CHAPTER 3
SILICON PATTERNING

MOS integrated circuits are constructed as a series of patterned layers on the surface of a silicon wafer. The generation of the *masks* used in patterning and the *wafer fabrication* process itself are complex, requiring special equipment and considerable expertise. The researcher who wishes to have his designs cast in silicon need not be concerned with all of the details of mask generation and wafer fab, yet he must be familiar enough with both to effectively deal with the vendors of those services.

3.1 An Introduction to Photolithography

The layers of a MOS IC are patterned by a *photolithographic* process. The layer-making process begins with the deposition or growth of some material, for example silicon dioxide, polysilicon, or metal, on the surface of the wafer. That material is coated with a thin layer of photosensitive chemicals, called *photoresist*, and exposed to ultraviolet light through a *mask*, which is a sheet of glass large enough to cover the silicon wafer. (The mask is coated on one side with opaque material that has been patterned to define certain areas on the particular layer.) If *negative photoresist* was used, those areas of resist that were exposed to light will be hardened, while *positive photoresist* is softened in the exposed areas. The exposure can take place with the mask pressed against the wafer (*contact photolithography*) or by projecting an image of the mask onto the wafer (*projection photolithography*). Projection techniques are becoming more widely used in spite of the extra equipment and maintenance needed since the masks are subject to less wear and contamination than contact masks. Consequently masks last longer and it is easier to control certain kinds of defects incurred in the photolithography steps.

Following exposure, the resist is *developed* by immersing it in a solvent that dissolves the unexposed (for negative resist) or exposed (for positive resist) portions, leaving the desired pattern. The patterned resist is hardened by baking at a low temperature and is then used to protect the covered areas of the wafer during the *etching* process. Two methods are common: in the older *wet etching* process the wafer is immersed in a bath of chemical etchant under controlled temperature conditions for a specific amount of time. Wet etching depends on the availability of an etchant that will dissolve the layer beneath the photoresist, yet not significantly attack the resist. For some materials, for example silicon nitride, the wet etchants dissolve photoresist as well as the desired material. Such materials require an intermediate pattern to be formed in another layer that serves as the actual etching mask. The intermediate material must be amenable to wet etching, yet resist the etchant for the layer beneath. In the case of silicon nitride, silicon dioxide is a suitable intermediate layer. *Plasma etching*, a *dry etching* technique, utilizes a stream of ions and electrons to blast away material. Plasma etching gives better results for fine geometries and also permits the direct use of resist as an etching mask.

After the etching step the remaining resist is removed, leaving a pattern in the underlying material. This sequence is repeated for the various layers of the circuit. About six photolithography/etching cycles are required to build up a typical Si gate NMOS circuit. The entire process entails over forty individual steps, outlined in section 3.3.

3.2 Mask Generation

There are two readily available techniques for generating masks: *Optical* and *Electron-beam* (named for the the form of energy used to expose the plates). Both of these methods lead to a set of *master plates*, which may be used to pattern the wafers directly. More commonly, a secondary set of plates called *working plates*, is used in the actual photolithography process. Working plates are printed directly from the masters, thus allowing one set of masters to be used to produce many wafers.

Of optical and E-beam mask generation, optical mask generation is the older of the two processes, offering low cost and wide availability. The *pattern generation* process (see below) is slow, however, and its speed is directly related to the complexity of the design. Because of the difficulty in controlling alignment during the step and repeat process and the danger of reticle defects, it is costly to include more than two different chip types on the same set of working plates. Electron-beam mask generation is free of these shortcomings, but is not yet widely available. The flexibility and fast turnaround afforded by E-beam closely matches the requirements of many research institutions.

3.2.1 Optically Generated Masters

The first step in creating a mask is plotting the files provided by the designer on a photosensitized glass plate. This first plate, called a *reticle*, differs from a master or working plate in that it contains only one copy of the relevant chip layer and is plotted at 10x the actual size of the chip. The plotting process takes place in a *pattern generator*. Typical of such machines is the Mann 3000, which projects (*flashes*) the image of a variable size rectangle on the reticle. The input to the machine is the size of the rectangle, or *aperture*, the x and y coordinates of the center and the angle with respect to the x axis.

