Chapter 2: LSI Fabrication

Sections:

The series of steps by which a geometric pattern or set of geometric patterns is transformed into an operating integrated system is called a *wafer fabrication process*, or simply a *process*. An integrated system in MOS technology consists of a number of superimposed layers of conducting, insulating, and transistor forming materials. By arranging predetermined geometric shapes in each of these layers, a system of the required function may be constructed. The task of the integrated system designer is to devise the geometric shapes and their locations in each of the various layers of the system. The task of the process itself is to create the layers and transfer into each of them the geometric shapes determined by the system design. In this chapter we describe the patterning sequence and how it is applied in a simple, specific LSI process: nMOS. A number of other topics are covered which are related to the processing technology, or are closely tied to the properties of the underlying materials. The final section of the chapter discusses the factors affecting the choice of a particular process technology.

Patterning

The overall fabrication process consists of the successive *patterning* of a particular *sequence of layers*. The sequence of steps, called *patterning*, by which geometrical shapes are transferred into a layer of the final system, is very similar for each of the layers. The overall process is more easily visualized if we first describe the details of the patterning of one layer. We can then describe the particular sequence of layers used in the process to build up an integrated system, without repeating the details of patterning for each of the layers.

A common step in many processes is the creation of a silicon dioxide insulating layer on the surface of a silicon wafer, and the selective removal of sections of the insulating layer

exposing the underlying silicon. We will use this step for our patterning example. The step begins with a bare polished silicon wafer, shown in cross section in figure 1. The wafer is exposed to oxygen in a high temperature furnace to grow a uniform layer of silicon dioxide on its surface, as shown in figure 2. After the wafer is cooled it is coated with a thin film of organic resist material as shown in figure 3. The resist is thoroughly dried and baked to insure its integrity. The wafer is now ready to begin the patterning.

Whatever the origin of the pattern to be transferred to the wafer surface, by the time of actual wafer fabrication it almost universally exists as a *mask*. A mask is merely a transparent support material coated with a thin layer of opaque material. Certain portions of the opaque material are removed, leaving opaque material on the mask in the precise pattern required on the silicon surface. Such a mask with the desired pattern engraved upon it is brought face down into close proximity with the wafer surface as shown in figure 4. The dark areas of opaque material on the surface of the mask are located where it is desired to leave silicon dioxide on the surface of the silicon. Openings in the mask correspond to areas where it is desired to remove silicon dioxide from the silicon surface. When the mask has been brought firmly into proximity with the wafer itself, its back surface is flooded with an intense source of ionizing radiation such as ultraviolet light or low energy x-rays. The radiation is stopped in areas where the mask has opaque material on its surface. Where there is no opaque material on the mask surface, the ionizing radiation radiation passes on through and into the resist, the silicon dioxide, and silicon. While the ionizing radiation has little effect on the silicon dioxide and the silicon, it breaks down the molecular structure of the resist into considerably smaller molecules.

After exposure to the ionizing radiation, the wafer has the characteristics shown in figure 5. In areas exposed to the radiation, the resist molecules have been broken down to much lighter molecular weight than that of unexposed resist molecules. The solubility of organic molecules in various organic solvents is a very steep function of the molecular weight of the molecules. Hence, it is possible to dissolve exposed resist material in solvents which will not dissolve the unexposed resist material. In this way the resist can be "developed" as shown in figure 6 by merely immersing the silicon wafer in a suitable solvent.

Thus far, the pattern originally existing as a set of opaque geometries on the mask surface has been transferred into a corresponding pattern in the organic resist material on the surface of the silicon dioxide. This same pattern can now be transferred to the silicon dioxide itself by exposing the wafer to a material which will etch silicon dioxide but not

Figure 1.

BARE WAFER

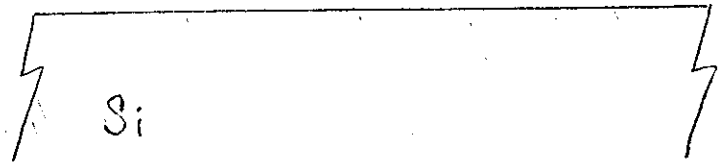


Figure 2.

OXIDATION

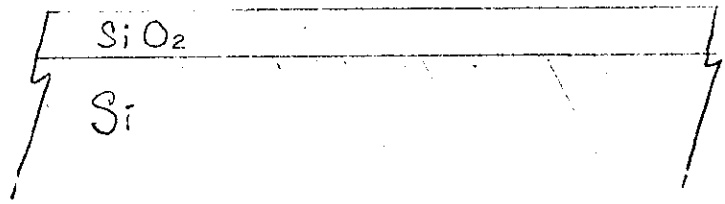


Figure 3.

COAT w/ RESIST

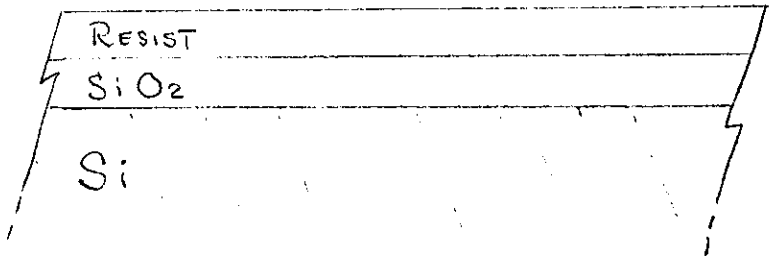


Figure 4.

MASK & EXPOSE

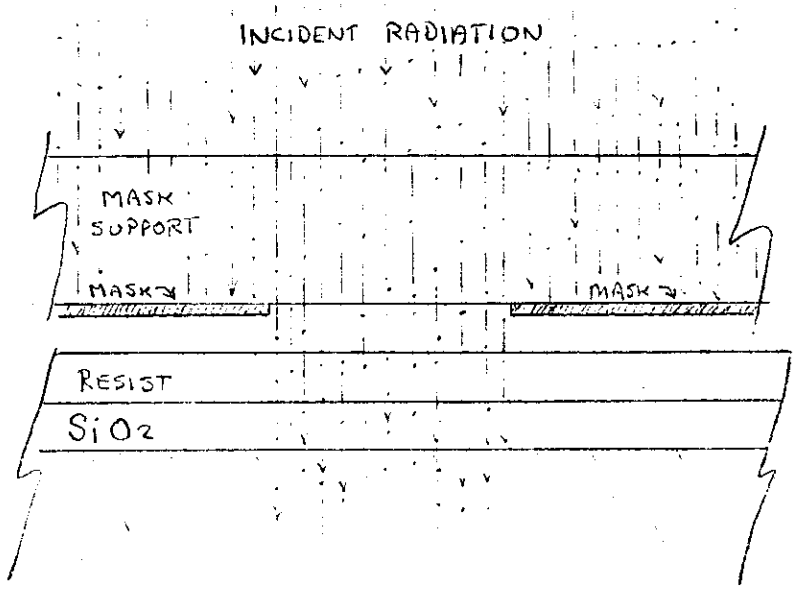


Figure 5.

EXPOSED RESIST

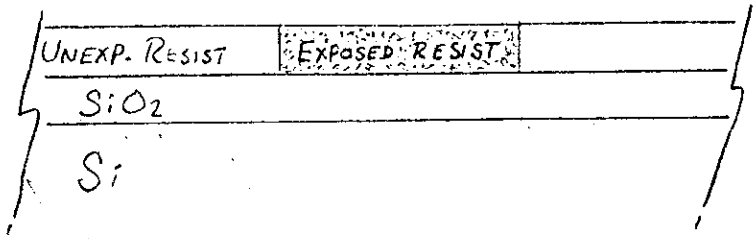


Figure 6.

DEVELOP RESIST

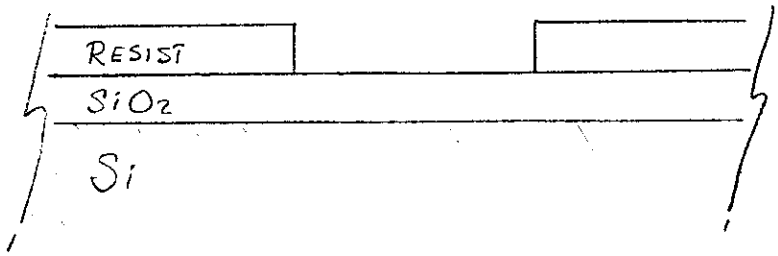


Figure 7.

