Lynn's group,

ARPA ORDER NO.: 189-1
7P10 Information Processing

R-1956-ARPA
November 1976

Basic Limitations in Microcircuit Fabrication Technology

Ivan E. Sutherland, Carver A. Mead, and Thomas E. Everhart

A Report prepared for

DEFENSE ADVANCED RESEARCH PROJECTS AGENCY

Rand
SANTA MONICA, CA. 90406

NATIONAL SECURITY INFORMATION

Unauthorized Disclosure Subject to Criminal Sanctions.

The research described in this report was sponsored by the Defense Advanced Research Projects Agency under Contract No. DAHC15-73-C-0181.

Reports of The Rand Corporation do not necessarily reflect the opinions or policies of the sponsors of Rand research.

ARPA ORDER NO.: 189-1
7P10 Information Processing

R-1956-ARPA
November 1976

Basic Limitations in Microcircuit Fabrication Technology

Ivan E. Sutherland, Carver A. Mead, and Thomas E. Everhart

A Report prepared for

DEFENSE ADVANCED RESEARCH PROJECTS AGENCY

Rand
SANTA MONICA, CA. 90406

Published by The Rand Corporation

PREFACE

This report presents the findings of a 6-month study undertaken for the Defense Advanced Research Projects Agency to ascertain what, if any, research ARPA might sensibly conduct in integrated microcircuit technology. The authors entered upon the study through a conviction that serious international competition in this technology may appear in the next few years, and a desire to ensure for the United States as favorable an opportunity to meet this competition as research can make available. Both the ubiquitous nature of micro-electronics in defense applications and the particularly severe special defense requirements for complex, low-power, micro-miniaturized circuitry make a commanding lead in this technology very important. The authors wished to assure themselves and ARPA that the existing research programs provide adequately for the forthcoming needs of the nation. The report details some high-leverage research areas, not now receiving government or private support, where relatively small, advanced research efforts may have substantial payoff. No endorsement of the study conclusions by ARPA is implied or intended.

During the course of the study, the authors visited major laboratories having a capability for making very small circuitry. In nearly every visit, discussions with the research personnel were complete and frank. The authors believe that they have seen all the major U.S. activities aimed at producing circuits of submicron dimensions.

It was quickly realized that U.S. industry is treating the new developments in microcircuit technology as a continuation of the coupled evolution of decreasing size of circuit features and increasing complexity of logical units that has been so effective in the past. The authors therefore asked each industrial group specific questions about its aspirations for very-small-size circuits of modest complexity. The responses were disappointingly conservative. There is such a wide variety of problems to be overcome in developing a submicron circuit technology, and those experienced in the field have seen so many "rocks in the road," that relatively slow progress is the most that they can foresee. Moreover, the only developments economically justifiable for private support must maintain as high a level of complexity as possible. The authors, as a group, believe that a more direct push toward very small circuitry, albeit of modest complexity, will pay off handsomely.

The three authors of this report bring a wide range of experience to bear on the study. Carver Mead, Professor of Electrical Engineering at the California Institute of Technology, is an expert in semiconductor physics; he has contributed importantly to an understanding of the fundamental physical principles that limit how small semiconductor circuitry can be made. Thomas Everhart, Chairman of the Department of Electrical Engineering and Computer Sciences at the University of California at Berkeley, is an expert in electron microscopy; he was one of the earliest builders of fine-resolution electron beam systems. Ivan Sutherland, a member of the Rand staff when this report was prepared and now Professor of Computer Science, California Institute of Technology, is an expert in systems design; he has done much fundamental work in large systems for computer graphics. Together they have sought to understand what limitations to microcircuit fabrication are fundamental. They have tried to "orthogonalize" the tasks in order to determine separate areas where progress can be made, identifying those areas adequately covered by existing development programs and highlighting those where relatively little work is being done.

The potential for future capability revealed by this study is truly impressive. There is every reason to believe that the integrated circuit revolution has run only half its course; the change in complexity of four to five orders of magnitude that has taken place during the past 15 years appears to be only the first half of a potential eight-order-of-magnitude development. There seem to be no fundamental obstacles to 10^7 to 10^8 device integrated circuits. The authors hope that their efforts may contribute to having such circuits sooner than would otherwise be the case.

Finally, the authors are indebted to the foresights of Rand's sponsors in ARPA, to the cooperation of the many technical experts in industry who gave so generously of their time and knowledge, and to the many people at Rand and at Science Applications, Inc. without whose help their efforts would not have borne fruit.

SUMMARY AND RECOMMENDATIONS

Today's microcircuit fabrication industry is operating against two fundamental limits: the wavelength of visible light and the number of elements that can be reproduced with a single alignment. The use of electron beams, ultraviolet light, and X-rays makes fabrication of submicron geometry devices possible. On the other hand, because no significant improvements in the number of devices reproduced per alignment are anticipated, substantial changes in the patterning processes are likely.

In spite of the revolutionary nature of the changes in fabrication and design methods imposed by submicron geometries, U.S. industry appears to be treating these changes as further incremental progress. There seems to be little evidence of work aimed at quickly reaching the fundamental limits to device size imposed by physical theory. More effort needs to be devoted to improving the organization of circuitry to provide the most computation per unit area of circuit. Unless positive steps are taken, the existing U.S. investment in today's fabrication methods may be made obsolete by the new fabrication technologies, producing less vigorous competition domestically, and placing the United States in a disadvantageous position in defense and international trade.

Although at first glance it appears that the microcircuit fabrication technology has adequate funding from private sources and ample economic justification to ensure continued private funding, the existing research efforts are aimed at a continuing gradual decrease in feature size and a corresponding gradual increase in performance, coupled tightly to ever-increasing complexity levels. The following four important activities, not now covered by private funds, should be considered for future funding:

(1) *Efforts aimed at making very small devices.* Such efforts would set aside for the time being the push toward more complicated devices and focus instead on making quite simple circuits with the smallest possible feature sizes. Such efforts would not only serve to verify the limits to transistor size predicted theoretically, but also

serve as a test bed for the fabrication and electronic design techniques required for these small dimensions.

(2) *Efforts aimed at measuring the limits of dimensional stability of silicon substrates and mask materials.* Such efforts would require very precise dimensional measurements over silicon wafers and mask materials before and after various processing steps. If the values of anomalous dimensional distortions were known, they would serve as an important input to the designers of replication processes. If processing steps that minimize distortions can be identified, they might form the basis of further improvements in replication precision.

(3) *Efforts aimed at predicting the optimum feature size, die size, and wafer size, given the constraints of the newly evolving technology.* It is apparent that the fundamental limits to pattern replication precision provided by the dimensional stability of silicon have been or soon will be reached. Further decrease in feature size will require multiple replication steps on each wafer, thus making wafer size independent of the pattern replication steps and presenting new freedoms and new difficulties in the manufacturing processes. The trend toward larger wafers has been driven by a desire to reduce the unit cost of handling, as has the drive to maintain a single alignment step per wafer. In a technology that can no longer satisfy both of these requirements, what is the most effective compromise to make? To what extent is this choice dictated by our existing capital investment in large wafers? If one were starting anew, as our international competitors are, what choices would one prefer to make?

(4) *Efforts aimed at understanding the system design implications of very-large-scale integrated circuits.* Indications are that great benefits may be obtained by improving the arrangement of memory and processing power implemented in the more complex circuits that will be available in the near future. Questions that need to be answered are: How should computations be organized so as to obtain maximum performance with minimum silicon area? What advantages can be gained by making "smart" memories that can compute as well as store? How can complex machines be configured to minimize the software burden on their users? How are organizations of 10^5 or 10^6 gates different in kind from today's

CONTENTS

I. INTRODUCTION

During the past decade the integrated microcircuit has been brought from a laboratory dream into a production reality. The development of this technology came about as a series of incremental improvements stimulated by active competition among a relatively large number of firms, and an eager marketplace. It is thus not surprising that a remarkable technology has emerged and that it has grown in an incremental fashion.

Today's microcircuit capability has envolved out of incremental improvements to the existing technology. But it is possible that the technology may soon reach a cul-de-sac, i.e., that further incremental steps will not lead to further marked improvements in economy or performance, even though such improvements are possible by other means. If this is so, and there is abundant evidence to support this belief, entirely new processes must be implemented, processes that may be very difficult for a firm competing in today's market to adopt. Just as the economically outmoded World War II steelmaking technology still persists in the United States today because newer, more economical methods are not readily accessible to an industry already heavily capitalized, so our current investment in particular fabrication methods may prove a detriment as circuit technology progresses to submicron dimensions.

The evolution of microcircuit fabrication has reached or almost reached two fundamental limitations. In size, the technology is rapidly approaching the limitations imposed by the wavelength of visible light. Five-micron circuit dimensions are used in routine production; masks for such circuits contain features only 10 wavelengths of light in size. In precision, the technology is rapidly approaching or possibly exceeding the dimensional stability of the silicon substrate. Precision mask alignments on the order of 1-micron accuracy are being used over 4-in. (10-cm) wafers, a precision of one part in 10^5.

The capital equipment required for further improvement in the technology is different in kind from that in use today. To overcome the size limitations imposed by the wavelength of light, an impressive collection of electron beam pattern-generation machines has been developed. These devices can make remarkably accurate and remarkably detailed patterns having features ranging down to the submicron level. To use these patterns efficiently, an array of pattern replication systems has been built by using X-rays, ultraviolet radiation, and electron imaging systems. These devices are very different from those used in current production.

Yet in spite of the need for entirely new production methods for circuits of finer dimensions, U.S. industry generally appears to persist in incremental development. Having removed the wavelength of light as a barrier to further minification, industry might be expected to push rapidly toward the 0.1-micron feature sizes predicted as the absolute minimum on the basis of solid-state theory. At present, however, there does not seem to be any effort to explore the physical limits of small size independent of device count (the number of components per circuit). Incremental improvement in feature size while maintaining or increasing device count is an appropriate route for industry to take, considering that industrial developments are driven by competitive economic forces and that too bold a technological step may prove economically fatal. But the knowledge gained by making devices with very small feature dimensions can be put to good use both in guiding our aspirations and in rationalizing our capital investments, while postponing, if necessary, the difficulties introduced by making complex circuits out of such devices.

