely difficult and time consuming. In general, this requires meticulous delayering, without disturbing underlayers, followed by optical or SEM inspection

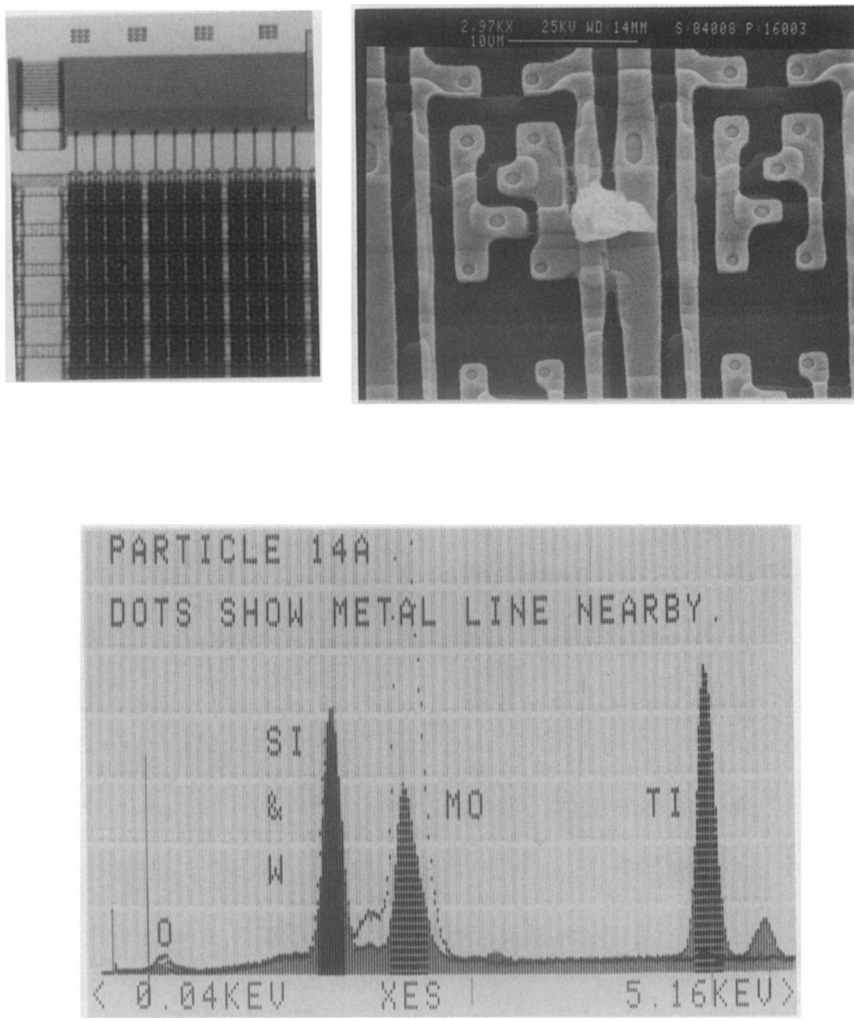
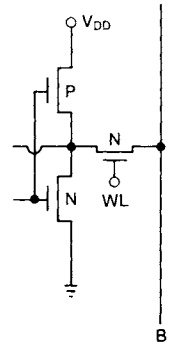


Fig. 6.24. SEM/X-ray classification of a particulate random defect.



CASE 1	CASE 2
SRAM	SRAM
Nonfunctional	Functional
Large standby current	Large standby current
PARAMETRIC	PARAMETRIC
High NMOS leakage	High PMOS leakage
Low n-channel V_t	Low p-channel V_t
NMOS punch-through	PMOS punch-through
V low	V low
DIAGNOSTIC	DIAGNOSTIC
No NMOS V_t implant	No PMOS punch-through implant
FAILURE	FAILURE
Confirmed by simulation	Confirmed by simulation

Fig. 6.25. Nonvisual defect classification from TEG vehicle SRAM and parametric data.

after each layer removal. This process of repeated delayering and inspecting is time consuming at best but must be done to fully characterize nonvisual defect sources. The best approach is intuitive defect interpretation from parametric data that suggests process fixes that are usually not confirmed until the process fix has been instituted. In other words, there is no substitute for experience and intuition when it comes to identifying nonvisual random defects. However, the SRAM–TEG does offer advantages in that it provides localized data for both parametrics and functional circuit operation, which aids the intuitive process.

The end result of postprocess yield enhancement is input to the yield model for the four basic defect classifications:

1. Random visual
2. Random nonvisual

3. Gross visual
4. Gross nonvisual

Classification of the defects in this way identifies major yield loss mechanisms so that an attack on reducing or eliminating them can be prioritized. The effectiveness of the postprocess defect classification can be measured by how well the resulting model yield compares with the experimental yield. The calculated SRAM yield based on postprocess defect modeling compared to the experimental yield over time, shown previously in Fig. 6.11, indicates excellent defect identification and classification. With good defect classification, defect sources can be identified and subsequently reduced, thereby increasing the yield to achieve a steep learning curve slope for the process development. Simultaneously the model becomes more and more accurate and useful for preprocess product and technology evaluations.