The nature of the reticle making process has a number of important implications for the designer. All shapes on the masks must be decomposed into simple rectangles. It should be noted that exposure time has a definite effect on the feature sizes on the reticle, in particular overexposed areas tend to "grow" slightly; for this reason the designer should avoid substantial overlap between flashes. The pattern generation process involves complex mechanical motion; proper sorting of the individual rectangles of which the chip is composed can speed up the pattern generation process considerably -- and thus lower the price (the bulk of the cost of optical mask generation is PG cost). For example a Mann 3000 PG machine is fastest at moving in the x direction, followed by aperture change, motion in the y direction, and finally, angle change. Unfortunately the optimum order is based on a complex function that depends on mechanical considerations as well as the pattern being flashed; in general this function is not known to the designer. Unless the designer has detailed knowledge about the PG machine being used, he is probably better off using a simple sorting algorithm (for instance lexicographic ordering based on what the particular PG machine is fastest at) than trying to second guess the pattern generator.

Optically generated masters require several special features on the reticle that are used during intermediate steps in mask making. A *parity mark*, consisting of an arrow or triangle, is sometimes included on each mask layer to help the operator orient the mask. The mark is placed outside of the boundary of the chip pattern. *Fiducials* are small crosses which also appear on each layer outside of the boundaries of the chip. These are used in the *step and repeat* process (see below). Often the parity marks and fiducials are provided by the mask house thus making it unnecessary and undesirable for the designer to supply them. Parity marks and fiducials appear only on the reticles and not on the finished master plates.

The reticles are used to make a set of master plates in a step and repeat machine that projects an image of the reticle (reduced 10x) onto a photosensitized plate. By precisely stepping the image across the master a matrix of images of the reticle is created. The fiducials are used to control the distance between exposures and to align the reticle images relative to one another. It is possible to interstep two different reticles on the same master, but it becomes increasingly difficult as the number of reticles goes up. We have not found any manufacturers willing to guarantee alignment specs for more that 3 reticles.

3.2.2 E-Beam Masters

As in the optical process, electron-beam mask generation equipment can be used to create reticles that are stepped and repeated on a master plate. More commonly, however, an entire master plate is written in one step. E-beam masks offer several advantages to researchers interested in fast turnaround: one-step mask generation (if the masters are used to directly expose the wafers), speed, flexibility, and reduced defects from certain causes. For instance, a defect on a reticle means that each and every chip will have the same defect, in addition, the step and repeat process is a potential source of defects (for example, alignment problems, defects from dust specks). Both of these problem areas are eliminated with e-beam masters.

Unlike optical pattern generation equipment, electron-beam exposure systems are raster oriented. The mask can be visualized as a piece of graph paper, where the squares are the same size as the e-beam diameter (typically 0.25μ or 0.5μ). All geometric data is ultimately converted into a *bitmap* (a rectangular array of 1's and 0's), which is placed on top of the graph paper — the squares containing 1's are exposed, those containing 0's, not. Conceptually, the exposures are made by sweeping the electron beam in a repeating "S" pattern from the lower left-hand corner of the mask, *blanking* and *unblanking* the beam according to the input stream of bits. To a first approximation, the beam visits each point on the mask regardless of whether the point is exposed, and so the *writing* time is independent of the design complexity. (In practice, this is not entirely true. Some machines are programmed to skip large blank areas, and so take less time to write sparse designs.)

For practical reasons, the writing sequence is not quite that straightforward. Assume that we wish to write and array of 8 identical chips (refer to Figure 3.2.1). The chip is divided into horizontal strips of fixed height and the geometric shapes within each strip are fractured into rectangles and trapezoids (or approximated by same). Software is available to convert conventional PG formats to this e-beam format, or the designer can generate the trapezoids and rectangles directly. The location of each strip of the chip, in this case there are four strips, along with other information is used to create a command sequence for writing the array.