The microcircuit industry is now at a turning point. New methods will be implemented to obtain smaller dimensions. New pattern generation and pattern replication equipment will be put into service. Basic choices presently being made will affect the entire industry for the next 20 years, and some of the relatively inexpensive research objectives undertaken now may have considerable future leverage. Not only will improvements in our national capability to produce microcircuits meet specific defense requirements for smaller, faster, and lower power circuits than are now available, they will also provide a better base for all defense electronics. This report outlines the fundamental principles that should guide the choice of such research topics.

II. IMPROVEMENTS IN INTEGRATED CIRCUITRY

Improvements in integrated circuit technology can be separated into two basic types: improvements in the devices themselves, i.e., functional improvements such as the speed of a memory; and improvements in the cost of the devices, i.e., fabrication improvements such as projection wafer exposure. The effects of these improvements on six important factors are summarized in Table 1 and are described in the following paragraphs. It is essential to consider these effects separately. Because many of them seem to interact, there has been a widespread tendency to confuse their separate implications.

FUNCTIONAL IMPROVEMENTS

Integrated circuits have been improved by minification of their features. Reduction in the size of circuit features not only permits more circuitry per unit chip area, but also improves the speed of the devices, particularly for metal oxide semiconductor circuits (MOS). Reductions in size have usually been accompanied by increases in the number of circuit elements connected together or fabricated together. This has often led to confusion between minification and complexity, a confusion to be avoided here.

Another improvement that has been made is in the size of chips that can be manufactured with adequate yield. These improvements have impact on both function and fabrication, as we shall see shortly. Working circuit chips have steadily become larger over the years, providing more area in which to pack the components and bonding pads and more power-dissipation capability. Larger chips also increase the lengths of the longest wiring runs, introducing both time delays and the need for more careful electrical design to prevent excessive voltage drops. Improvements in the size of chips derive from improvements in the defect density of the materials used and in the number of defects introduced in pattern replication. Functionally, chip size is important because it permits more logic capability without the need for off-chip interconnections.

Table 1

FUNCTIONAL EFFECTS AND FABRICATION LIMITATIONS RESULTING FROM IMPROVEMENTS IN INTEGRATED CIRCUITRY

Factor	Functional Effects	Fabrication Limitations	Recommendations
1. *Feature Size:* Dimensions of circuit elements and spaces between them.	Smaller → faster; changes voltage and impedance levels.	Limited by fabrication process and/or alignment precision.	Explore the limits permitted by semiconductor physics.
2. *Chip Size:* Overall dimensions of a complete circuit.	Affects permissible power dissipation, number of bonding pads, and length of longest conductors.	Now limited by yield; relation to feature size unknown.	Industry will do it.
3. *Component Count:* Number of active devices on a chip.	Device and circuit improvements have made component count go up faster than accounted for by feature size and chip size alone.		Industry will do it.
4. *Replication Precision:* Relative stability of each element of a pattern to any other.		Limited by dimensional stability--sufficient for a whole wafer at present dimensions, i.e., 1 μm in 10 cm or 10^{-5}.	Understand the fundamental limitations imposed by anomalous deformation of semiconductor materials.
5. *System Capability or Functional Complexity:* Measure of compute power of a device.	For a given component count, compute power may be greatly improved by system design cleverness; for simple memory this has already been done.		Explore new organizations; understand fundamental limits imposed by wiring geometry, device performance, and speed of light.
6. *Wafer Size:* Size of substrate processed as a unit.		Was equal to replication size. Will become just a handling consideration.	Understand the optimum choice of *feature size*, *chip size*, and *wafer size* given the constraints of the submicron manufacturing technologies.

A third area of improvement is in circuits and devices. The component count available in integrated circuits has been increased over and above the level accounted for by the improvements in chip area and feature size. We have become more clever in making devices and circuits and in packing them together so as to get more devices per unit silicon area. According to Gordon Moore (see Appendix B) these improvements in component count have been more significant in the past decade than improvements in either feature size or chip area. Moore predicts, however, that this source of improvement is nearly exhausted.

Functional improvements have also been made in circuit organization and machine architecture. An integrated circuit is intended to perform a certain function and its user cares only that the function be performed faithfully and that the device be as inexpensive as possible. With circuits reaching component counts of over 10^4, and soon to reach from 10 to 100 times that number, it is not surprising that new ways of organizing circuits are being found that will function equally well with fewer components, or that will provide much more function per component. Even more important is the measure of logical function per unit circuit area, since in most circuits most of the space is taken up by wires rather than by components. As component count increases, more wires are required and these wires, on average, must be longer, so that more of the circuit area has to be devoted to wiring unless great care is taken in organization.

Historically, much of the progress in circuit organization has been made by using regular implementations to implement complex functions, e.g., read-only memories for implementing multiplication, and the use of serial rather than parallel arithmetic. The potential for future gain in computing power through organization is very high for two reasons. First, since most of the logic elements in traditional systems do nothing most of the time, there is enormous room for improvement in their duty cycle. Better organization can reduce wire length or introduce latches to lessen the uncertainty in the arrival times of data and thus permit data to be transmitted at a faster rate, more nearly utilizing the full speed potential of the logic elements involved.

Second, the greatly reduced cost of logical circuitry afforded by the integrated circuit makes it economical to duplicate computation functions and deploy them geometrically close to the data elements on which they operate, thus avoiding the expensive and slow wires that are traditionally placed between memory and computing elements. Today's computers consist of a memory and a computing element separated by the barrier imposed by a memory bus; better organization should be able to eliminate this barrier.

FABRICATION IMPROVEMENTS

One area in which improvements in integrated circuit fabrication technology have been made is in feature size. Improvements in the absolute resolution of printing processes and in the resolution of fabrication processes have made it possible to reproduce smaller and smaller features reliably in the circuits. The features now being used by the industry are on the order of 5 microns in size, rapidly approaching the limits imposed by the 0.5-micron wavelength of visible light. Further improvement in feature size will depend on new fabrication processes. The current efforts aimed at using electron beams, ultraviolet radiation, and X-rays are a direct response to this need.

As mentioned above, improvements have also been made in the size of the circuit chips that can be produced with adequate yield. Defect densities in substrate materials, in masks, and in the replication processes have steadily decreased, improving the yield for a given chip size or, if yield must be held constant, permitting larger chip areas.

How the relationship between chip size and yield depends on feature size, particularly for very small features, is not known, and projections of future functional capability must be suspect until we know more about the defect mechanisms for submicron devices. We would like to believe that chip area can be held constant while feature dimensions are decreased dramatically. If pattern defects are mostly due to relatively large particles of dust or scratches, this will be possible. If, however, submicron devices are subject to a whole new set of defect mechanisms, the development of more complex circuits may be delayed.

Important improvements have been made in the precision with which a pattern can be replicated. Pattern replication precision should be treated in dimensionless terms, since a pattern remains the same when scaled to different sizes. The precision required for pattern replication is, of course, directly related to the number of circuit elements replicated; more precision means that more circuit elements can be "printed" at each replication step. Today's equipment provides precisions on the order of one part in 10^5. Considering that the coefficient of expansion of common metals and silicon is on the order of one part in 10^5 per degree, these precisions imply temperature compensation, careful control of mechanical loading, and very careful alignment. The precision of the existing microcircuit pattern replication capability exceeds anything available in the photographic or publication industries by an order of magnitude.

Because many circuits may be replicated by a single step, improvements in replication precision affect the economics of production and not the characteristics of the circuits themselves. Replication has commonly been done on a full wafer basis, and so improvements in precision have often been accompanied by an increase in wafer size or a decrease in feature size. It is important to identify replication precision improvements as being separate from other types of fabrication improvement, not only because the newly evolving short wavelength replication methods have different precisions from those to which we are accustomed, but also because today's precisions are approaching the fundamental limitations imposed by the dimensional stability of silicon. Thus further decreases in feature size may require that the total area of the pattern replicated in a single step be decreased so that the precision required in the replication step will not exceed one part in 10^5. We may well have reached the point where the complexity of patterns to be replicated (but not the complexity of the circuit chips) is approaching maximum. Replication patterns should begin to decrease in size with decreasing feature size, whereas until now they have been increasing in size in spite of decreases in feature size.

Wafer size is another area in which improvements have been made.

The size of the substrate wafer used as a production unit in integrated circuit fabrication has been steadily increasing over the past decade, making it possible to produce more circuits per unit of labor because more circuits can be placed on a single wafer. As long as pattern replication accuracy keeps pace with the requirements imposed by decreasing feature size and increasing wafer size, a single replication exposure per wafer can be used. However, further decrease in feature size or increase in wafer size will necessitate multiple replication steps per wafer, a prospect that may appreciably change the economics of using large wafers.

III. FUNDING THE FACTORS

Having orthogonalized those factors that lead to improvements in semiconductor circuitry, we can now identify the major efforts currently underway and recommend support for activities not adequately covered. In the following discussion, we will elaborate on the four recommendations summarized in Table 1 regarding the efforts that should be directed toward factors 1, 4, 5, and 6 in that table. We believe that the existing private funding is not and will not be adequate to accomplish the work that needs to be done in these areas.

FEATURE SIZE (FACTOR 1)

Our first recommendation is that the limits of feature size (factor 1) permitted by semiconductor physics be explored. There is much to be learned about the fabrication processes, defect mechanisms, and performance of such small devices, independent of the difficulties introduced by making complex chips. It may be, for example, that a whole new set of defect mechanisms exists for very small devices that would cause us to alter, dramatically, our expectations of chip yield. Then there is the possibility that devices with 0.1-micron features may prove sufficently faster than larger devices to warrant their production even in simple circuits. Finally, simple techniques for making very small devices may be found that capitalize on the small size and are not applicable to larger devices. It is certain that experience with very small devices will serve to verify theoretical predictions, and experience in making them will provide valuable insights into the course that our minification efforts should take. For example, subthreshold currents become larger as size is decreased and probably will make dynamic devices with very small features unusable at room temperature.

Early experience with small circuits helps us to avoid two difficulties that could be encountered with a continuing gradual decrease in feature size. Very small circuits will require voltage levels much lower than those now in use. With some agreement on what these ultimate standards might be, we may be able to avoid a series of standards changes. As will be shown later in this report, in order to make very small circuits with electron beams, it will be necessary to use electron-sensitive "resists" similar to those now available; resists that are more sensitive to electron radiation are relevant only to devices with intermediate-size features. Experience in the manufacture of very small circuits may obviate the need for developing more sensitive resists. It is possible that from efforts aimed at making small devices, albeit in small numbers, there will emerge a complexity at the new size scale that will leapfrog the combined efforts to decrease size and increase complexity now prevalent in industry.