ETCH

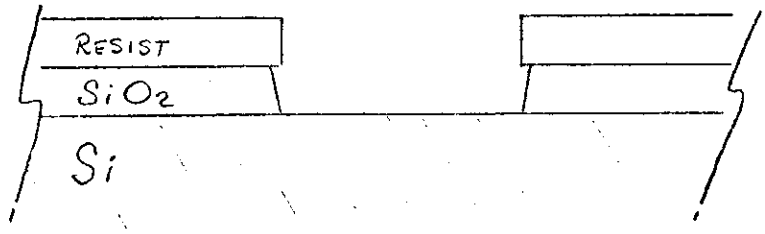
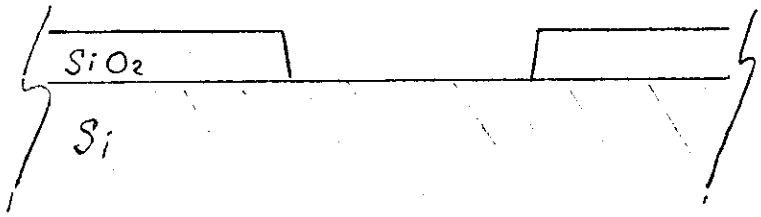


Figure 8.

REMOVE RESIST



attack either the organic resist material or the silicon wafer surface. This etching step is usually done with hydrofluoric acid, which easily dissolves silicon dioxide. However, organic materials are very resistant to hydrofluoric acid, and it is incapable of etching the surface of silicon. The result of this etching step is shown in figure 7.

The final step in the patterning is the removal of the organic resist material. This can be done by using strong organic solvents which will dissolve even unexposed resist material, by using strong acids such as chromic acid which actively attack organics, or by exposing the wafer to atomic oxygen which will oxidize away any organic materials present on its surface. All three techniques have been used to remove the resist materials. Once the resist material is removed, the finished pattern on the wafer surface is as shown in figure 8. Notice that we have transferred the geometric pattern which originally existed on the surface of the mask directly into the silicon dioxide on the wafer surface. While a foreign material was present on the wafer surface during the patterning process it has now disappeared and the only materials present are those which will in fact be part of the finished wafer.

A similar sequence of steps is used to selectively pattern each of the layers of the integrated system. These differ only in the details of the etchants used, etc. Thus as we study the processing of the various layers, the reader need not visualize all the details of the patterning sequence for each layer, but only recognize that a mask pattern for a layer can be transferred into a pattern in the material of that layer.

Scaling of Patterning Technology

As discussed in chapter 1., semiconductor devices could be at least an order of magnitude smaller in linear dimension than those typically manufactured in 1977 and still function correctly. The fundamental dimensional limitation is approximately a one quarter micron channel length, corresponding to a length unit λ (to be discussed under design rules) of approximately 0.1 micron. This limitation appears to apply to both bipolar and MOS technologies. It has been possible for several years to create sub-micron lines using electron beam and x-ray techniques, and there is considerable research and development under way to bring these patterning technologies into general manufacturing use. At present the lithography system appearing most promising for the manufacture of sub-micron devices uses masks made of very thin supporting materials and a heavy metal such as gold as the opaque material. There are resist materials having resolutions of small fractions of a

micron and which are sensitive to electron beams. The opaque material on the thin support may thus be patterned by using direct writing of a scanned electron beam to expose resist on its surface, followed by etching of the resulting selected areas of material. This mask can then be positioned above the silicon wafer and soft x-rays used to expose resist material on the wafer surface. The most desireable source for such soft x-rays appears to be synchrotron radiation from an electron storage ring, operating at approximately 500MEV. Electron storage rings are exceedingly efficient sources of soft x-rays in the one to two hundred volt quantum energy range. This range is optimum for exposure of resists of the sort which have inherent resolutions in the sub-micron region. In addition, sychrotron radiation can be extracted from a storage ring in such a way that it is exceedingly highly collimated. Exposure times for such soft x-rays used with the kind of resists useful at sub-micron resolution are not appreciably longer than those required today with ultra-violet light and standard photo resists. X-rays have sufficiently short wavelength that they exhibit the very desireable property of creating essentially straight sides on the exposed resist material, making them ideal in a patterning process.

The Silicon Gate n-Channel MOS Process

We now describe the particular sequence of patterned layers used to build up nMOS integrated circuits and systems. Figures 9 through 14 illustrate a simple but complete sequence of patterning and processing steps which are sufficient to fabricate a complete, integrated system. The example follows the fabrication of one simple circuit within a system, but all other circuits are simultaneously implemented by the same process. The example used is the basic inverter circuit. The top illustration in figures 9 through 14 shows the top view of the layers of the circuit layout. The lower illustration in each of those figures shows the cross section through the cut indicated by the downward arrows. The vertical scale in these cross sections has been greatly exaggerated for illustrative purposes.

The opening in the opaque material of the *first* mask is shown by the green outline in the top portion of figure 9. This opening exposes all areas that will be eventually be the diffusion level. It includes the sources and drains of all transistors in the circuit, together with the transistor gate areas, and any diffusion level circuit interconnection paths. This mask is used for the first step in the process, the patterning of silicon dioxide on silicon as described in the previous section. The resulting cross section is shown in the lower portion of figure 9.

The second step in the process is to differentiate transistors which are normally "on" (depletion mode) from those which are normally "off" (enhancement mode). This is done by overcoating the wafer with resist material, exposing the resist material through openings in a *second* mask, and developing it in the manner shown in figure 10. This patterning step leaves an opening in the resist material over the area to be selectively turned into normally on transistors. The actual conversion of the underlying silicon is then done by implanting ions of arsenic or antimony into the silicon surface. The resist material, where present, acts to prevent the ions from reaching the silicon surface. Therefore, ions are only implanted in the silicon area free of resist. The implanted layer, which causes a slight n-type conductivity in the underlying silicon, is shown by the yellow box in figure 10. Once the depletion areas are defined, the resist material is removed from the surface of the wafer.

The wafer is then heated while exposed to oxygen, to grow a very thin layer of silicon dioxide over its entire surface. It is then entirely coated with a thin layer of polycrystalline silicon, usually called *polysilicon* or *poly* for short. Note that this polysilicon layer is everywhere insulated from the underlying materials by the layer of thin oxide, and additionally by thicker oxide in some areas. The polysilicon will form the gates of all the transistors in the circuit and will also serve as a second layer for circuit interconnections. A *third* mask is used to pattern the polysilicon by steps similar to those previously described, with the result shown in red in figure 11. The left-most polysilicon area will function as the gate of the pull down transistor of the inverter we are constructing, while the square to the right will function as the gate of the depletion mode pull up transistor.

Once the polysilicon areas have been defined, n-type regions can be diffused into the p-type silicon substrate, forming the sources and drains of the transistors and the first level of interconnections. This step is done by first removing the thin gate oxide in all areas not covered by the (red) polysilicon. The wafer is then exposed to n-type impurities such as arsenic, antimony or phosphorus at high temperature for a sufficient period of time to allow these impurities to convert the exposed underlying silicon to n-type material. The areas of resulting n-type material are shown in green. Notice, in the *cross section* of figure 12., that the red polysilicon area and the thin oxide under it act to prevent impurities from diffusing into the underlying silicon. Therefore the impurities only reach the silicon substrate in areas not covered by the polysilicon and not overlain by the thick original oxide. In this way the active transistor area is formed in all places where the patterned polysilicon overlies the thin oxide area defined in the previous step. The diffusion level