The first step in writing is converting the trapezoids and rectangles for a given strip into a bitmap, this process, called *corefill* (because the bitmap is loaded into core) is relatively time consuming. For this reason, it is only done once for a given strip. The machine then writes every area on the mask that is covered by that strip, before it converts another. In our simple example the machine would write identical strips a,b,c,d,e,f,g,h in that order, then convert the next strip and continue the process. Mechanically, the mask (affixed to a *stage*) is moved in the x direction, while the electron beam scans in the y direction along short *scan lines.*

For more complex arrays, the only penalty paid is in corefill time, since the writing time is more or less constant. If chips 1,3,5,7 (the odd group) are identical to each other but different from the chips in the even group the machine might first corefill with the bottom strip of the odd chip. Strips a,d,f,g are written in that order. Corefill would proceed with the bottom strip of the even chip, and then strips b,c,e,h would be written. Using this technique, have combined as many as 8 different chip types on the same set of masks. Such an undertaking would be impossible if optical masks were employed. Aside from the great expense of generating 8 reticles, each reticle would have to be perfectly aligned through 8 step and repeat cycles.

3.2.3 Working Plates

When needed, working plates can be made from the masters by contact printing. In cases where a large number of working plates are required the mask house may make several sets of

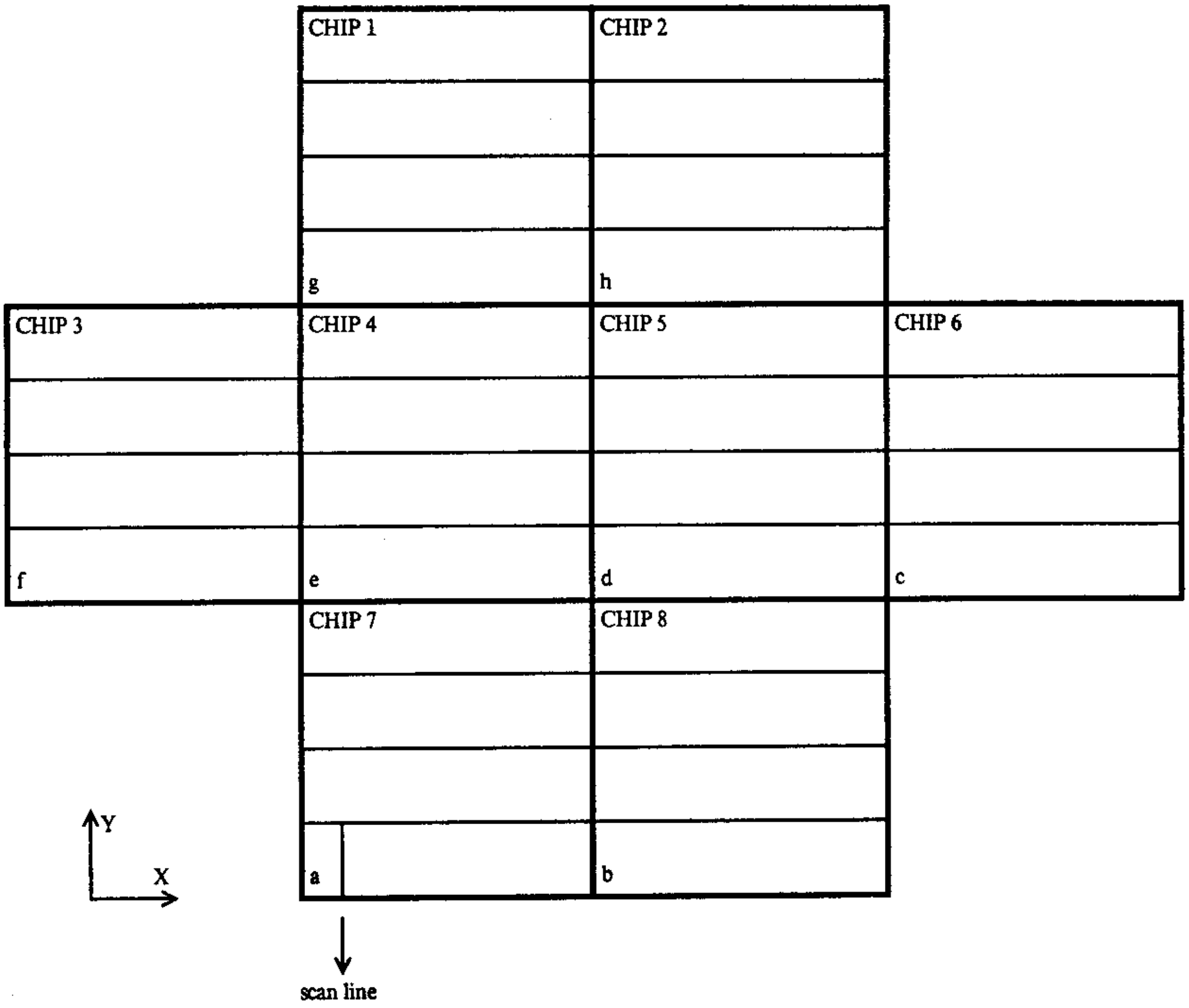


Figure 3.2.1 E-Beam Writing Sequence

(after [Varian 1979])

submasters and print the working plates from them.

To improve the quality of the masks, reduce cost, and shorten turnaround time, it may be possible to use the masters directly for wafer fabrication. This approach is particularly attractive with masters generated on electron-beam exposure systems since it reduces mask making to a one-step process with a possible turnaround time of a few days. Once the decision is made to use master plates directly, it is generally not possible to have working plates made from those masters. This is because the copying process causes the dimensions on the copies to differ from the masters, so the mask house must compensate for these changes in advance (i.e. as the masters are being made).

3.2.4 Mask Specification

The researcher is faced with the task of specifying many details so that the masks will be made correctly. Appendix A contains copies of instructions that we have sent to mask houses, providing an overview of the type of data needed. It is essential that those wishing to have wafers fabricated understand the mask generation and wafer fabrication processes in enough detail to make reasonable decisions. The following paragraphs briefly cover some tradeoffs and decisions the researcher must make.