Devices with very small features are more important for defense applications than for commercial exploitation. In many defense applications, size, weight, and power-consumption limitations are severe, so that very small devices are essential. Also, very small devices have the potential of providing an enormous amount of compute power, not only because of the numbers of them that might be interconnected but also because of their high speed. This power will be vital to a number of signal-processing applications important to defense needs. Moreover, privately funded efforts are not trying to develop devices as small as are theoretically possible because of their economic need to maintain complexity and their inability to bear an additional development burden. If new developments are to result in very-small-size devices, new funding will be required.

CHIP SIZE (FACTOR 2)

There is adequate private investment for improving chip size (factor 2). The economic reasons for increasing chip size are so obvious that every semiconductor house in the country is continually trying to improve yields and thus make larger and larger chip areas available. Theoretical efforts aimed at gaining a fundamental understanding of the ingredients that affect yield are already in progress under ARPA sponsorship at the University of Florida. There seems to be no justification for recommending additional research efforts in this area.

COMPONENT COUNT (FACTOR 3)

The benefits of improving component count (factor 3) are readily apparent, and the semiconductor industry is already considerably motivated by the competitive economics in this field. The relatively rapid development of commercial semiconductor memory technology, for example, is the direct result of improvements in component count and chip size.

REPLICATION PRECISION (FACTOR 4)

Our second recommendation is that research efforts be aimed at a better understanding of the fundamental limitations to replication precision (factor 4) imposed by the properties of semiconductor materials. Replication precision, as separate from feature size, chip size, etc., seems to be poorly understood by the industry. How much distortion is experienced by silicon as it is passed through the severe environments required for semiconductor processing is not well known, nor is there a clear understanding of the factors that influence these distortions. A program of research aimed at developing a basic understanding of these mechanisms would be invaluable in guiding the development of the industry. Although the beginnings of such efforts are evident at several industrial laboratories, the approach taken in each case is quite ad hoc, merely being aimed at solving the problem well enough to take care of the complexity of the replications now being contemplated. We believe that a solid national understanding of this problem will be valuable.

SYSTEM CAPABILITY (FACTOR 5)

Our third recommendation is for research leading to an understanding of the organizational factors that influence system capability (factor 5). While it is widely recognized that wiring dominates the cost of computing equipment today, there is little theory on which to base designs. The logic minimization theory that is available minimizes relay points or gates, not wires. The geometric and topological problems introduced by the need for wires to pass one another are formidable and present a major design obstacle to obtaining circuits that implement complex logical functions. The traditional methods of organizing computing machines were developed in a technology in which the separation between memory and logical processing was clear, a separation no longer necessary or even desirable. Traditional design methods have treated logical functions and wiring geometry as independent of one another; this approach can be very costly in microcircuit technology where wiring dominates the cost of logic.

Organizational factors also influence our ability to make new designs easily. Traditional electronic packaging methods have served not only to house elementary electronic functions, but also as logical separations in the design process. The designer of a "component" packages that component not only physically by providing a housing for it, but also logically by providing a functional description that states its terminal behavior and shows typical applications. Integrated circuit technology has reached a level of complexity where such functional descriptions are often too complicated to be fully useful. In effect, by placing more and more logic on a single circuit chip we are eliminating the natural design separation points previously provided by the package. We will have to replace these points with arbitrarily chosen divisions of the logic design process for the integrated circuit itself. The division points will be much like those used in another homogeneous logical medium: software. We need to learn how to choose such divisions wisely so that the generality of the functions provided by a designer on one side of an arbitrary division point will be available to a designer using his design, but will not overwhelm him; such design interfaces need to be both simple and general.

The influence of organization on performance is in some sense an extension of the "device and circuit cleverness" ideas put forth by Gordon Moore (see Appendix B). According to Moore, much of the increase in component count achieved in the past decade has resulted from our growing cleverness in electronic circuit design and in fitting components together. He points out, however, that circuit and device cleverness is about exhausted as a source of improvement in component count. While this is undoubtedly true, we believe that compute power, not component count, is the proper measure of value. We maintain that at a higher level of organization the same kinds of cleverness that resulted in higher component counts will enable us to get more compute power per component, particularly in systems in which there are many thousands of components and, more important, many thousands of wires.

The need for new organizations of information-processing equipment is overwhelming in almost all areas of defense. Literally tons of data go unprocessed each day because the means for processing them are inadequate. Machines able to interpret pictures, to search large data bases quickly for "interesting" coincidences, to extract signals from noise in radar, sonar, and other sensors, to perform target identification, and to model complex military and economic processes are invaluable to defense. Organizational innovations such as commingling computation and memory functions may serve to bring our computation capability much nearer to the as-yet-poorly-understood limits imposed by the size of our devices, their switching speed, and the speed of light.

We are recommending, therefore, that ARPA initiate a research program to determine how computing machines can best be organized, based on the implications of today's semiconductor technology. Such research should lead to the development of new kinds of organizations and new theories of organization that will meet the level of complexity demanded by future semiconductor devices.

WAFER SIZE (FACTOR 6)

Our fourth recommendation is for a research effort that will result in a better understanding of the optimum choice of feature size, chip size, and wafer size (factor 6), given the constraints of the newly emerging manufacturing technologies. We are concerned that the choices being made by industry in these areas are driven primarily by historical factors that may no longer be relevant. It will be very difficult, for example, to retreat from the 4-in. wafer. However, a careful examination of the choices now being made by industry may have a major effect on guiding the industrial developments in the next decade. Such a study would probably be a one-time effort, rather than a continuing project.

IV. A SIGNAL-TO-NOISE RATIO VIEW OF INTEGRATED CIRCUIT FABRICATION

A signal-processing view of the processes involved in fabricating an integrated circuit has much to offer in terms of understanding the fundamental limits that apply to this technology. Such a view is suggested naturally by the image-processing steps involved in the photolithography. But a signal-processing view is also useful in describing the various dimensional stability limitations of the materials involved.

PIXEL NOISE

The patterns used to fabricate integrated circuits are all composed of purely black and purely white areas. Such images can be divided by a raster of suitable size into a large number of square picture elements, or "pixels," each of which can be described with one binary bit of information. We therefore think of such a pattern as being a very large number of bits, even though its repetitive nature and its circuit properties make it possible to describe it more compactly in other forms. We will call this the pixel view of pattern replication.

The pixel view is useful for thinking about spot defects in the replication process. Spot defects can be thought of as noise imposed on the information contained in the pixel pattern, just as "snow" on a television picture is noise in the video signal. Given a certain level of noise in a pattern replication process, it is very probable that the circuit involved will fail to work. One could produce a monte carlo simulation that would predict the tolerance of integrated circuits to this kind of noise, provided the circuit dimensions and the statistical properties of the noise were known.

Unfortunately, predicting failure rates on the basis of pattern noise is difficult because the causes of this kind of noise are usually mechanical. Plates get scratched, a mote of dust or a hair mars the pattern, or a crystal defect occurs in the substrate. The statistical properties of these kinds of "noise"--defects that tend to be long and thin--are difficult to describe mathematically, and are quite unrelated to the kinds of Gaussian noise with which information theory is most able to deal. On the other hand, recognizing pattern defects as a kind of raster noise is a useful way to think about the problem.

DIMENSIONAL NOISE

Two kinds of dimensional noise can occur in the pattern replication process. The first is introduced by systematic or random dimensional distortions of the material bearing or receiving a pattern. If one accurately measures the distance between two identifiable points on a silicon substrate as it is passed through various processing steps, a statistical variation in the measured distance will be discerned, even when compensations are made for linear expansion with temperature. These distortions form one kind of dimensional noise, which limits the resolution of the patterns that can be replicated.

A second kind of dimensional noise is introduced by inaccuracies in the alignment of the pattern being replicated. This noise limits the size of the elements that can be reproduced. As the area of the patterns being replicated increases, accurate alignment over the entire pattern area becomes more difficult because of the systematic alignment errors and the distortions in the patterns themselves. Ultimately, both the alignment errors and the pattern distortions limit the number of picture elements that can be reproduced reliably at one step. Because these errors affect the dimensions of various parts of the pattern, we will call this the dimensional view of pattern replication.

The dimensional signal-to-noise ratios required for integrated circuit processing are so high that great care must be taken in the design of equipment to generate and replicate patterns. Two developments worthy of note circumvent the high signal-to-noise ratio requirements by dividing the dimensional accuracy problem into two separate parts. In the EBES system built by Bell Telephone Laboratories, the dimensional accuracy is obtained in part by a moving mechanical stage whose position is measured very accurately by a laser interferometer, and in part by direct deflection of the electron beam. In the fly's eye lens CRT systems, such as those used by General Electric in their BEAMOS memory and by Eiichi Goto (Japan) to generate precision artwork, there is a separation between coarse and fine deflection that permits the burden of the overall precision to be shared by two less-precise processes, or by processes that are precise in different ways.

Similarly, the process of producing whole wafer artwork is commonly divided into two steps: reticle generation followed by step and repeat. Reticle generation is a complex process at modest precision; step and repeat is a simple process at high precision. Again, the self-alignment mechanisms used by some of the newer electron optical generation and replication systems make the overall precision problem more manageable. However, in considering any pattern generation or replication process with an overall precision of one part in 10^5, it is essential to understand clearly how that precision is obtained, for one can be sure that it will always be "with difficulty."

In current, integrated circuit fabrication technology, dimensional signal-to-noise ratios on the order of 100 dB are in regular use. This is a remarkable precision for a mechanical process at any scale, considering the dimensional signal-to-noise ratios for other technologies shown in Table 2.