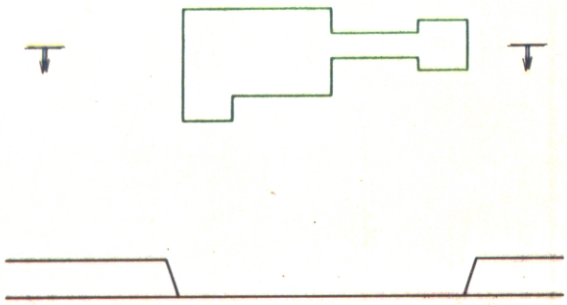


Fig.9. Patterning SiO_2

Fig.10. Ion Implantation

Fig.11. Patterning Polysilicon

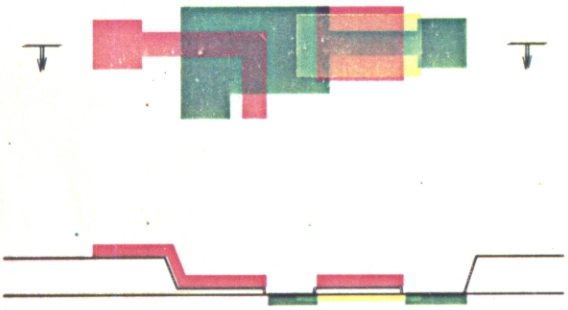


Fig.12. Patterning Diffused Region

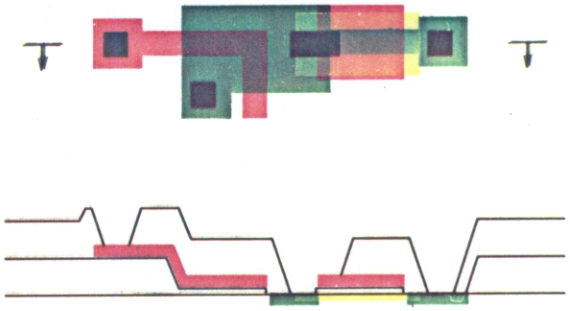


Fig.13. Placing Contact Cuts

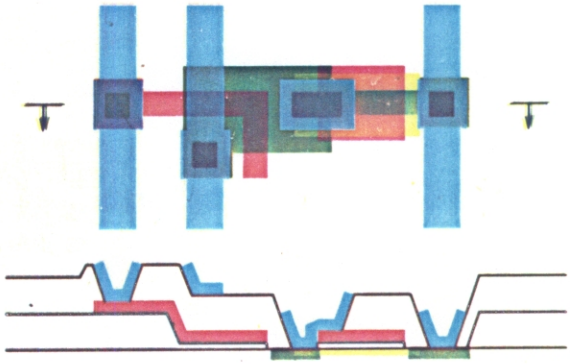


Fig.14. Patterning the Metal Layer

sources and drains of the transistors are automatically filled in between the polysilicon areas and extend up to the edges of the thick oxide areas. The major advantage of the silicon gate process is that it does not require a critical alignment between a mask which defines the green source and drain areas and a separate mask which defines the gate areas. Rather, the transistors are formed by the intersection of the two masks, and the conducting n-type diffused regions are formed in all areas where the green mask is not covered by the red mask.

We now have all the transistors of the basic inverter circuit defined. There remains only the task of electrically connecting them. Required are a connection from some other circuit in to the input gate, a connection between the gate and source of the pullup, and appropriate connections to VDD and GND. These interconnections will be made with a metal layer that can make contact to both the diffused areas and the polycrystalline areas. However, in order to ensure that the metal does not make contact to underlying areas except where intended, another layer of insulating oxide is coated over the entire circuit. At the places where the overlying metal is to make contact to either the polysilicon or the diffused areas, the overlying oxide is selectively removed by the patterning process as previously described. The result of coating the wafer with the overlying oxide and removing this oxide in places where it is desired to make contacts, is shown in figure 13. In the top view, the black areas are those defined by openings in the contact mask, the *fourth* in the process's sequence of mask patterns. In cross section notice that in the contact areas all oxide has been removed down to either the polycrystalline silicon or the diffused area.

Once the overlying oxide has been patterned in this way, the entire wafer is coated with metal, usually aluminum, and the metal is patterned with a *fifth* mask to form the conducting areas required by the circuit. The top view in figure 14 shows three metal lines running vertically, the left most connecting to the input gate of the inverter, the center one being ground, and the right-most one forming the VDD connection to the inverter. The peculiar structure* formed by the metal square slightly to the right of center connects the polysilicon gate of the depletion mode pull up transistor to its source and to the drain of the pull down transistor. Rather than making two separate contacts from the metal line to the pullup's polysilicon gate region and to the adjacent diffusion region, area can be conserved by coalescing the contacts into the compact arrangement shown. This geometrical arrangement is known as a *butting contact* and will be used extensively throughout the text.

*In general, it is good practice to avoid placing contacts over active transistor area whenever possible. However, butting contacts in the location shown here reduce the area and simplify the geometry of the basic inverter and

many other circuits, and have been so placed by the author's in many systems run on a number of commercial fabrication processes without ill effects. A more conservative approach would be to place the butting contact adjacent to, rather than over, the active pullup area.

The inherent properties of the silicon gate process allow the blue metal layer to cross over either the red polysilicon layer or the green diffused areas, without making contact unless one is specifically provided. The red polysilicon areas, however, cannot cross the green diffused areas without forming a transistor. The transistors formed by the intersection of these two masks can be either enhancement mode if no yellow implantation is provided, or depletion mode if such an implantation is provided. Hence, the enhancement mode transistors are defined by the intersection of the green and red masks while the depletion mode transistors are defined by the intersection of the green, red and yellow masks.

If we wish to fabricate only a small number of prototype system chips and to have access to the metal level for the probing of test points, this completes the wafer fabrication sequence. However, when fabricating large numbers of chips of a debugged design, the wafer surface is usually coated with another layer of oxide. This step, called overglassing, provides physical protection for the devices in the system. A *sixth* mask is then used to pattern contact cuts in the overglassing at the locations of relatively large metal contact pads.

Each wafer contains many individual chips. The chips are separated by scribing the wafer surface with a diamond scribe, and then fracturing the wafer along the scribe lines. Each individual chip is then cemented in place in a package, and fine metal wire leads are bonded to the metal contact pads on the chip and to pads in the package which connect with its external pins. A cover is then cemented over the recess in the package which contains the silicon chip, and the completed system is ready for testing and use.

Yield Statistics

Of the large number of individual integrated system chips fabricated on a single silicon wafer, only a fraction will be completely functional. Flaws in the masks, random particles on the wafer surface, defects in the underlying silicon, etc., all cause certain devices to be less than perfect. With present design techniques, any single flaw of sufficient size will kill an entire chip.

The simplest model for the *yield*, or the fraction of the chips fabricated which do not contain fatal flaws, assumes (naively) that the flaws are randomly distributed over the wafer, and that one or more flaws anywhere on a chip will cause it to be non-operative. If there are N fatal flaws per unit area, and the area of an individual chip is A, the probability that a chip has n flaws is in the simplest case just given by the Poisson distribution, $P_n(NA)$. The probability of a good chip is:

$$P_0(NA) = e^{-NA} \qquad \text{[eq.1]}$$

While this equation does not accurately represent the detailed behavior of real fabrication processes, it is a good approximate model for estimating the yield of alternative designs at any given point in time. The exponential is such a steep function that a very simple rule is possible: chips with areas many times 1/N will simply never be found without flaws. Areas must be kept less than a few times 1/N if one flaw will kill a system. There may be design forms which will permit systems to work even in the presence of flaws. If such forms can be developed, the entire notion of yield will be completely changed and much larger chips will be possible.

Once a wafer has been fabricated, each chip must be tested to determine if it is functional. Testing of simple combinatorial logic networks is straightforward and may be done completely, even for very complex networks. - - -*reference?* - - - Complete testing of complex systems with internal sequencing is not in general possible, and most LSI chips manufactured, even at 1977 levels of complexity, are not economically testable even for a small fraction of their possible internal states.

As time passes and the number of devices per chip continues to increase, it becomes important to consider including special functions in the design of LSI systems to improve their testability. The basic problem is to linearize an otherwise combinatorial problem. One approach to this is:

(i) Define the entire system as a set of register to register transfer blocks, i.e. successive stages of storage registers with combinational logic between them.
(ii) Provide for reading and writing from the external world to/from each of the storage registers.