Mask *polarity* — whether the plates for each layer should be *opaque field* (clear features) or *clear field* (opaque features) – must be specified by whomever orders the plates. The requisite polarity for each mask is specified by the fab line; typically a mixture will be called for. The choice depends on the process step and the type of photoresist used. The photoresist is selected partially on the basis of requisite linewidths for the fabrication process. More important, however, is the field area involved with the particular plate. A speck of dust on an otherwise clear area of the working plate will cause a pattern to be made in the photoresist. If negative resist is being used the speck will make a hole in the resist that will enlarge somewhat due to undercutting in the subsequent etching step. Positive resist will leave a small dot where the speck was; this dot will probably be etched into oblivion. The fabrication line decides which of these factors to trade off in choosing the polarity of the working plates.

The opaque material covering the plates may be photographic emulsion, iron oxide, or chromium. Emulsion is the least expensive but relatively easily damaged in the contact photolithography process. Chromium and iron oxide give better line resolution and are very hard, but they are more expensive. The glass plate itself is available in various grades, the least expensive being *"green glass"*. *Low-expansion glass* yields masks that are more stable, but may run several times the cost of green glass. There are also several options in glass thickness. All of the alternatives are provided so that users may choose masks that will provide the best performance (which includes cost, defect density, repeatability, etc.) for their application.

The plate specifications may be dictated by the fabrication line, as they may be used to working with a particular type of plate. In this case the researcher has little choice but to order (and pay for) whatever the fab line requires. In the event that the fab line has no preference, a low-cost option may be quite adequate, since masks for a research chip set are rarely used for more that one run of wafers.

Often the mask house or fab people require some special features to be included on the mask set. *Critical dimensions* (CD's) are simple lines or crosses of a fixed size appearing on each layer; they are used by the mask house to adjust exposure and developing time to insure that these marks and hence other features on the mask are the correct size. Additional features to be put on each mask may include an identification code for the process step. The fabrication line may have specific codes that they wish placed on the masks. In order to register the layers during wafer fabrication, a set of *alignment marks* (see section 4.3) must be included on the masks. The alignment marks may be a specific set for the fab line or may be designed by the researcher.