Table 2

DIMENSIONAL SIGNAL-TO-NOISE RATIOS OF VARIOUS INDUSTRIAL PROCESSES

Process	Accuracy	Signal-to-Noise Ratio
Carpenter	1/8 in. over 10 ft	60 db = 1 in 10^3
Conventional color printing	200 screen over 20 in.	72 db = 1 in 4 × 10^3
Machine shop	0.002 in. over 10 in.	74 db = 1 in 5 × 10^3
Automobile pistons	0.0002 in. over 2 in.	80 db = 1 in 10^4
Surveying manual means	1/10 in. over 100 ft	80 db = 1 in 10^4
Navigation	1 mi over 10,000 mi	80 db = 1 in 10^4
Routine optical components	Fraction wavelength over several centimeters	100 db = 1 in 10^5
Step-and-repeat camera	1 μm over 10 cm	100 db = 1 in 10^5
Integrated circuit fabrication	1 μm over 10 cm	100 dB = 1 in 10^5
Special optical components	1/20 wavelength over many inches	126 dB = 1 in 2 × 10^6
Laser interferometer	0.16 μm over 60 m	126 dB = 1 in 2 × 10^6 (frequency stability limit)
Speed of light measurement	0.33 ppm	130 dB = 1 in 3 × 10^6
Frequency counter HP5345	---	160 dB = 1 in 10^8
Time measurement cesium beam standard	---	220 dB = 1 in 10^{11}

V. PATTERN REPLICATION AND GENERATION

An integrated circuit is essentially a set of patterns of impurities, oxide layers, and conductors laid down on a semiconductor substrate. These patterns are usually placed on the circuit by pattern replication processes from masks. The masks are themselves replicas of some original pattern that was recorded as opaque and transparent areas by a pattern generation process from a description of the pattern originally conceived by a designer. Usually, pattern generation is done by a digitally controlled recording device. In this report pattern generation refers only to the process of first recording the pattern in physical form from some abstract description of it; we will not concern ourselves with how the description was obtained. Pattern replication refers to the process of copying a pattern from one physical form to another by optical, chemical, physical, electronic, or mechanical means.

PATTERN REPLICATION

In today's integrated circuit technology, very precise pattern replication processes are used. Such precision has enabled the industry to proceed to ever finer feature dimensions while simultaneously increasing the overall size of the patterns replicated. Because the component count per chip allowed by present yield limitations is much lower than the limit imposed by the available pattern replication steps, the industry has been able to produce many chips simultaneously and thus achieve very low costs. If yields permitted a component count per chip larger than the limit imposed by pattern replication steps--as may well be the case in the future--multiple pattern replication steps would be required to produce different areas of the same chip. The precision available in pattern replication provides a natural boundary for growth of component count in integrated circuits; it will be substantially more difficult, but not impossible, to produce chips more complex than is permitted by this precision.

The number of components that can be fabricated in a single circuit have a natural limit determined by replication precision. The ultimate precision available in pattern replication steps is limited by the materials used to hold the patterns. It appears that the anomalous dimensional changes in silicon subjected to the high temperatures required for integrated-circuit processing will limit precision to about 10^5 resolution elements on a side or approximately 10^8 devices per chip. Chips with larger component counts must of necessity be assembled from subunits linked together by areas permitting misregistration. The connection areas will have conductors considerably larger than those within the subunit, and the size of these conductors will limit the number of interconnections between subunits. If the precision of newer fine-line replication technologies is substantially less than that of current optical techniques, the natural subunit may have only about 10^6 devices.

Figure 1 shows the relationship between replication precision and the difficulty of replication. If the precision is adequate to cover

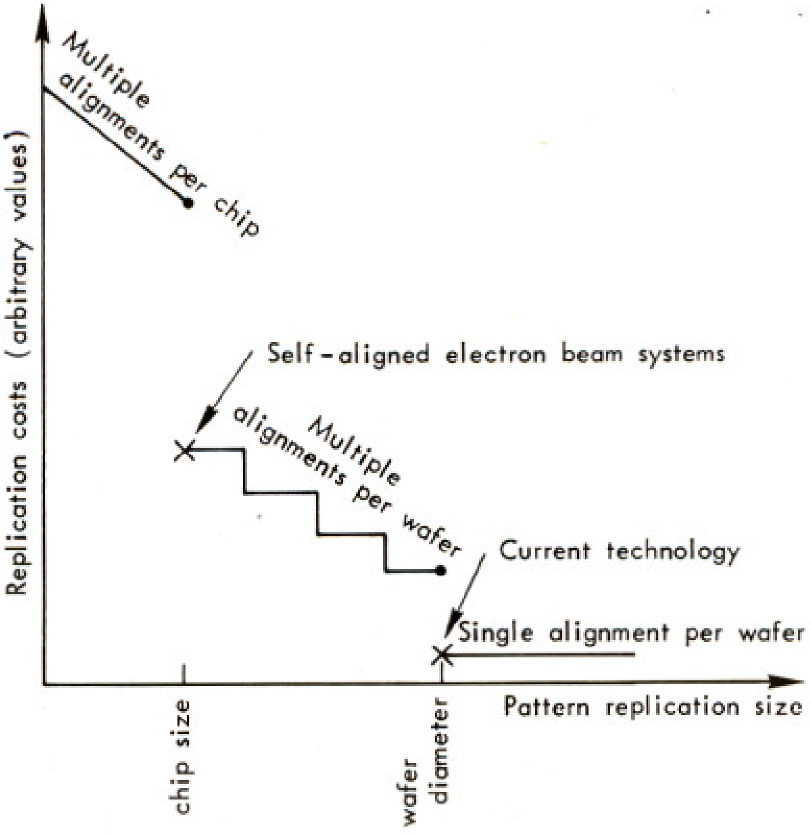

Fig. 1--Schematic representation of the effect of pattern replication precision on replication costs an entire wafer, replication is relatively easy. If the precision is inadequate for an entire wafer but adequate for an entire chip, the cost of replication will depend on how many replications are required to cover the wafer, and will thus vary in staircase fashion with changes in precision. If the precision is inadequate to cover even the size of a single chip, multiple replications per chip are possible but will be substantially more expensive. Current practice involves pattern replication over an entire wafer. Most of the self-aligned electron beam systems now under development operate with one alignment per chip.

PATTERN GENERATION

Pattern generation, unlike replication, needs to be done only to the precision required by the component count of the circuit, or part of a circuit, for which the pattern is being generated. In today's technology, patterns for a single circuit need to be generated with a precision of only about one part in 10^4. These individual patterns are then combined by a step-and-repeat process to form highly precise masks, which are accurate to about one part in 10^5 and are ultimately used to replicate the patterns onto the semiconductor. As component count increases, both the need for accuracy in pattern generation and the number of features to be recorded in a pattern will also increase.

The time that it takes a pattern generator to record a pattern depends, of course, on the number of parts that are in the pattern and the speed of the pattern generator. As will be explained below, pattern generation times increase more than linearly with decreases in feature size. Pattern replication, if it can be done at all, can be done in a time that is independent of the actual pattern replicated for a given feature size. This time is usually much less than that required for pattern generation. Moreover, equipment to generate patterns must be digitally controlled as well as precise, whereas pattern replication equipment can often be less costly analog equipment similar to a camera. To the extent that replication devices must include complex alignment procedures, however, they become more complex and costly. If the alignment procedures are very complex, the distinction between pattern generation and pattern replication steps may become diffused. Alignment procedures are an important part of the development of the new replication technologies.

The newly evolving submicron fabrication technologies are based on new mechanisms for pattern generation and replication. It is generally agreed that pattern generation is best done with computer-controlled electron beam recording, but there is essentially no agreement on the design details of such recorders. Two major categories of recording systems exist: raster scan and vector control. The raster scan systems methodically cover the area of the pattern to be generated, turning the electron beam on or off as demanded by the requirements of the pattern being recorded. The vector control systems deflect the beam to locations specified by the needs of the pattern.

Although the raster-scan technique is simple in concept, and avoids critical constraints on the linearity and hysteresis of the deflection system, it has two major disadvantages: (1) The video rate of a given pattern is very high; the rise time must be less than the pixel time. (2) The format makes it difficult to adjust exposure to compensate for local variations in pattern complexity. The vector scan, on the other hand, requires exceedingly tight control of the deflection system but preserves the locality of shape. This locality permits the beam current to be dynamically adjusted for interiors of large areas and makes adjustment for proximity effect at adjacent edges easier.

It thus appears that if the deflection problems can be solved, vector scan systems will provide about 10 or more times the throughput of raster scan systems for line dimensions of $\sim$0.3 μm or smaller.

SPEED LIMITS IN PATTERN GENERATION

Both raster-scan and vector-control pattern generators face basic limitations in operating speed. The rate at which a pattern can be generated can be limited by two factors: (1) Given that the pattern contains a certain amount of information, the bandwidth of the systems that control the electron beam place limits on how fast the pattern can be transmitted. (2) Resist sensitivity sometimes limits the writing rate of the electron beam.

Many complex phenomena are involved when electrons expose a resist by losing energy in it. The electron scatters elastically and inelastically, changing direction and losing energy. The resist changes physically and chemically, and the pattern left after development depends on all of these effects plus those introduced by the development process. However, there are certain fundamental relationships that are useful to observe.

Let us assume that the resist requires a dose of Q coulombs/cm^2 for correct exposure. In order to be certain that a given pixel is exposed, i.e., to ensure adequate pixel signal-to-noise level, at least a minimum number of electrons, N_m, must strike and lose their energy in each pixel. This is fundamental and important. Since $Q/e \geq (N_m/\ell_p{}^2)$, where ℓ_p is the pixel linear dimension and e is the electron charge, Q *must increase* as ℓ_p decreases, for the probability that each pixel will be correctly exposed to remain constant. Stated another way, based on pixel signal-to-noise considerations, the minimum total number of electrons needed to expose reliably a pattern of a given complexity, i.e., with a given number of pixels, is *independent* of the size of pixels. More sensitive resists are useful for larger pixels; less sensitive resists *must* be used for smaller pixels. This argument assumes that an electron's energy is lost within a pixel, i.e., that the transverse scattering is considerably smaller than a pixel, and that the beam size is at least as small as a pixel.