The storage locations are first tested independantly for their ability to reproducibly store

data or control information. If all storage locations pass this test, each combinatorial logic block can be tested separately, by use of its input and output storage locations. Such a test becomes essentially linear in the number of components, and may be accomplished in an acceptable time period, even for extremely complex systems. However, without access to the individual storage locations, testing rapidly becomes hopeless. For this reason even present day microprocessors are very incompletely tested. When one is used for a while, an apparantly new and sudden malfunction may simply be the first occurrance of a particular state of control and data in the system, and thus may represent the first time the device had been "tested "under those conditions.

From experience gained in testing memory parts, it is known that the behavior of one circuit can be influenced by the state of a nearby circuit. For example, a memory cell may be able to remember both a logic-1 and a logic-0 if its neighbor is at a logic-0, but may be able to retain only a logic-0 if its neighbor is at a logic-1. Failures of this type are dependent upon the data patterns present in the system, and are known as *pattern sensitive* failures. In a reasonable (or even an unreasonable) time, it is not possible to exercise even a minute fraction of all the combinations of bit patterns of many integrated systems. What is done instead is to apply our knowledge of the physics of such failures, and construct a *model* for possible failure modes. In the memory example, we may conclude that any flaw not visible optically will be unable to reach beyond the immediate locality of the cell involved. Hence, pattern sensitivity in the behavior of a particular cell may be introduced by other cells in the same row or column of an array of memory cells, or by nearest diagonal neighbors. A test for pattern sensitivity under this model is quite fast, being only slightly worse than a linear function of the number of devices on the chip.

In order to test for pattern sensitive failures, we must construct a physical model for the possible failure mechanisims. This model will inevitably include the physical proximity of other signals. For this reason, any practical test for pattern sensitive failures must be based on a knowledge of the physical location of the various elements of the subsystem being tested. The task of preparing such tests is thus greatly eased by regularity in the design and physical layout of a system.

Scaling of the Processing Technology

In order to have a complete process for sub-micron transistors, it is necessary not only to

make patterns in the resist material but to transfer these patterns to the underlying layers in the silicon and silicon dioxide. This has traditionally been done by wet etching processes. However, wet etching processes do not scale well into the sub-micron range. Alternatives are currently being developed which appear workable. Etching with plasmas (i.e. glow discharges of gaseous materials resulting in free ions of great chemical activity) is already used in a number of advanced processing laboratories. It is known that very controllable etching can be achieved in this way and it seems likely that essentially no wet processing will be used in the construction of sub-micron devices. Ion implantation, an ideal method for achieving controled doses of impurity ions in the silicon surface, is already a common production technique in essentially all MOS processing facilities. Metal layers for sub-micron lithographic processes must be thicker in relationship to their width than today's commercial processing technology allows. A possible solution to this problem may be the use in metal patterning of a process known as ion milling. In this process ions of modest energy sputter away any metal not covered with resist material, yielding much steeper sides on the metal thus patterned than do current wet etching processes.

It thus appears that the basic technological pieces exist to enable development of a complete patterning and wafer fabrication process at sub-micron dimensions. In reality, the ultimate submicron process will not emerge full-blown, but dimensions will gradually be reduced. As one after another of the myriad of technological difficulties are surmounted. The sketch we have given is rather an artists conception of the possibility of such an ultimate process. We do believe, however, that the evolution of this process is of fundamental importance to the entire electronics industry.

Design Rules

Perhaps the most powerful attribute of modern wafer fabrication processes is that they are *pattern independent*. That is, there is a clean separation between the processing done during wafer fabrication and the design effort which creates the patterns to be implemented. This clean separation requires a clean definition to the designer of the capabilities of the processing line. This specification usually takes the form of a set of permissible geometries which may be used by the designer with the knowledge that they are within the resolution of the process itself and that they do not violate the device physics required for proper operation of transistors and interconnections formed by the process. When reduced to their simplest form, such geometrical constraints are called *design rules*. The constraints are of the form of minimum allowable values for certain *widths*, *separations*, *extensions*, and *overlaps* of geometrical objects patterned in the various levels of a system.

As processes have improved over the years, the absolute values of the permissible sizes and spacings of various layers have become progressively smaller. There is no evidence that this trend is abating, in fact there is every reason to believe that at least another order of magnitude of linear dimensional shrinkage is possible. For this reason we present a set of design rules in dimensionless form, as constraints on the allowable ratios of certain distances to a basic length unit. The basic unit of length measurement used is equal to the fundamental resolution of the process itself. This is the distance by which a geometrical feature on any one layer may stray from another geometrical feature on the same or on another layer, all processing factors considered and an appropriate safety factor added. It is set by phenoma such as overetching, misalignment between mask levels, distortion of the silicon wafer ("runout") due to high temperature processing, over or underexposure of resist, etc. All dimensions are given in terms of this elementary distance unit. We refer to this unit as the *length-unit* λ. In 1977 the length-unit $\lambda \sim 3$ microns.

The rules given below have been abstracted from a number of processes over a range of values of λ, corresponding to different points in time at different fabrication areas. They represent somewhat of a "least common denominator " likely to be representative of nMOS design rules for a reasonable period of time, as the value of λ decreases in the future.

A typical minimum for the line width W_d of the diffused regions is 2λ, as shown in figure