The seemingly simple concept of mask *parity*, which specifies up from down and left from right, has turned out to be quite confusing to keep straight in practice. A number of factors confound the issue:

1. Some mask generation equipment uses a left handed coordinate system (e.g. Mann) others, right handed.
2. The initial data may be reversed (mirrored) a number of times, depending upon the details of the process used to arrive at the working plates. Thus, working plates may come out reversed, even though the input data was not.
3. The same mask manufactures may make masks for digital and analog circuits in MOS, bipolar and other technologies. Thus the natural frame of reference for the designer is largely irrelevant to the mask house, and he should not count on the mask people having any intuition about the orientation of shapes on the masks.

We have found that the most effective way to specify mask parity is to include some text in the same location on each layer, and to instruct the mask house as to how the text should appear on the final plates (be they working plates or masters). Usually up and down is not an issue (the chip can be rotated 180^{o}), so two things must be specified. The first is whether the text is to appear *right reading* or *wrong reading* (normal or reversed), and the second is which side of the plate the text should be viewed from (chrome side or non-chrome side). A MOS wafer processed with masks whose text is WRONG reading when viewed from the chrome side will have features normally oriented. Other schemes for specifying parity are possible, for example arrows that "point upward and left", but few are as unambiguous as a string of text.

It is important to verify the correctness of masks as much as possible. When optical mask generation is used, it is usually possible to order color enlargements of each reticle; these *blowbacks* are typically about 100x-150x actual (chip) size. The layers can be checked individually and in combination by superimposing the films on one another. *Black and clear* transparencies (usually

8½" x 11") may also be made at the same time. They are sometimes used in the interaction between the operators on the fab line and the designer to indicate the location of features on the mask such as alignment marks. At the present time film blowbacks are not generally available from E-beam masks. Large checkplots are the only recourse and provide a reasonable means of checking individual layers, but are not particularly helpful in checking combinations.

Before the mag tape can be sent off to the mask house some information must be obtained from the fabrication line regarding their process. Varying etch conditions may cause the fab line to request that features on the masks for certain layers be altered (i.e. stretched or shrunk) by a constant amount, for example 0.5 micron around any border, in order to produce the desired dimensions on the silicon. These dimensional adjustments can be made in one of three ways:

1. The circuit designer can be required to change his design to take into account the over- or under-etching at the fabrication line. This entails considerable work on the part of the designer each time the circuit is implemented on a different fab line, but has the advantage that the designer retains complete control of the layout geometry.
2. Software could be provided to input the original design file and produce a new design file that had the borders of features expanded or contracted in the appropriate way. This approach may require the use of complex algorithms in order to correctly modify the original file since minimum spacing design rules may be violated by enlarging adjacent features while gaps and discontinuities may be introduced by shrinking features which abut in the original design.
3. The mask house may be able to effect the changes by adjusting exposure time and other parameters in the mask generation process.

Once all of this information has been collected and reduced to a set of files on mag tape and some written instructions, the mask house takes over. When the masks are returned they are passed in turn to the fabrication line along with more instructions. The total elapsed time for mask making and wafer fabrication can be 8-12 weeks. During this time the designer should be preparing for the day when finished wafers are delivered.

3.3 Wafer Fabrication

A number of wafer fabrication lines offer "standard" n-channel Si-gate MOS processes. In principle, one need only produce a set of masks that is compatible with the fab process and about $3,000, and the fab line will deliver a *minimum run* (about 20) of finished wafers. The availablility of these processes allows IC designers to take a black box view of NMOS processing.

Complementary MOS may soon reach the same "stable" state. Today several firms offer CMOS processes, but they vary widely. In addition, many low-level tradeoffs may be made in

CMOS, for example a particular gate may be implemented in complementary form or using n-channel or p-channel devices exclusively. Another problem arising in current CMOS processes is that the layout that most cleanly reflects the functional topology of the system often has to be distorted in order to move transistors of the same polarity close together. This is so that they can be placed into an area of the same substrate polarity. These factors make it difficult to design independently of the fab line, and to provide an overview of the process.