Appendix C exhibits the fundamental considerations of electron beam formation and focusing that cause the time, τ, required to expose a pixel to N_m electrons to increase as the pixel linear dimension ℓ_p decreases. As shown by the left-hand curve in Fig. 2, $\tau \propto \ell_p{}^{-8/3}$. To correctly expose a real resist of sensitivity Q coulombs/cm^2, a fixed number of electrons per unit area must strike the resist, and the time required to expose such a resist is $\tau_R \propto \ell_p{}^{-2/3}$. A family of curves corresponding to such real resist exposure is also shown in Fig. 2. For a given probability that each pixel will be correctly exposed, these curves for a real resist cannot extend to the left past the limiting curve. As we proceed to the right of the limiting curve along a curve for constant sensitivity, Q, the number of electrons striking each picture

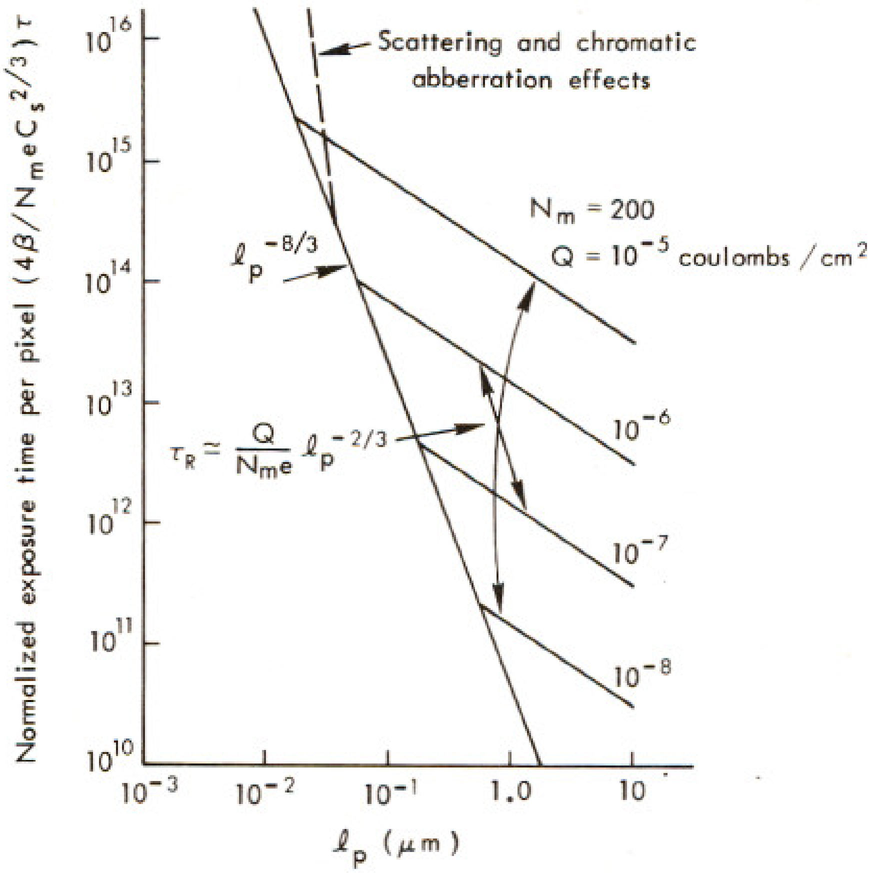

Fig. 2--Normalized exposure time/pixel vs pixel dimension

element increases, improving the pixel signal-to-noise ratio. Because the normalization factor on the ordinate of Fig. 2 includes N_m, the vertical positioning of the τ_R curves depends on the value of N_m actually chosen. For binary exposure, the probability that a pixel struck by 200 electrons is not correctly exposed is less than 10^{-12}; if struck by 100 electrons, a pixel has a probability of incorrect exposure of 3×10^{-7}, enough to cause many errors in a pattern of 10^{10} pixels. Hence, we have set $N_m = 200$ in the τ_R curves of Fig. 2.

These curves predict that for $Q = 10^{-8}$ coulombs/cm^2, pixels smaller than $\ell_p = 1.0$ μm should be possible, and for $Q = 10^{-6}$ coulombs/cm^2, pixels below $\ell_p = 0.1$ μm should be attainable, based on signal-to-noise ratio considerations alone. Resists such as polymethyl methacrylate processed for high resolution by the correct choice of developer have demonstrated line widths less than 0.1 μm. The fundamental point to emphasize is that slow resists are necessary to get higher resolution, a result familiar to all photographers. Note that if all electron energy is not dissipated within the pixel (due to lateral scattering, for example), the exposure time per pixel increases and the solid curve in Fig. 2 moves toward the dashed curve. Inclusion of quantitative information on scattering and aberrations in addition to spherical aberration will cause the actual limiting curve to move toward the right at small pixel dimensions, as shown in Fig. 2.

It is instructive to consider an example. Suppose that we want to know how long it will take to expose a chip on an integrated circuit consisting of 10^8 pixels, with a pixel dimension $\ell_p = 500$ Å. We may choose to use a tungsten hairpin cathode, a LaB_6 cathode, or a field emission cathode, as discussed by Broers (1972). Furthermore, we may choose to scan a relatively small area of the surface and move the object being exposed often (or continuously, as in EBES); or we may choose to scan a larger field and move the object less frequently. Normally, a vector scan is chosen in the latter case, and a raster scan in the former (as in EBES). If the smaller field is chosen, the electron optics can be optimized for a higher current and current density at the object being exposed.

From Broers (Fig. 1 or 10 and Fig. 11) we determine that the current in a beam of 500 Å diameter from a tungsten cathode, standard LaB_6 cathode, and a field emission gun followed by a lens is 1.3×10^{-11}, 6×10^{-10}, and 7×10^{-9} amps, respectively. For the stated resolution of 500 Å, our previous argument suggests that $Q \geq 6.4 \times 10^{-6}$ coulombs/cm^2. If, for this calculation, we do not consider the settling time of the beam and the repositioning-registration time of the stage for vector scan systems, and we assume that the rise-time of a raster scan system can be arbitrarily fast, the time that the beam must be on the chip to expose 1 pixel, 10^8 pixels, or 1 cm^2 is given in Table 3.

Many assumptions have been made in arriving at the numbers in Table 3. For example, it is assumed that the vector scan must expose every pixel, whereas in practice only 5 to 50 percent of the pixels are exposed; hence the *exposure time* for the vector scan system is very

Table 3

COMPARISON OF EXPOSURE TIMES AT 500 Å RESOLUTION FOR DIFFERENT SOURCES AND DIFFERENT ELECTRON-OPTICAL PARAMETERS

Source	Exposure Time			Conditions
	For 1 pixel = τ(μsec)	For 10^8 pixels = 10^8 τ(sec)	For 1 cm^2 = 4 × 10^{10} τ(hr)	
Tungsten (hairpin)	9.7 μsec	970 sec	107 hr	*Vector Scan* C_s = 12 cm, C_c = 5 cm [Broers Fig. 11] Beam Voltage = 25 kV
LaB_6 (standard)	0.26 μsec	26 sec	2.9 hr	
Field effect (gun & lens)	0.018 μsec	1.8 sec	0.2 hr	
Tungsten (hairpin)	0.16 μsec	16 sec	1.7 hr	*Raster Scan* C_s = 1.8 cm, C_c = 1 cm, [Broers Fig. 1] Beam voltage = 25 kV
LaB_6 (standard)	0.021 μsec	2.1 sec	0.23 hr	

Source	Exposure Time	Conditions
Conventional X-ray	1000 sec	100 mA, 10 kV 30 cm working dist
Synchrotron radiation	1 sec	10 mA 500 MeV beam

pessimistic (i.e., too long). The exposure times can be shortened by increasing the beam brightness. For a given resolution in the resist, the resist sensitivity has a maximum value $Q_m = N_m e/\ell_p^2$; using resists more sensitive than this will decrease resolution or the confidence in exposure, and hence the yield. Thus a more sensitive resist than $Q = 6.4 \times 10^{-6}$ coulombs/cm^2 in the above example is not useful, and will not usefully decrease exposure times.

ALIGNMENT LIMITATIONS FOR PATTERN REPLICATION

Pattern replication for submicron dimensions, unlike pattern generation, does not suffer any important fundamental speed limitations, but it does suffer various kinds of resolution and precision limitations. We have become accustomed to pattern replication processes with accuracies on the order of one part in 10^5, a level difficult to maintain with some of the new methods. Noncontact pattern replications are highly desirable, because placing a mask in contact with the integrated circuit damages the mask and forces one to use multiple generations of masks, which interposes stages of replication otherwise unnecessary. While conformable masks alleviate the problem to some extent, they do not eliminate it. Noncontact printing requires highly collimated sources of radiation and sophisticated alignment schemes. So far, the most ideal radiation source seems to be ultraviolet synchrotron radiation. However, both ultraviolet and X-rays require masks that are much thinner than the glass plates currently in use, and the dimensional stability of such masks is not likely to be any better. Decreased accuracy in replication means that smaller areas will be exposed at a single replication step. If wafer size is maintained, this implies multiple exposures of each wafer.

Alignment for the replication steps is now done manually by a light microscope. For alignments down to 0.1 micron or even finer, it still appears possible to use light as an alignment mechanism. Many systems use scanning electron beams in various configurations for automatic alignment. In addition to an evacuated chamber, they require fairly complicated sensing and actuating systems and are thus quite expensive.

Some of the electron-beam pattern-generation systems now being used for making masks can also be used to expose wafers directly. Point-by-point serial electron-beam writing equipment references the beam to registration markers on the wafer so that the orientation and size of the pattern can be held within acceptable tolerances during exposure. By using a laser interferometer to measure beam position with respect to a given origin, it is possible to expose an area of $\gtrsim 10^5$ pixels $\times \gtrsim 10^5$ pixels ($\gtrsim 10^{10}$ pixel^2). Without a laser interferometer, areas of $\sim 10^6$ pixel^2 to 10^8 pixel^2 can be exposed after each registration, depending on the stability of the electrical signals used to accelerate and deflect the beam, the electron optical corrections, etc.

The time required for each registration determines how many registrations are economically feasible within the processing time devoted to a wafer. As feature sizes continue to decrease, exposure times in electron-beam pattern-generation equipment of necessity increase, so that automatic alignment becomes a small fraction of total exposure time. Just as a constant number of electrons per pixel are required to provide an adequate signal-to-noise ratio for the self-alignment process, so are a certain number of electrons required to reduce dimensional signal-to-noise ratios to a desired level. Thus, for patterns of more than a certain number of pixels, self-alignment should be acceptable for each pattern independent of feature size.

If the fractional substrate distortion,

$$\left| \frac{r_m - r_m'}{r_m} \right| ,$$

where r_m is the measured distance between two features and the prime indicates a measurement after processing, exceeds the ratio of feature size to pattern size, misregistration is bound to occur. To avoid this, local registration is essential for smaller circuit dimensions. For patterns exceeding 10^3 to 10^4 pixels on a side, laser interferometry servo control will be required in addition to electron-beam-deflection control. For patterns simpler than 10^3 or 10^4 pixels on a side, open-loop electronic control following alignment will suffice and is substantially faster and less expensive than servo control.