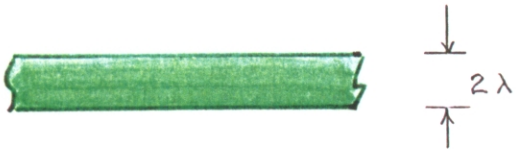


Fig.15. $W_d/\lambda \geq 2$

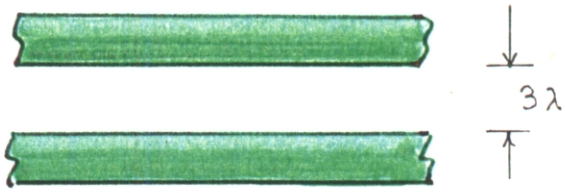


Fig.16. $S_{dd}/\lambda \geq 3$

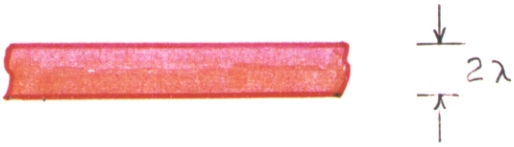


Fig.17. $W_p/\lambda \geq 2$

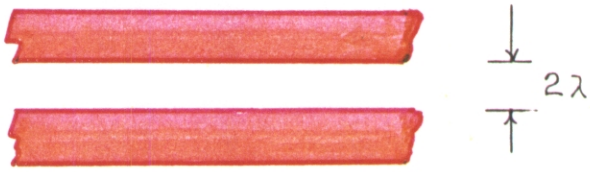


Fig.18. $S_{pp}/\lambda \geq 2$

Fig.19. $S_{pd}/\lambda \geq 1$

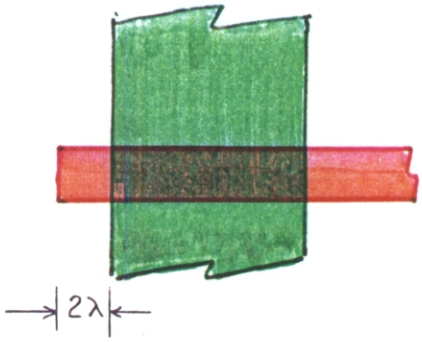


Fig.20. $E_{pd}/\lambda \geq 2$

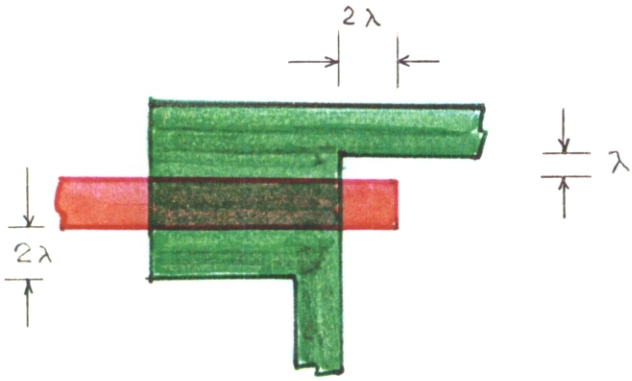


Fig.21. Example of Several Rules

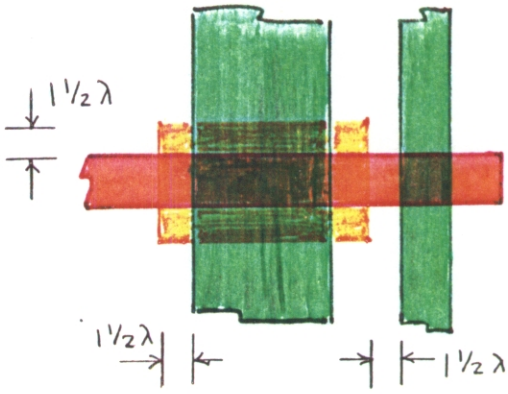


Fig.22. $S_{ig}/\lambda \geq 1\frac{1}{2}$ $E_{ig}/\lambda \geq 1\frac{1}{2}$

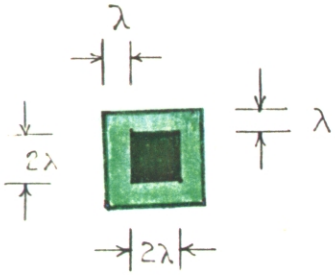


Fig.23. $W_c/\lambda \geq 2$, $E_{dc}/\lambda \geq 1$

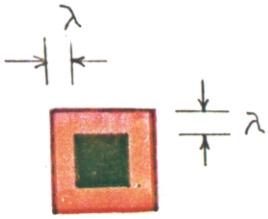


Fig.24. $E_{pc}/\lambda \geq 1$

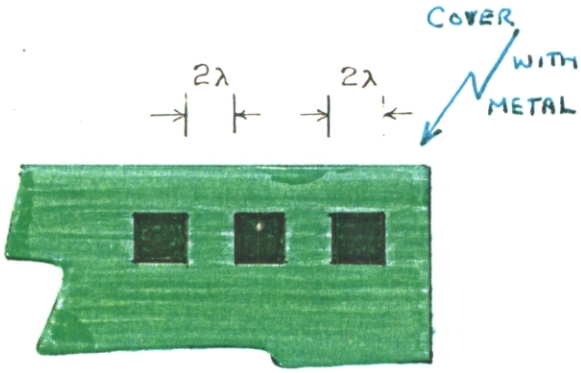


Fig.25. $S_{cc}/\lambda \geq 2$

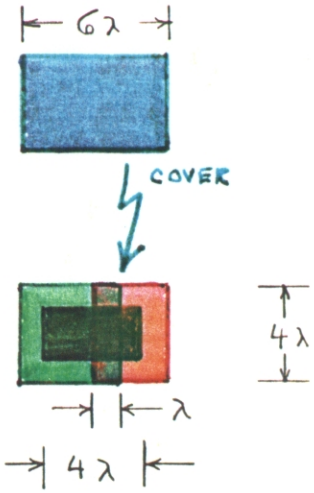


Fig.26. $O_{pd}/\lambda = 1$, & DETAILS OF BUTTING CONTACT

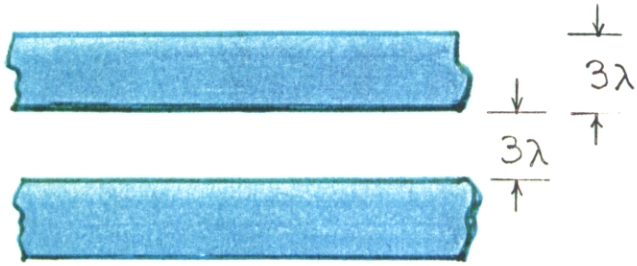


Fig.27. $W_m/\lambda \geq 3$, $S_{mm}/\lambda \geq 3$

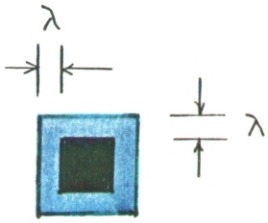


Fig.28. $E_{mc}/\lambda \geq 1$

15. The spacing required between two electrically separate diffused regions is a parameter which depends not merely upon the geometric resolution of the process, but also upon the physics of the devices formed. If two diffused regions pass too close together, the depletion layers associated with the junctions formed by these regions may overlap and result in a current flowing between the two regions when none was intended. Also, the actual geometrical boundary of the diffused region is in fact somewhat larger than the geometrical boundary of the mask due to the fact that diffusion occurs sideways as well as downwards into the silicon. To this dimension must be added the width of the depletion* layers associated with two adjacent diffused areas. * See later section on voltage variation with capacitance. In typical processes a safe rule of thumb is to allow 3λ of separation, S_{dd}, between any two diffused regions which are unconnected, as shown in figure 16. The width of a depletion layer associated with any diffused region depends upon the voltage on the region. If one of the regions is at ground potential, its depletion layer will of necessity be quite thin. In addition some processes provide a heavier doping level at the surface of the wafer between the diffused areas in order to alleviate the problem of overlap of depletion layers. In cases where either very low voltage exists on both diffused regions or where a heavily doped region has been implanted in the surface between the diffused areas, it is often possible to space diffused areas 2λ apart. However, this should not be done without carefully checking the actual process by which the design is to be fabricated.

The minimum for the width W_p of polysilicon lines is similarly 2λ. No depletion layers are associated with polysilicon lines, and therefore the separation, S_{pp}, of two such lines may be as litle as 2λ. These are illustrated in figures 17 and 18.

We have so far considered the diffused and polysilicon layers separately. Another type of design rule concerns how the two layers interact with each other. Figure 19 shows a situation where a diffused line is running parallel to an independent polysilicon line, to which it is not anywhere connected. The only consideration here is that the two unconnected lines not overlap. If they did they would form an unwanted capacitor. Avoidance of this overlap requires a separation S_{pd} of only λ between the two regions as shown in figure 19. A slightly more complex situation is shown in figure 20, where a polysilicon gate area intentionally crosses a diffused area, thereby forming a transistor. In order to make absolutely sure that the diffused region does not reach around the end of the gate and short out the drain to source path of the transistor with a thin diffused area, it is necessary for the polysilicon gate to extend a distance E_{pd} of at least 2λ beyond the nominal boundary of the diffused area, as shown in figure 20.

A composite of several of these design rules is shown in figure 21. Note that the minimum width for a diffused region applies to diffused regions formed between a normal boundary of the diffused region and an edge of a transistor as well as to a diffused line formed by two normal boundaries. This situation is illustrated in the lower left corner of the figure.

As we have seen in figure 10, ion implantation in the region which becomes the gate of a transistor will convert the resulting transistor into the depletion mode type. It is important that the implanted region extend outward beyond all four boundaries of the gate region, as shown in figure 22. To avoid any possibility that some small fraction of the transistor might remain enhancement mode, the yellow ion implantation region should extend a distance E_{ig} of at least 1½ λ beyond each edge of the gate region. The separation S_{ig} between an ion implantation region and an adjacent enhancement mode transistor gate region should also be at least 1½ λ. Both situations and their design rules are illustrated in the figure.

A contact may be formed between the metal layer and either the diffused level or the polysilicon level by means of the contact mask. A set of rules apply to the amount by which each layer must provide an area surrounding any contact to it, so that the contact opening not find its way around the layer to something unintended below it. Since no physical factors apply here other than the relative registration of two levels, a very simple set of design rules results. Each level involved in a given contact must extend beyond the outer boundary of the contact cut by λ at all points, as illustrated by extension distances E_{dc}, E_{pc}, and E_{mc}, in figures 23, 24, and 26. The contacts themselves, like the minimum width lines in the other levels, must be at least 2λ long and 2λ wide (W_c). This situation is illustrated for the diffusion and polysilicon levels in figures 23 and 24. When making contact between a large metal region and a large diffused region, many small contacts spaced 2 λ apart should be used, as shown in figure 25.

A special case arises when it is desired to connect a polysilicon layer directly to a diffused layer, using the butting contact. The detailed geometric layout of the butting contact is shown in figure 26. In its minimum sized configuration, it is composed of a square region of diffision 4 λ on a side, overlapped by a 3 λ by 4 λ rectangle of polysilicon. A rectangular contact cut, 2 λ by 4 λ in size, is made in the center of this structure. The structure is then overlayed with metal, thus connecting the polysilicon to the diffusion. The rules involved in figure 26 are identical to those given so far, with the addition of a

minimum of one λ overlap, O_{pd}, of the diffused and polysilicon layers in the center area of the contact.