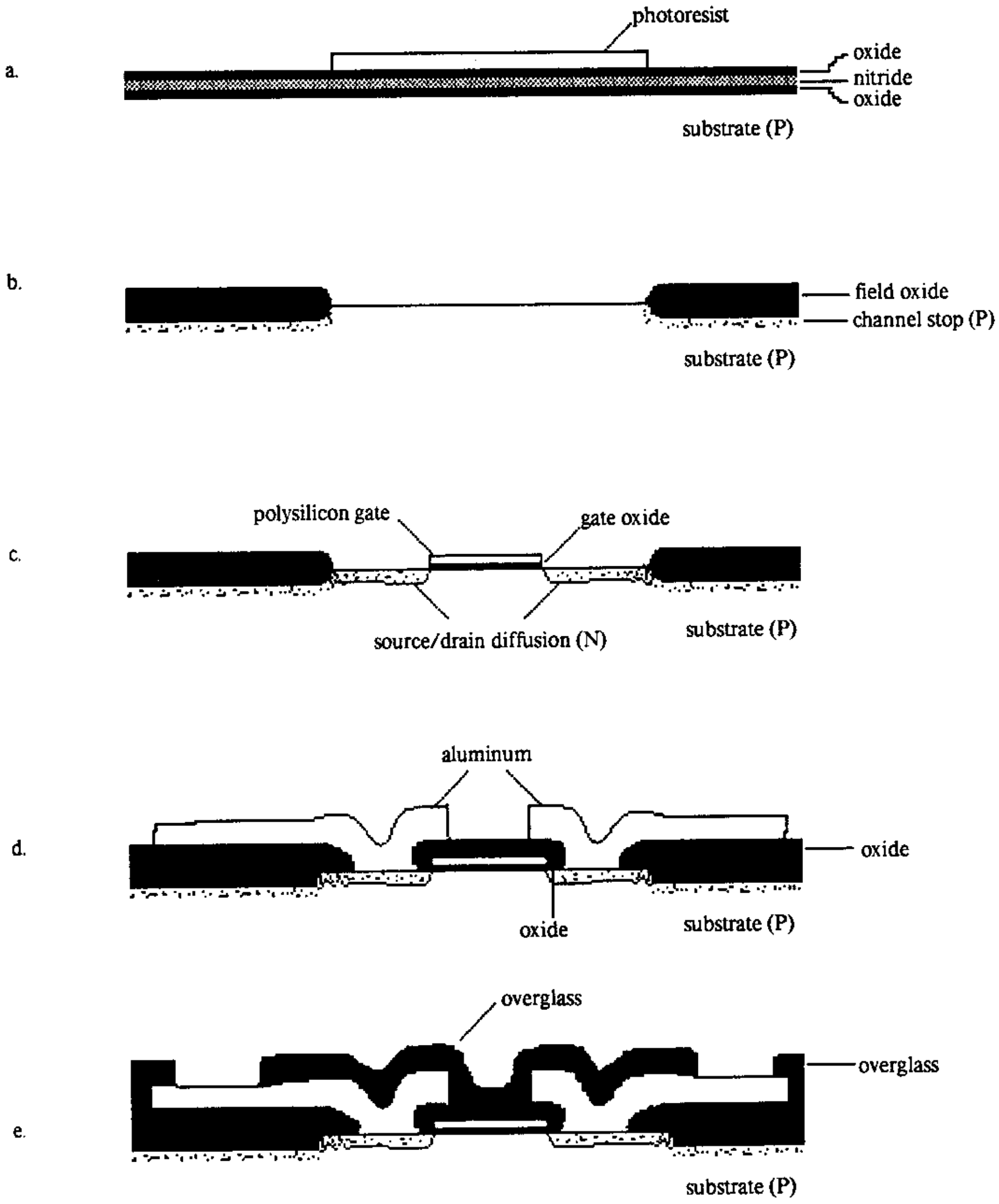
3.3.1 The Si-gate NMOS Process

The process discussed here is a standard Si-gate n-channel MOS process, such as available from a number of fabrication lines. The reader need not be concerned with learning all of the details of the process, indeed most of the decisions concerning processing are made by the fabrication line. The designer may not know which particular techniques the fab line uses, and may not care as long as his standard circuits exhibit normal performance. This section is presented to provide background for those interested in what really happens behind the clean room doors. It may be helpful for the reader to make a sketch of the state of the wafer after each step as this section is covered.