VI. CIRCUIT ORGANIZATION

Much of the improvement in the functional characteristics of integrated circuits over the next decade can be expected to come from better circuit organization. Whenever many thousands of components are used to perform a function, the organization of those components becomes a major factor in their collective performance. Integrated circuits are now at a level of complexity where careful organization pays off handsomely, but we do not yet clearly understand the fundamental design constraints imposed by circuit technology.

The first thing one observes in a complex integrated circuit is that wires occupy most of its area. As a circuit increases in component count, the percentage of its area devoted to wires will also increase, unless the circuit is carefully organized. The reason is that not only do the number of wires increase in direct proportion to the component count, but also each wire tends, on average, to be longer. As Appendix A shows, for the upper limit of random interconnection, the space per component required for wiring increases linearly with the number of components; the area cost of *each component* thus goes up solely because of the number of components, not because of the component itself. The gains to be achieved by arranging components in rows and columns with local wiring between them therefore increase with increasing component count.

The second thing one observes about complex integrated circuits is that their ultimate computation speed is limited by the rate at which information can be transmitted from one place to another. But the rate at which information can be transmitted from one place to another is limited because the conductors have capacitance and must be driven by sources of finite resistance. There is an advantage, therefore, in physically arranging information so that data elements that must be combined in a computation are located close to each other and close to the circuits that perform the computation. The difficulty of computation tasks should not be thought of in terms of megabits transmitted or computed per second, but rather in terms of megabit *meters* per second.

Heretofore, we have designed computing equipment with a rather clear-cut physical as well as functional separation between memory and computing, a separation no longer warranted by our technology. We now have a strong motivation to commingle processing and memory functions in order to derive the most computational output from our equipment. The motivation is increased by the fact that both memory and computation are most economically provided by the very same semiconductor technology. Unfortunately, we know relatively little about how best to organize our logical elements to make use of these capabilities.

In the past, suggestions for combining computing and memory have been implemented in a context in which the fundamental costs of the functions have been drastically different, i.e., thousands of bits of memory could be obtained for the cost of a few computation functions. But for the chip, the costs of processing and memory are more nearly the same. Wiring costs dominate both of these functions, because wiring not only occupies most of the chip area but also introduces most of the propagation delay. This is a whole new context in which organization needs to be explored.

In the construction of large software systems, and in the design and fabrication of complex systems such as oil refineries and aircraft, we have learned that careful organization is essential to success. Whenever many thousands of parts are involved, and all of them must function perfectly for correct overall operation, the conception and planning of the interrelationships between the parts must be done with great care. In software, design concepts such as "structured programming" have been found very useful. Such concepts are really just reflections of good engineering practice: the form of the solution should follow the form of the problem, and parts of the design that are treated separately must be truly separable. In circuitry, as opposed to the logical design of software, conflicts for wiring space become a major design consideration, just as conflicts for piping space are a major consideration in the design of oil refineries, ships, and aircraft.

Careful organization of the design task, as well as of the design itself, may be required as the complexity of hardware design begins to approach the complexity of the software systems with which we have had such abysmal disasters. In fact, it may be argued that the distinctions between computer hardware and software are beginning to vanish. Software, after all, is a medium in which one can describe logical processes and have them performed; it is characterized by a very high design cost and a very low replication cost. Silicon microcircuit technology is also characterized by high design cost and low replication cost. While there are those who maintain that the solution to the software problem lies in improving integrated circuits, we believe that unless care is taken now, the design of the integrated circuit may itself become "the software problem."

We have found relatively little evidence that system designers are doing any fundamental thinking about organization. The theoretical results discussed in Appendix A are the only ones we know of that deal with the problem of organizing wiring. It would be nice if a body of theory about the geometric aspects of computing could be developed, and if that theory could be applied to practical devices. As circuit complexities involving several tens of thousands of components are reached, there is no doubt that improved organizations can make large differences in design cost and functional performance.

Appendix A
HOW BIG MUST AN INTEGRATED CIRCUIT CHIP BE?

Complex component interconnections are of two types: regular wiring patterns and irregular wiring patterns. Regular wiring patterns are those in which the wires are arrayed in rows or columns between the cells of an array of similar logical elements. Such regular patterns are used to implement memories, read-only memory cells, adders, array multipliers, bit maps, and a host of other useful logical functions.

Irregular wiring patterns are used when insufficient regularity is available in the function being implemented. Collections of logic gates to implement control functions for computing machinery, for example, are often implemented as irregular wiring patterns. At a higher level, irregular wiring patterns are found as the interconnections between subunits composed of regular wiring. Irregular patterns of wiring are difficult to design, difficult to inspect, difficult to certify as correct, and, as we shall see, wasteful of chip space.

In order to model the statistics of irregular wiring patterns, let us examine a random wiring model. We will assume that there are N points on a two-dimensional surface that are to be interconnected by a known, but random, pattern of wires. We shall try to estimate, given a center-to-center wire spacing, w, how much area will be occupied by the wires. We will assume for the moment that the wiring pattern involves at least two layers of wiring so that wires may cross each other, and that most wiring runs are arranged either vertically or horizontally in the available space. The statistics for a random wiring pattern will serve as an upper bound on the amount of space required for better organized wiring, since any effort devoted to the random pattern will surely pack it more closely together.

Experience with the layout of printed circuit boards, integrated circuit chips, and highway networks tells us that the critical congestion problem will occur at the center of the layout. We will therefore estimate the number of wires that cross the midline of the layout, realizing that there must be enough space along the midline to accommodate these wires. Any wire that crosses the midline will be connecting a point or points on one side of the midline with a point or points on its other side. We will assume that the wiring layout has been done with at least enough common sense to permit a wire to cross the midline only once regardless of how many points on each side of the midline it interconnects.

We will be interested in how the expected number of midline crossings depends on the number of points connected together into a single "net" by each wire. Let us first consider nets involving only two points. Given N points to interconnect, there are N/2 such wires. Of these, one-half will cross the midline, since only in half of the cases will the two points to be interconnected lie on opposite sides of the midline. We can therefore expect N/4 wires to cross the midline. Now let us consider nets involving three points. There are N/3 such nets. Of these, one-eighth will involve exclusively points on one side of the midline and one-eighth will involve exclusively points on the other side, leaving three-fourths of the wires to cross the midline. Since (3/4) × (N/3) = N/4, we can again expect N/4 wires to cross the midline! For nets of four points, the expected number of crossings is (1 - (1/16) - (1/16)) × (N/4) = 7N/32, again very close to N/4. In fact, as Table 4 shows, the expected number of midline crossings is a very slowly varying function of the number of points in the net. For nets of most interesting sizes, we can therefore conclude that *given N points to interconnect, about N/4 wires can be expected to cross the midline of the layout.* This result was published, with embellishments, by Sutherland and Oestreicher (1973) in a paper entitled: "How Big Should a Printed Circuit Board Be?" It is a remarkably simple and powerful result.

Knowing how many wires will cross the midline of a random wiring layout enables us to determine how much space to provide for them. Naturally, any layout more systematic than random will require less space, and so we have an upper bound. Sutherland and Oestreicher successfully used their result to choose the size and component count of a family of printed circuit boards in such a way as to make the layout

Table 4

EXPECTED NUMBER OF ESSENTIAL MIDSECTION CROSSINGS AS A FUNCTION OF NET SIZE

Net Size	Expected Crossings per Net	Expected Crossings
n	$C = 1 - P_1^n - P_2^n$	$W_m = \frac{N}{n} (1 - P_1^n - P_2^n)$
2	$1 - 1/2^2 - 1/2^2 = 1/2$	$\frac{N}{2} \times 1/2 = \frac{N}{4} = 0.25N$
3	$1 - 1/2^3 - 1/2^3 = 3/4$	$\frac{N}{3} \times 3/4 = \frac{N}{4} = 0.25N$
4	$1 - 1/2^4 - 1/2^4 = 7/8$	$\frac{N}{4} \times 7/8 = \frac{7N}{32} = 0.2185N$
5	$1 - 1/2^5 - 1/2^5 = \frac{15}{16}$	$\frac{N}{5} \times \frac{15}{16} = \frac{15N}{80} = 0.188N$
N	1	$\frac{N}{N} \times 1 = 1 = 1$

problem easy. More important, however, is that this result provides us with an understanding of the growth laws for wiring complexity.

To consider how the area occupied by wiring changes with complexity, let us determine the area per point interconnected on a two-dimensional surface that is occupied by wiring. If there are N points to interconnect, there will be N/4 wires crossing the midline of the layout, and the layout must therefore be $(wN/4)^2$ in area if the wire center-to-center spacing is w. For each point interconnected, then, an area of $(w/4)^2N$ will be required for wiring. As long as the size of the points interconnected is larger than $(w/4)^2N$, the points interconnected will occupy an area larger than the wiring. As the number of points to be interconnected is increased, the *area per point* occupied by wiring increases; through no fault of the individual interconnection points, the cost of interconnecting *each* of them increases linearly with their numbers. When enough points are involved (a remarkably small number) so that $(w/4)^2N$ exceeds the size of an individual point, the area required for the layout will be dominated by the area occupied by wiring. This is the regime in which all integrated circuits are designed, in which many printed circuits lie, in which the back panels (Semore Cray's "mat") of the largest computers are built, and which causes most of downtown Los Angeles to be paved with overcongested freeways.

Relief from the congestion of two-dimensional wiring can be obtained by resorting to three dimensions. Obviously, providing more levels of wiring serves the same purpose as reducing wire spacing, w, if the points to be interconnected are still arrayed in a plane array. If, however, we had a mechanism for building truly three-dimensional circuits similar to the biological circuits found in the human nervous system, the growth law would be more favorable.

For a three-dimensional arrangement of N points, again N/4 wires can be expected to cross the midplane. Each such wire and the space around it, let us say, has a cross-sectional area of w^2, and so a cube whose side is $w(N/4)^{1/2}$ on a side will suffice to hold the wiring. Such a cube has volume $w^3(N/4)^{3/2}$, of which $(w/2)^3N^{1/2}$ must be assigned to each point interconnected. In three dimensions, then, the volume needed for random wiring attributable to each point interconnected increases only as the square root of the number of points interconnected; whereas in the plane, the area for random wiring attributable to each point increases linearly with the number of points interconnected.