In considering the design rules for the metal layer, notice that this layer in general runs over much more rugged terrain than any other level, as can be seen by referring to the cross section of figure 14. For this reason it is generally accepted practice to allow somewhat wider minimum lines and spaces for the metal layer than for the other layers. As a good working rule 3λ widths (W_m), and 3λ separations (S_{mm}) between independent metal lines should be provided, as shown in figure 27. The metal layer must surround the contact layer in much the same way that the diffused and polysilicon layers did. Since the resist material used for patterning the metal generally accumulates in the low areas of the wafer, it tends to be thicker in the neighborhood of contact than elsewhere. For this reason metal tends to be slightly larger after patterning in the vicinity of a contact than elsewhere. Thus it is generally sufficient to allow only one λ of space around the contact region for the metal, as for the other two layers. The rule for metal surrounding contacts is shown in figure 28.

The above design rules are likely to remain valid as the length-unit λ scales down in size with the passage of time. Occasionally, for specific commercial fabrication processes, some one or more of these rules may be relaxed or replaced by more complex rules, enabling slight reductions in the area of a system. While this may be important for certain competitive products such as memory systems, it has the disadvantgage of making the system design a captive of the process specific design rules. Extensive redesign and checking is required to scale down such a design as the length-unit scales down. For this reason, we recommend use of the dimensionless rules given, especially for prototype systems. Designs implemented according to these rules are easily scaled, and may have reasonable longevity.

Formal Description of Design Rules

- - - someone needs to figure this out, so we can put something in here - - -

Electrical Parameters

The design rules permit the designer to create any pattern, with the knowledge that the appropriate transistors, lines, etc., will be created by the process from his original patterns.

To complete a design it is necessary to also know the electrical parameters of the transistors, diffused layers, polysilicon layers, etc., so that the performance of the circuit can be evaluated. The resistances per square of the various layers and the capacitance per square micron with respect to underlying substrate are shown in Table 1. Note that the resistance of a square of material contacted along two opposite sides is independant of the size of the square, and equals the resistivity of the material divided by its thickness. The tabulated values are typical of processes running in 1977. As the circuit dimensions are scaled *down* by dividing by a factor α, the parameters scale approximately as shown in the table.

Resistances:	Metal	~ 0.1 ohms/□	Resistances/square scale
	Diffusion	~ 10 ohms/□	*up* by α., as dimensions
	Poly	~ 15-100 ohms/□	scale *down* by α
	Transistor	~ 10^4 ohms/□	
Capacitances:	Gate-channel	~ 4×10^{-4} pf/μ^2	Capacitances/micron^2 scale
	Diffusion	~ 0.8×10^{-4} pf/μ^2	*up* by α., as dimensions
	Poly	~ 0.4×10^{-4} pf/μ^2	scale *down* by α
	Metal	~ 0.3×10^{-4} pf/μ^2	

Table 1. Typical MOS Electrical Parameters (1977).

The relative values of resistance of metal, diffusion, poly, and drain to source paths of transistors are quite different. Diffusion and good polysilicon layers have approximately one hundred times the resistance per square area of the metal layer. A fully turned on transistor has approximately one thousand times the resistance of the diffused and polysilicon layers. The capacitances are not as wildly different as the resistances of the various layers. Compare the capacitances in Table 1 to the gate to channel capacitance, as a reference. The diffused areas typically have one fifth the capacitance per square micron. Polysilicon on thick oxide has approximately one tenth, and the metal layer slightly less than one tenth, of the gate-channel capacitance per square micron.

The relative values of the resistances and capacitances are expected to not vary dramatically as the processes evolve towards smaller dimensions. One note of warning: There is a wide range of possible values of polysilicon resistance for different commercial processes. Polycrystalline silicon suffers from inordinately high resistances at the crystal grain boundaries if the doping level in the polysilicon itself is not held quite high. This disease

does not affect the diffused layers themselves. For this reason, any processing which tends to degrade the doping levels in the diffused and polysilicon layers, affects the polysilicon resistance much more dramatically than the resistance of the diffused area. It is in general difficult to design circuits which are optimum over the entire range of polysilicon resistivity. If the designer's circuit is to be run on a variety of fabrication lines, it is desireable for the circuit to be designed in such a way that no appreciable current is drawn through a long thin line of polysilicon. In an important example in Chapter 5., polysilicon lines are used as busses along which information flows. The timing of these busses can be dramatically affected by the resistance of the polysilicon. However, the protocol used on these busses has the polysilicon lines precharged during one period of a clock and then pulled low by the appropriate bus source during a following clock period. In this way the circuit is guaranteed to work independent of the resistance of the poly. However, it may be considerable slower in processes of high poly resistivity.

Dependence of Capacitance on Voltage

In Table 1 we gave typical capacitance for the various layers to the substrate. These capacitances are those which would be measured if the voltage on the particular layer were zero (relative to the substrate). The dependence upon voltage of the capacitance of the different layers may sometimes be important and we will now discuss how this dependence arises. References R1 and R2 are good sources for those wishing more background information on the concepts of device physics used in this section and the later sections of this chapter.

When a negative voltage is applied to an n-type diffused region relative to the p-type bulk silicon, the negative electrons are pushed out of the n-type layer into the bulk and a current flows. In integrated systems we are careful to never allow the voltage on the n-type diffused regions to be more negative than the p-type bulk. Diffused regions are biased positively with respect to the p-type bulk, resulting in a reversed biased p/n junction. With the exception of a small leakage current, the reverse biased p/n junction acts merely to isolate one diffused region from another. The p-type bulk of our integrated system has a small number (typically 10^{15}-10^{16} per cubic centimeter) of impurity atoms. When a voltage is applied to an n-type diffused region, it's influence is felt well out into the p-type bulk. Positive charge carriers in the p-type bulk and repelled from the positively charged n-type layer, thereby exposing negatively charged impurity ions. The region surrounding

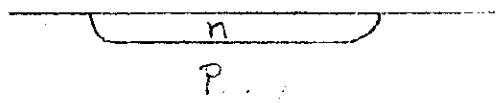


Fig. 29a. n-Type Diffusion in p-Type Bulk Silicon

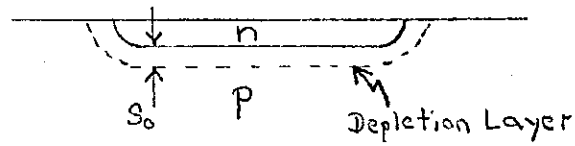


Fig. 29b. Depletion Layer

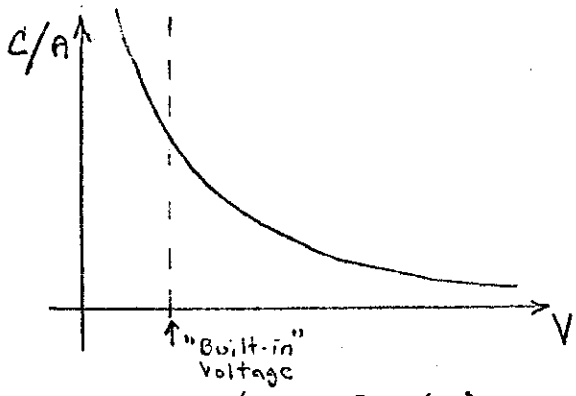


Fig. 30. C/A as fcn(V)

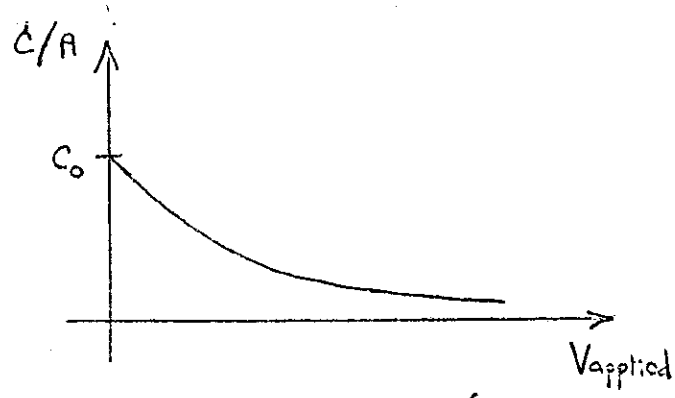


Fig. 31. C/A as fcn($V_{applied}$)