REFERENCES

1. J. A. Cunningham, Using the learning curve as a management tool, *IEEE Spectrum* 7(6), 45 (1970).
2. J. Cunningham and J. Jaffe, Insight into RAM costs aids memory system design, *Electronics* **48**(25), 101 (1975).
3. A. B. Glaser and G. E. Subak-Sharpe, "Integrated Circuit Engineering," Ch. 16, p. 771. Addison-Wesley, Reading, Massachusetts, 1979.
4. D. M. H. Walker, "Yield Simulation for Integrated Circuits, Kluwer Academic Publishers, Hingham, Massachusetts, 1987.
5. W. J. Bertram, Yield and reliability, *in VLSI Technology*, S. M. Sze, ed., Ch. 14, p. 611. McGraw Hill, New York, 1983.
6. B. T. Murphy, Cost–size optima of monolithic integrated circuits, *Proc. IEEE* **52**(11), 1537 (1964).
7. T. Okabe, M. Nagata, and S. Shimada, Analysis on yield of integrated circuits and a new expression for yield, *Electrical Engineering in Japan* **92**(6), 135 (1972).
8. J. E. Price, A new look at the yield of integrated circuits, *Proc. IEEE* **58**(7), 1290 (1970).
9. R. B. Seeds, Yield, economic, and logistic models for complex digital arrays, IEEE International Convention Record, Part 6, p. 60 (1967).
10. C. H. Stapper, Jr. Defect density distribution for LSI yield calculations, *IEEE Trans. Electron Devices* **ED-20**, 655 (1973).
11. C. H. Stapper, Jr. On a composite model to the I. C. yield problem, *IEEE J. Solid State Circuits* **SC-10, 537** (1975).
12. C. N. Alcorn, VLSI multilevel wiring defect monitor *IEEE V-MIC Conference Proceedings*, p. 252, June 1984.
13. A. V. Ferris-Prabhu, Modeling the critical area in yield forecasts *IEEE J. Solid State Circuits* **SC-20**(4), 874 (1985).
14. C. H. Stapper, Jr., Modeling of defects in integrated circuit photolithographic patterns, *IBM Journal of Research and Development* **28**(4), 461 (1984).
15. W. L. Morgan and J. R. Burnett, Concepts for world class manufacturing plants, *Semiconductor International* 7(6), 137 (1984).

16. C. H. Stapper, Jr., Modeling of integrated circuit defect sensitivities *IBM Journal of Research and Development* **27**(6), 549 (1983).
17. C. H. Stapper, Jr., F. M. Armstrong, and K. Saji, Integrated circuit yield statistics, *Proc. IEEE* **71**(4), 453 (1983).
18. T. E. Mangir, Sources of failures and yield improvements for VLSI and restructurable interconnects for RVLSI and WSI: Part 1—sources of failures and yield improvement for VLSI, *Proc. IEEE* **72**(6), 690 (1984).
19. F. Engle, Design and process criteria for modern ion implanters, *Microelectronic Manufacturing and Testing* **9**(6), 43 (1986).
20. E. W. Moore, Contamination of technological components by human dust, *Microcontamination* **3**(9), 64 (1985).
21. R. K. Lowry, J. H. Lin, G. M. Grove, and C. A. Vicroy, Analysis of human contamination pinpoints sources of IC defects *Semiconductor International* **10**(8), 73 (1987).
22. S. Gunawardena, U. Kaempf, B. Tullis, and J. Vietor, SMIF and its impact on cleanroom automation *Microcontamination* **3**(9), 55 (1985).
23. J. Burnett, Process equipment issues in leading edge practice *Microcontamination* **3**(9), 16 (1985).
24. N. Durrant and P. Jenkins, Defect density reductions utilizing wafer edge resist removal, *Proc. Kodak Microelectronics Seminar Interface '84* October 29–30, San Diego, California, 1984.
25. M. Yost and A. Steinman, Electrostatic attraction and particle control *Microcontamination* **4**(6), 18 (1986).
26. D. W. Johnson, Evaluating low particulate chemicals, *Semiconductor International* **8**(4), 168 (1985).
27. S. Powell, Chemical purity: trends in VLSI manufacturing, *Microcontamination* **3**(8), 17 (1985).
28. M. A. Mitchell, Defect test structures for characterization of VLSI technologies, *Solid State Technology* **28**(5), 207 (1985).
29. W. Lukaszek, W. Yarbrough, T. Walker, and J. Meindl, CMOS test chip design for process problem debugging and yield prediction experiments, *Solid State Technology* **29**(3), 87 (1986).
30. "Federal Standard Cleanroom and Work Station Requirements Controlled Environment," U.S. Federal Standard No. 209B, Washington, D.C., 1973.
31. C. Alcorn, D. Dworak, N. Haddad, W. Henley, and P. Nixon, Kerf test structure designs for process and device characterization, *Solid State Technology* **28**(5), 229 (1985).
32. H. G. Parks, C. E. Logan, and C. A. Fahrenz, VLSI defect detection, classification, and reduction from in-process and post-process SRAM inspections, "Semiconductor Fabrication: Technology and Metrology, ASTM STP 990, Dinesh Gupta, ed. American Society for Testing and Materials, 1988.
33. N. M. Salvo, Practical considerations in contamination source control: how to make the best of a retrofit cleanroom, *Microcontamination* **3**(8), 43 (1985).
34. B. J. Tullis, A method of measuring and specifying particle contamination by process equipment, *Microcontamination* **3**(11), 67 (1985).
35. C. H. Stapper, Jr., P. P. Castrucci, R. A. Maeder, W. E. Rowe, and R. A. Verhelst, Evolution and accomplishments of VLSI yield enhancement at IBM, *IBM Journal of Research and Development* **26**(5), 532 (1982).
36. J. D. Reyes, Deprocessing locates IC failure causes *Semiconductor International* **10**(6), 108 (1987).
37. K. Imai and T. Hashimoto, Submicron lithography in mass-production lines, *Semiconductor World* **6**(5), 77 (1987).

Index