Appendix B

PROGRESS IN DIGITAL INTEGRATED ELECTRONICS*

Gordon E. Moore

Intel Corporation

Complexity of integrated circuits has approximately doubled every year since their introduction. Cost per function has decreased several thousandfold, while system performance and reliability have been improved dramatically. Many aspects of processing and design technology have contributed to make the manufacture of such functions as complex single-chip microprocessors or memory circuits economically feasible. It is possible to analyze the increase in complexity plotted in Fig. B.1 into different factors than can, in turn, be examined to see what contributions have been important in this development and how they might be expected to continue to evolve. The expected trends can be recombined to see how long exponential growth in complexity can be expected to continue.

A first factor is the area of the integrated structures. Chip areas for some of the largest of the circuits used in constructing Fig. B.1 are plotted in Fig. B.2. Here, again, the trend follows on exponential quite well, but with a significantly lower slope than the complexity curve. Chip area for maximum complexity has increased by a factor of approximately 20 from the first planar transistor in 1959 to the 16,384-bit charge-coupled device memory chip that corresponds to the point plotted for 1975; while complexity, according to the annual doubling law, should have increased about 65,000-fold. Clearly much of the increased complexity had to result from higher density of components on the chip, rather than from the increased area available through the use of larger chips.

Density was increased partially by using finer-scale microstructures. The first integrated circuits of 1961 used line widths of 1

*This paper was delivered at the International Electron Devices Meeting, in Washington, D.C., December 1975.

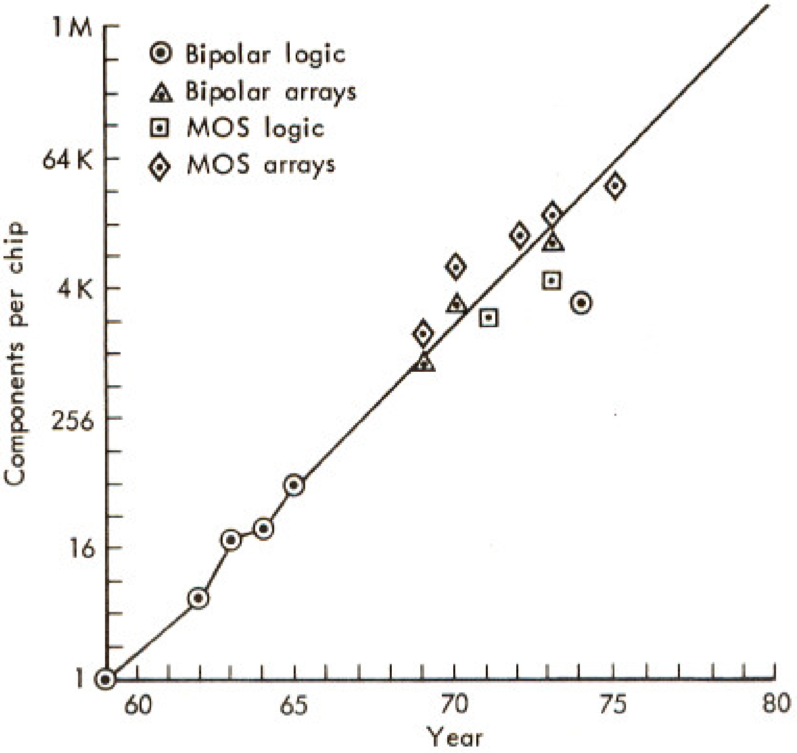

Fig. B.1--Approximate component count for complex integrated circuits vs year of introduction

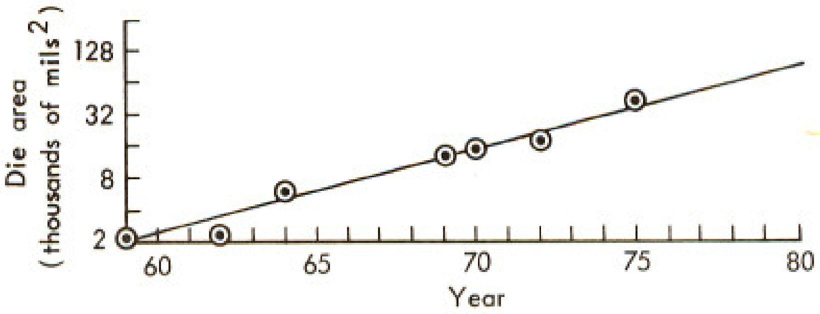

Fig. B.2--Increase in die area for most complex integrated devices commercially available

mil (≃ 25 micrometers), while the 1975 device uses 5 μm lines. Both line width and spacing between lines are equally important in improving density. Since they have not always been equal, the average of the two is a good parameter to relate to the area that a structure might occupy. Density can be expected to be proportional to the reciprocal of area, so the contribution to improved density versus time from the use of smaller dimensions is plotted in Fig. B.3.

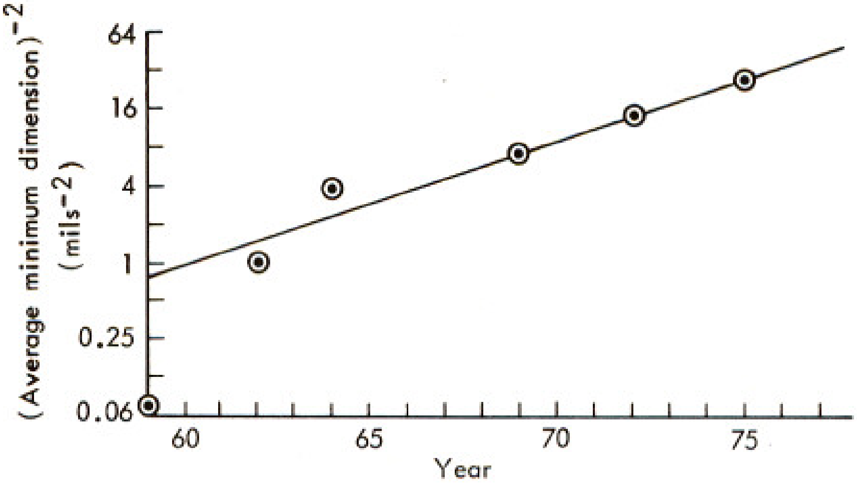

Fig. B.3--Device density contribution from the decrease in line widths and spacings

Neglecting the first planar transistor, where very conservative line width and spacing was employed, there is again a reasonable fit to an exponential growth. From the exponential approximation represented by the straight line in Fig. B.3, the increase in density from this source over the 1959-1975 period is a factor of approximately 32.

Combining the contribution of larger chip area and higher density resulting from geometry accounts for a 640-fold increase in complexity, leaving a factor of about 100 to account for through 1975, as is shown graphically in Fig. B.4. This factor is the contribution of circuit and device advances to higher density. It is noteworthy that this contribution to complexity has been more important than either increased chip area or finer lines. Increasingly, the surface areas of the integrated devices have been committed to components rather than to such inactive structures as device isolation and interconnections, and the components themselves have trended toward minimum size, consistent with the dimensional tolerances employed.

Can these trends continue?

Extrapolation of the curve for die size to 1980 suggests that chip

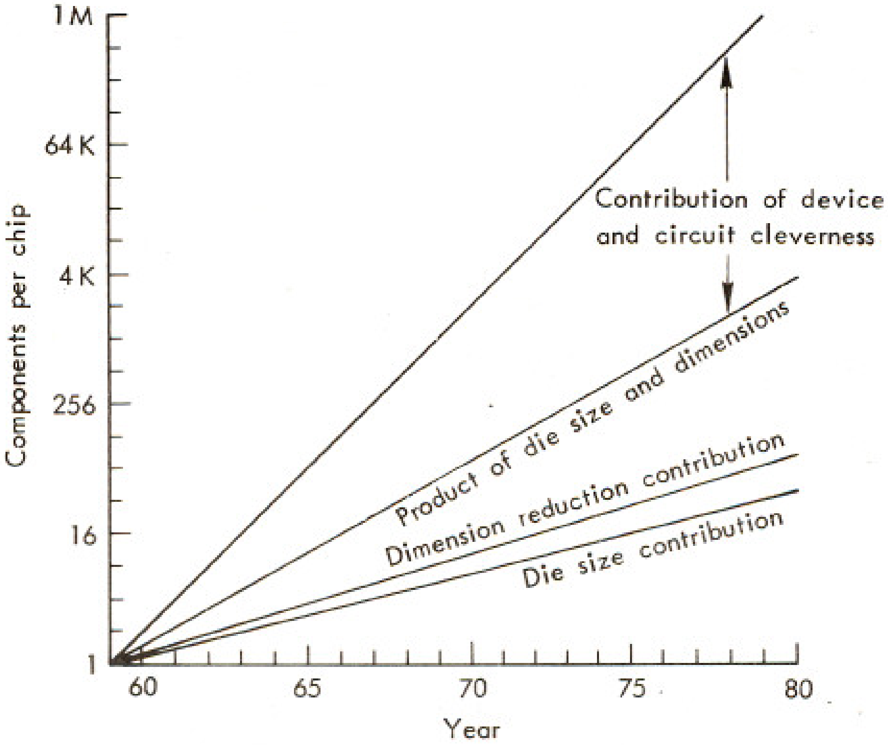

Fig. B.4--Decomposition of the complexity curve into various components

area might be about 90,000 mil^2, or the equivalent of 0.3 in.^2. Such a die size is clearly consistent with the 3-in. wafer now widely used by the industry. In fact, the sizes of the wafers themselves have increased about as fast as die size during the time period under consideration and can be expected to continue to increase. Extension to larger die size depends principally on the continued reduction in the density of defects. Since the existence of the type of defects that harm integrated circuits is not fundamental, their density can be reduced as long as such reduction has sufficient economic merit to justify the effort. I see sufficient continued merit to expect progress to continue for the next several years. Accordingly, there is no present reason to expect a change in the trend shown in Fig. B.2.

With respect to dimensions, in these complex devices we are still far from the minimum device sizes limited by such fundamental considerations as the charge on the electron or the atomic structure of matter. Discrete devices with submicrometer dimensions show that no basic problems should be expected at least until the average line width and spaces are a micrometer or less. This allows for an additional factor of improvement at least equal to the contribution from the finer geometries of the last 15 years. Work in nonoptical masking techniques, both electron beam and X-ray, suggests that the required resolution capabilities will be available. Much work is required to be sure that defect densities continue to improve as devices are scaled to take advantage of the improved resolution. However, I see no reason to expect the rate of progress in the use of smaller minimum dimensions in complex circuits to decrease in the near future. This contribution should continue along the curve of Fig. B.3.