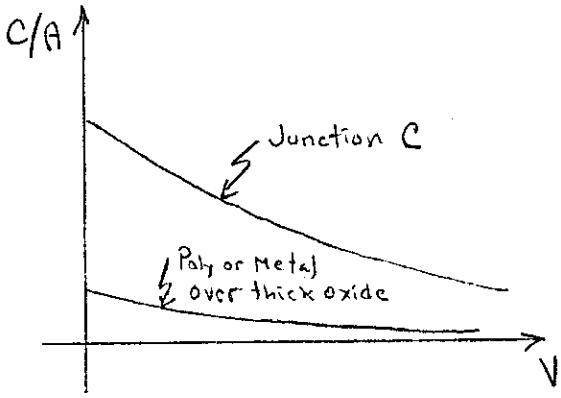


Fig. 32. Capacitance of Poly or Metal over Thick Oxide

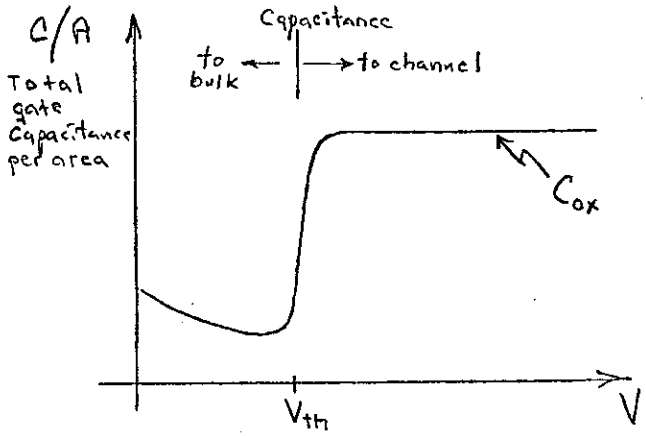


Fig. 33. MOS Gate Capacitance as fcn V

the n-type diffused layer which has been depleted of positive charge carriers is referred to as a depletion layer and is shown schematically in Figure 29b. As the voltage on the n-type layer is increased, charge carriers are pushed further back from the junction between the n-type layer into the p-type bulk, widening the depletion layer and exposing more charged impurity ions. The charge thus induced in the depletion layer as the voltage on the n-type diffused region is increased is responsible for the capacitance of the n-type diffused region relative to the substrate.

We will now consider a unit area of the junction. The total charge in the depletion layer per unit area is proportional to the number per unit volume of impurity ions in the bulk (N), and the width, s_0, of the depletion layer.

$$\text{Total charge/area} \propto Ns_0$$

The electric field in the region is proportional to the charge per unit area.

$$\text{Electric field} \propto \text{charge/area} \propto Ns_0$$

The voltage between the n-type diffused layer and the p-type bulk on the far side of the depletion layer is proportional to the electric field times thickness of the depletion layer, and therefore to the density of negatively charged ions in the depletion layer times the square of the width of the depletion layer.

$$\text{Voltage} \propto \text{electric field} \times s_0 \propto Ns_0^2$$

The capacitance per unit area is just the charge per unit area divided by the voltage across the depletion layer. From the above equations the capacitance is proportional to the square root of the density of impurity atoms in the p-type bulk divided by the voltage on the across the junction.

$$\text{Capacitance/area} = Q/V \propto 1/s_0 \propto (N/V)^{1/2}$$

This relationship is plotted in Figure 30. Notice that the capacitance tends towards infinity as the voltage across the junction tends to zero. It would seem that this large capacitance would be disastrous for the performance of our integrated systems. However, this proves not to be the case. When the p/n junction was formed the n-type region had an excess of

negative charge carriers while the p-type bulk had an excess of positive charge carriers. When the two were brought together to form the junction, there was no voltage to prevent charge carriers of either type from flowing over into the opposite region. This initial flow caused the n-type layer to become more positive than the p-type layer. This flow ceased when just enough voltage built up to stop. In silicon the voltage required to prevent the flow of charge carriers in such a situation is approximately 0.7 volt. Thus the true voltage across the junction is this initial "built-in" voltage plus the voltage we apply in our circuit. The variation of the capacitance per unit area with applied voltage is shown in Figure 31. An approximate equation which can be used to calculate the junction capacitance C_j per unit area of diffused layers as a function of the applied voltage is given by.

$$C_j = 4.5\times10^{-12}[N/(V+0.7)]^{1/2} \ pf/\mu^2$$

In this equation, N, the density of impurity ions in the p-type bulk, should be given in number per cubic centimeter. The voltage is in volts and the capacitance per unit area is evaluated in picofarads per square micron. This equation is adequate for most design purposes.

Aside from the diffused regions, there are two other situations where the capacitance is of interest. The first is poly or metal over thick oxide and the second is the gate of an MOS transistor. We will discuss poly or metal over oxide first. Figure 32 illustrates once more the capacitance per unit area of a junction over the p-type bulk. If the poly or metal layer was lain on an oxide much thinner than the depletion layer, its capacitance would be nearly the same as that of the corresponding p/n junction. However, if an oxide is interposed whose thickness is of the order of the depletion layer thickness, the capacitance of the poly or metal line will be decreased. The formula which applies in this case is given by:

$$1/C_{total} = 1/C_j + 1/C_{ox}$$

A typical dependence is shown in Figure 32. For an oxide thickness d, $C_{ox} = 3.5\times10^{-2}/d$, where the thickness d is given in angstrom units (10^{-4} microns), and the result is in picofarads per square micron as before.

The most spectacular voltage dependence of a capacitance in the technology we will be using is that of the gate of an MOS transistor. When the gate voltage V_{gs} is less than the threshold voltage V_{th}, the capacitance of the gate to the bulk is just that given above for

metal or poly over oxide since the voltage on the gate merely depletes positive charge carriers back from the channel area. However, when the voltage on the gate reaches the threshold voltage of the transistor, negative charge is brought in under the gate oxide from the source of the transistor and the capacitance changes abruptly from the small valve associated with depleting charges in the bulk to the much larger oxide capacitance between the gate and the channel region. Further increase in voltage on the gate merely increases the amount of mobile charge under the gate oxide with no change in the width of the depletion layer underneath the channel. Hence, the character of the gate capacitance changes abruptly as the gate voltage passes through the threshold voltage.

The dependence of the total gate capacitance on gate voltage is shown in Figure 33. The capacitance from channel to bulk is completely separate from the gate to channel capacitance. It is associated with the depletion layer underneath the channel region, and is almost identical to that of a diffused region of the same area. When the gate voltage is below the threshold voltage, the gate to channel capacitance disappears altogether and we are left only with the small parasitic overlap capacitances between the gate and the source and drain regions.

Current Limitations in Conductors

One limit which is not covered in either the design rules or the electrical parameters section is that associated with the maximum currents through metal conductors. There is a physical process called *metal migration* whereby a current flux through a metal conductor, exceeding a certain limit, causes the metal atoms to move slowly in the direction of the current. If there is a small constriction in the metal, the current density will be higher and therefore more metal atoms will be carried forward from that point, narrowing the point still more. Hence, metal migration is a destructive mechanism causing open circuits in the metal layer carrying heavy currents.

For metals like aluminum this limit is a few times 10^5 amperes per square centimeter. This limit does not interfere too drastically with the design of integrated systems in current MOS technologies. However, many metal conductors in present integrated systems are operated near their current limit, and currents do not scale well as the individual elements are made smaller. Applying the scaling rules developed earlier, we found that the power per unit area is independent of the scale down ratio. However, the supply voltage decreases and therefore

the current per unit area increases as the devices are scaled down. For this reason it will not be possible to use processes for very large scale integrated systems where the metal thickness scales in the same way as other dimensions in the circuit. Much work will likely be done to develop processes enabling fabrication of metal lines of greater height relative to width than is presently possible.