The wafer of *p-type* (*100 crystal orientation* for lowest interface state density) silicon is scrubbed and a thin layer of silicon dioxide (hereafter called "oxide") is thermally grown on the surface. This layer serves as a mechanical buffer zone for the silicon nitride (Si_3N_4) that follows. The buffer zone is needed to relieve stress caused by differences in the coefficients of thermal expansion of silicon and silicon nitride. A layer of Si_3N_4 is deposited by *chemical vapor deposition*, then another layer of oxide is grown. Photoresist is applied over the entire surface and the wafer is exposed to ultraviolet light through the *diffusion layer* mask. The resist is developed, leaving open areas over the *field* region (figure 3.3.1a). The top layer of oxide is etched away wherever there is no photoresist using a hydrofluoric acid solution. After the resist is removed this top layer of oxide is used as a mask for patterning the nitride since the photoresist alone will not stand up to the chemicals used in the wet etching of silicon nitride. A third etching step is used to remove the bottom layer of oxide. *Ion implantation* is used to place the *channel stop* region and a thick *field oxide* is grown over those areas. The field oxide and the channel stop are *self-aligned* with respect to the source/drain diffused areas (the nitride covers the source/drain areas during channel stop implant and prevents oxidation of the underlying silicon during field oxide growth). The remaining nitride and the thin oxide under it are removed resulting in the profile shown in figure 3.3.1b.

Next a layer of photoresist is applied and the wafer is exposed through the *depletion mode implant* mask. The resist is developed, leaving open spaces in the gate regions of the depletion load transistors. Another ion implantation step occurs here (using the resist as a mask) to alter the threshold voltages of the depletion load transistors. The resist is removed and a thin layer of *gate oxide* is grown. If there are *buried contacts* used in the IC design more photoresist is applied, the wafer is exposed through the buried contact mask, the resist is developed, the gate oxide is etched

Figure 3.3.1 Si Gate NMOS Processing Steps



(not to scale)

away in the contact areas, and the resist is removed. This allows the *polysilicon gate* material to contact the substrate in selected areas.

A layer of polysilicon is deposited from a chemical vapor and a thin layer of oxide is grown on top of that to provide a surface that photoresist will adhere to. Resist is applied and the wafer is exposed through the polysilicon layer mask. The development of the resist leaves the gates of the transistors covered; the uncovered areas of oxide and polysilicon are etched away (a little field oxide is also removed). After the resist is removed the *source* and *drain* regions are doped (figure 3.3.1c) in a phosphine gas atmosphere. Since the edges of the polysilicon gates define where the source/drain regions begin these features are also self-aligned. Here self-alignment results in a significant reduction in parasitic capacitance due to the near zero gate to source/drain overlap. A thick layer of oxide containing P_2O_5 is deposited over the surface of the wafer. This layer is reflowed for better coverage of the steps in the surface and a layer of photoresist is applied. *Contact hole* areas are defined using the contact cut mask and the oxide is etched away where the metal layer will contact the underlying features. After the resist is removed the source and drain are doped again (this is to prevent a phenomenon called *spike-through* — essentially shorting of the aluminum contacts and the substrate through the shallow source and drain regions). A layer of aluminum is evaporated onto the surface of the wafer, followed by the application of more photoresist. Exposure (and subsequent development) through the metal layer mask leaves resist protecting the metal runs and contacts. The uncovered aluminum is etched away and the resist is removed (figure 3.3.1d). The wafer is then *annealed* (heated at a low temperature) to remove radiation damage resulting from the electron beam that is used to heat the aluminum during the evaporation process.

A thick layer of oxide is deposited on the entire surface of the wafer to provide physical protection. Windows to the bonding pads are etched through this layer in another photolithography step using the *overglass layer* mask. At this point (figure 3.3.1e) the wafer is finished, ready to be broken apart, bonded and tested.

References

[Varian 1979]
Varian/Extrion Division, "Ee˚BES-40 Electron Beam Lithography System", Varian Corp., Gloucester, MA., August 1979.