With respect to the factor contributed by device and circuit cleverness, however, the situation is different. Here we are approaching a limit that must slow the rate of progress. The CCD structure can closely approach the maximum practical density. This structure requires no contacts to the components within the array, but uses gate electrodes that can be at minimum spacing to transfer charge and information from one location to the next. Some improvement in overall packing efficiency is possible beyond the structure plotted as the 1975 point in Fig. B.1, but it is unlikely that the packing efficiency alone can contribute as much as a factor of 4, and this only in serial data paths. Accordingly, I am inclined to suggest a limit to the contribution of circuit and device cleverness of another factor of 4 in component density.

With this factor disappearing as an important contributor, the rate of increase of complexity can be expected to change slope in the next few years, as shown in Fig. B.5. The new slope might approximate a doubling every 2 years, rather than every year, by the end of the decade.

Even at this reduced slope, integrated structures containing several million components can be expected within 10 years. These new devices will continue to reduce the cost of electronic functions and extend the utility of digital electronics more broadly throughout society.

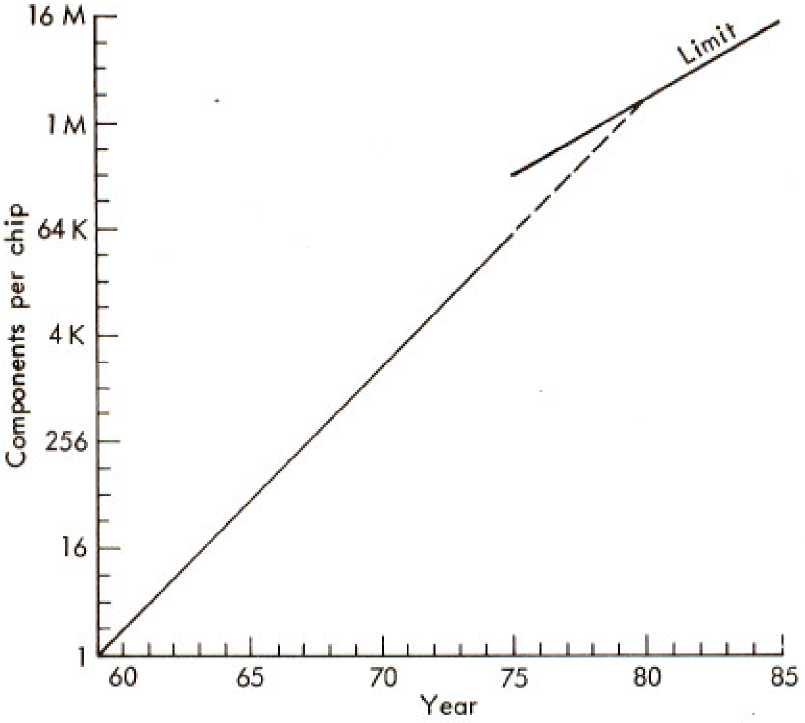

Fig. B.5--Projection of the complexity curve reflecting the limit on increased density through invention

Appendix C

EXPOSURE TIME VERSUS PICTURE ELEMENT SIZE

Consider the time required to expose a pattern with a focused scanning electron beam. The electron beam with current density $J(A/cm^2)$ must strike a pixel for time τ (sec) to produce exposure Q (coulombs/cm^2) $= J\tau$. The beam current density $J = J_c(eV/kT)\alpha^2$ by Langmuir's law, where J_c, T, and V are cathode current density, temperature, and beam accelerating voltage, e and k are the electronic charge (1.6×10^{-19} coulombs) and Boltzmann's constant (1.38×10^{-23} J/°K), and α is the beam convergence angle.

By increasing α, the current density exposing the pattern increases, which is desirable. However, if α is increased too far, the beam spot diameter increases because of the spherical aberration of the focusing system. An optimum value of α occurs when the diameter of the disk of confusion due to spherical aberration, $d_s = 0.5\ C_s\alpha^3$ (C_s is the spherical aberration coefficient), is set equal to the gaussian spot diameter, $d_s = d_g = \ell_p/\sqrt{2}$. Using the normal approximation of adding spot diameters in quadrature, the total spot size then is $d = (d_s^2 + d_g^2)^{1/2} = \ell_p$, the pixel dimension. The optimum convergence angle is then

$$\alpha_{opt} \approx \left[\frac{\sqrt{2}\ \ell_p}{C_s}\right]^{1/3} ,$$

and the exposure in time τ is

$$Q = J\tau = J_c\,\frac{eV}{kT}\left[\frac{\sqrt{2}\ \ell_p}{C_s}\right]^{2/3}\tau = \frac{\beta\pi 2^{1/3}}{C_s^{\ 2/3}}\,\ell_p^{\ 2/3}\tau, \qquad (1)$$

where β is the electron optical brightness ($J_c eV/\pi kT$). Equation (1) gives the change density deposited in a spot of diameter ℓ_p in time τ. For resist exposure, this charge density must equal the resist sensitivity under the exposure conditions used.

To ensure that each pixel is correctly exposed, a minimum number of electrons must strike each pixel. Since electron emission is a random process, the actual number of electrons striking each pixel, n, will vary in a random manner about a mean value, $\bar{n}$. Adapting the signal-to-noise analysis found in Schwartz (1959) to the case of binary exposure of a resist, one can show straightforwardly that the probability of error for large values of the mean number of electrons/pixel $\bar{n}$ is $e^{-\bar{n}/8}/[(\pi/2)\bar{n}]^{1/2}$. This leads to the following table of probability of error of exposure:

$\bar{n}$	50	100	150	200
Probability of error	2.2×10^{-4}	3×10^{-7}	4.7×10^{-10}	7.8×10^{-13}

To be conservative, we choose $\bar{n} = 200$, which should mean that, on average, no pixels in a field of 10^{10} pixels are incorrectly exposed due to randomness, as long as each electron striking a pixel causes at least one exposure event in the resist. For a pixel of dimension ℓ_p, the minimum number of electrons striking it (= 200 here) to provide adequate probability of exposure is N_m, and the charge density is then $Q = N_m e/\ell_p^2$. Substituting into (1) gives

$$N_m e = \frac{\beta\pi 2^{1/3}}{C_s^{2/3}} \tau \ell_p^{8/3}. \qquad (2)$$

To determine how noise limits pixel dimension, arrange (2) so that normalized exposure time depends on pixel dimension; note that $2^{1/3}\pi \simeq 4$:

$$\left[\frac{4\beta}{N_m e C_s^{2/3}}\right] \tau = \ell_p^{-8/3}. \qquad (2a)$$

A corresponding equation for real resist exposure is

$$\left[\frac{4\beta}{N_m e C_s^{2/3}}\right] \tau_R = \frac{Q}{N_m e} \ell_p^{-2/3}. \qquad (1a)$$

Here the same normalization was chosen for τ to facilitate plotting (1a) and (2a) on the same figure of τ vs ℓ_p (see Fig. 2 of the text).

Appendix D

SITES VISITED

Bell Telephone Laboratories, Murray Hill, New Jersey (12/10/75)

W. O. Baker
E. I. Gordon (host)
D. R. Herriott
D. Maydan
R. F. W. Pease
L. Tompson
P. A. Turner

Hughes Research Laboratories, Malibu, California (11/26/75--Mead & Sutherland)

A. Chester
H. L. Garvin
R. Henderson
F. Ozdemir
R. L. Seliger
G. F. Smith (host)
P. A. Sullivan
M. Waldner
E. D. Wolf

IBM Corporation, Yorktown Heights, New York (1/22/76)

W. A. Bohan
A. N. Broers
T. H. P. Chang
D. L. Critchlow
C. D. Cullum
R. E. Gomory
M. Hatzakis
M. B. Heritage
J. W. Newitt (host)
V. Sadagopan
S. Triebwasser
J. Wilczynski
H. Yu

Intel Corporation, Santa Clara, California (2/12/76--Everhart & Mead)

Gordon Moore

Lincoln Laboratory, MIT, Lexington, Massachusetts (1/21/76)

H. I. Smith (host)
J. I. Raffel

RCA Laboratories, Princeton, New Jersey (12/11/75)

J. Herzog (host)
J. Scott
W. Webster

Texas Instruments Incorporated, Dallas, Texas (12/15/75)

G. Barnell
F. Bucy
T. Blocker
N. Einsproch
J. Pankratz (host)
R. Stratton

University of Florida, Gainesville, Florida (12/05/75)--Sutherland)

D. P. Kennedy (host)
A. D. Sutherland

Eiichi Goto, Faculty of Science, University of Tokyo, Tokyo, Japan

Visited Rand on 10/17/75 and gave a talk entitled: "A Double Deflection Cathode Ray Tube"

BIBLIOGRAPHY

Broers, A. N., "Electron and Ion Probes," in R. Bakish (ed.), *Electron and Ion Beam Science and Technology*, The Electro-chemical Society, 1972, pp. 3-25.

Dennard, R. H., et al., "Design of Ion-Implanted MOSFET's with Very Small Physical Dimensions," *IEEE Journal of Solid-State Circuits*, Vol. SC-9, No. 5, Oct. 1974, pp. 256-268.

Hoeneisen, B., and C. A. Mead, "Fundamental Limitations in Microelectronics--I. MOS Technology," *Solid-State Electronics*, Vol. 15, 1972, pp. 819-829.

Hoeneisen, B., and C. A. Mead, "Limitations in Micro-electronics--II. Bipolar Technology," *Solid-State Electronics*, Vol. 15, 1972, pp. 891-897.

Keyes, Robert W., "Physical Limits in Digital Electronics," *Proceedings of the IEEE*, Vol. 63, No. 5, May 1975, pp. 740-767.

Moore, Gordon E., "Progress in Digital Integrated Electronics." A paper delivered at the International Electron Devices Meeting in Washington, D.C., December 1975. To appear in the *Proceedings*. (Reprinted here as Appendix B.)

Schwartz, M., *Information Transmission, Modulation, and Noise*, McGraw-Hill Book Company, Inc., New York, 1959, pp. 382-384.

Sutherland, Ivan E., and Donald Oestreicher, "How Big Should a Printed Circuit Board Be?" *IEEE Transactions on Computers*, Vol. C-22, No. 5, May 1973, pp. 537-542.