Short pulses of current are known to contribute much less to metal migration than steady direct current. Nanosecond pulses of currents two orders of magnitude higher than the dc limit mentioned above may be carried in metal conductors without apparent damage. Therefore, switching current may not be as damaging to metal conductors as a steady current.

These effects strongly favor processes like CMOS which do not require static dc current, and favor design methodologies which maximize system function per unit dc current.

Choice of Technology

Before proceeding to the chapters on system design, let us briefly examine some alternative technologies. Then, using the knowledge developed in these first two chapters, we will discuss the reasons for selecting nMOS as the single technology used to illustrate LSI systems in this text. Some of the factors which must be considered in choosing a technology include circuit density, richness of available circuit functions, performance per unit power, the topological properties of circuit interconnection paths, suitablility for total system implementation, and general availability of processing facilities.

As the technology advances, more system modules can be placed on the same sized chip. An ultimate goal is the fabrication of large scale systems on single chips of silicon. For this goal to be attained, it is necessary that any signal which is required in the system other than inputs, outputs, VDD, and GND, must be generated in the technology on the chip. In other words no subsystem can require a different technology for the generation of its internal signals. Thus such technologies as magnetic bubbles are ruled out for full integrated systems because they are not able to create the signals required for all operations in the on-chip medium.

We believe that for any silicon technology to implement practical large scale systems, it

must provide two kinds of transistor. The rationale for this observation is as follows. In order to provide some kind of nonlinear threshold phenomenon there needs to be a transistor which is normally off at the lowest voltage used in the system. Bipolar technologies use NPN transistors for this purpose. nMOS technology uses n-channel enhancement mode devices. In addition to the transistor which is normally off when its control element is at the lowest voltage, a separate type of transistor needs to be supplied to allow the output of a driver device to reach the highest voltage in the circuit (VDD). In the bipolar technologies PNP lateral devices are used to supply this function, in the n-channel technology a depletion mode device is used, and in complementary MOS technology a p-channel enhancement mode device is used. Any of these three choices allow output voltages of drivers to reach VDD and thus meet the criterion mentioned above.

To date there have emerged three technologies which are reasonably high in density and scale to submicron dimensions without an explosion in the power per unit area required for their operation. These are the n-channel silicon gate process, the complementary MOS silicon gate process, and the integrated injection logic, or I^2L, process[2]. Although present forms of the I^2L technology lack an additional level of interconnect which is available in the silicon gate technologies, there is no inherent reason that such a level could not be provided. It is important to note that the types ot array functions which can be created in a technology are greatly enhanced by increased flexibility of interconnect. I^2L has the advantage over nMOS that the power per unit area (and hence the effective τ of its elementary logic functions) can be controlled by an off chip voltage. The decision concerning what point on the speed vs power curve to operate may thus be postponed until the time of application (or even changed dynamically).

The nMOS scaling has been described previously. Any technology in which a capacitive layer on the surface unduces a charge in transit under it to form the current control "transistor" will scale in the same way. Examples include Schottky Barrier Gate FET's (MESFETs), Junction FET's, and CMOS.

There are certain MOS processes (VMOS, DMOS) of an intermediate form in which the channel length is determined by diffusion profiles. While competitive at present feature sizes, these likely are interim technologies which will present no particular advantage at submicron feature sizes.

The scaling of the bipolar technology[1] is quite different from that of the MOS technologies.

For completeness, we include here a discussion of the scaling of bipolar devices, which may be of interest to those familiar with those technologies.

Traditionally, bipolar circuits have been "fast" because their transit time was determined by the narrow base width of the bipolar devices. In the 1950's, technologists learned how to construct narrow bipolar transistor base regions by forming them as the difference between two impurity diffusion profiles. This technique allowed very precise control of the distance perpendicular to the silicon surface, and therefore permitted the construction of very thin base regions with correspondingly short transit times. Since current in a bipolar device flows perpendicular to the surface, both the current and the capacitance of such devices are decreased by the same factor as the device surface dimensions are scaled down, resulting in no change in delay time performance. The base widths of high performance bipolar devices is already as thin as device physics allows. For this reason, the delay times of bipolar devices is expected to remain approximately constant as their surface dimensions are scaled down.

The properties of bipolar devices may be analyzed as follows. The collector current is due to the diffusion of electrons from emitter to collector. For a minority carrier density N(x) varying linearly with distance x, from N_0 at the emitter to zero at the collector (where x=d), the current per unit area is:

$$I/A = q(2D)dN/dx = q(2D)N_0/d = q(2kT/q)\mu N_0/d \qquad \text{[eq.2]}$$

where the diffusion constant $D = \mu kT/q$. The factor of two multiplies the diffusion constant in eq.2 because high performance bipolar devices operate at high injection level (injected minority carrier density much greater than equilibrium majority carrier density). Now, the inherent stored charge in the base region is:

$$Q/A = N_0 d/2 \qquad \text{[eq.3]}$$

Therefore, the transit time is

$$\tau = Q/I = d^2/[4\mu kT/q] \qquad \text{[eq.4]}$$

The form of equation 4 is exactly the same as that for MOS devices (eq.1., chapter 1.), with the voltage in the bipolar case being equal to 4kT/q (at room temperature kT/q = 0.025

volts). A direct comparison of the transit times is shown in Table 2.

Table 2. Transit Time:

$$\tau = (\text{Distance})^2/(\text{Mobility} \times \text{Voltage})$$

	MOSFET:	MESFET, JFET:	Bipolar:
Distance:	channel length	channel length	base width
Voltage:	~ VDD/2, many kT/q	~ VDD/2, many kT/q	4kT/q
Mobility: cm^2/v-sec, (Si)	surface mobility, ~800	bulk mobility, ~1300	bulk mobility, ~1300

At the smallest dimensions to which devices can be scaled, the base width of bipolar devices and the channel length of FET devices are limited by the same basic set of physical constraints, and are therefore similar in dimension. The voltage on the FET devices must be many times kT/q to achieve the required nonlinearity. Hence at the ultimate limiting small dimensions the two types of devices have roughly equivalent transit times. At these limiting dimensions, choices between competing technologies will be made primarily on the grounds of the topological properties of their interconnects, the functional richness of their basic circuits, simplicity of process, and ability to control dc current per unit area. As supply voltages are scaled down to the 1 volt range, MOS devices become similar in most respects to other FET type devices, and it is possible that mixed forms (MOS-JFET, MOS-MESFET, Bipolar-MESFET, etc.) may emerge as the ultimate integrated system technologies.

We have chosen to illustrate this monograph with examples drawn from the n-channel silicon gate depletion mode load technology. The reasons for this choice in 1977 are quite clear. In addition to meeting the required technical criteria we have described, this technology provides some important practical advantages to the student and to the teacher. It is the only high density technology which has achieved universal acceptance across company and product boundaries. The reader who wishes to implement his designs can go to essentially any company involved in fabricating silicon wafers and run his parts without fear that slight changes in the process or the vagaries of relationships with a particular firm will cut off his source of supply. It is also presently the highest density process available. This certainty of access to fabrication lines, the more generally widespread knowledge of

nMOS technology among members of the technical community, its density, and its performance similarity with bipolar technology in its ultimate scaling, are all important factors supporting its choice for this text on LSI Systems. However, the principles and techniques developed in this text can be applied to essentially any technology.

References

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

2. F. M. Klaassen, "Device Physics of Integrated Injection Logic", *IEEE Transactions on Electron Devices*, March 1975, pp. 145- , and cited papers by Hart & Slob, and by Berger & Weidmann.

Reading References

R1. A. S. Grove, "Physics and Technology of Semiconductor Devices", J. Wiley and Sons, 1967, is the early classic text on process technology and device physics.

R2. W. G. Oldham, "The Fabrication of Microelectronic Circuits", *Scientific American*, September, 1977, provides an excellent overview of the fabrication process.

R3. I. E. Sutherland, C. A. Mead, T. E. Everhart, "Basic Limitations in Microcircuit Fabrication Technology", ARPA Report R-1956-ARPA, November, 1976, contains a quantitative discussion of the many limiting factors in fabrication.

R4. R. S. Muller, T. I. Kamin, "Device Electronics for Integrated Circuits", Wiley, 1977, provides insight into the device physics relevant to current integrated circuit practice.