CVD process. Following the application of a layer of photoresist, Mask #1 is used to pattern the surface. Mask #1, which is often called the *moat*, or n^+ *diffusion* mask, defines in photoresist the drain, source, and channel regions of all transistors as well as any other regions where n^+ implants are desired. After exposure, development removes the photoresist layer in areas that are not to be moat (i.e., the *complement of the moat*, or the *antimoat*). A top (Mask #1 pattern) and cross-sectional view at this stage of what will be the two transistors appear in Fig. 2A.1*a*. The Si_3N_4 is then etched from the areas not protected by the photoresist. A high-energy implant of p-type impurities (typically boron) is then applied to the entire wafer. The remaining photoresist protects the moat regions from this implant. This heavy implant is used to raise the threshold voltage in the antimoat region (often called the *field*) and to provide electrical isolation between adjacent devices. After this field implant and a drive in diffusion, the remaining photoresist is stripped. A thick layer of SiO_2 (about 10,000 Å) is then thermally grown by the oxidation process over the wafer. This layer is formed in the field, but no oxidation can take place in the region protected by the Si_3N_4 because Si_3N_4 does not oxidize. The thick field oxide layer is termed a local oxidation layer and is often called *LOCOS*. The oxidation consumes some of the substrate silicon. The second cross section in Fig. 2A.1*a* shows the state of the wafer following growth of the field oxide. This corresponds to Step 10 in the process scenario of Table 2A.1. Following removal of the Si_3N_4, the thin layer of SiO_2 under the Si_3N_4 is stripped and another SiO_2 layer is grown.

With the moat now protected only by the very thin SiO_2 layer, a light n-type implant over the entire wafer can be applied (optional) to set the threshold voltage of the enhancement devices. This implant is light enough so that all p-type regions remain p-type. If used, the implant is applied to the entire wafer to avoid the need for an additional mask.

A heavier selective implant is required in regions that are to serve as the channels of depletion transistors. To achieve this, a second layer of photoresist is applied to the entire wafer, and the second mask, Mask #2 (termed the *implant* mask), is used to pattern the photoresist so that only the channel regions of depletion transistors are unprotected. Another n-type implant is used to make the exposed regions n-type with the remaining photoresist serving as an implant mask. After stripping the photoresist, the wafer is as shown in Fig. 2A.1*b*. This corresponds to the status of the wafer at Step 19 in Table 2A.1.

Direct contacts between the lower polysilicon layer and moat are termed buried contacts. In the process scenario of Table 2A.1 this is listed as an optional step and is not used in the layout shown in Fig. 2A.1 although it could be used, if available, to reduce the area required for contacting the gate of M2 to its source. To make a buried contact, the thin gate oxide must be patterned and etched to remove the insulating SiO_2 layer and create paths (vias) through which the following polysilicon layer can contact the moat. Mask #A is used to pattern the buried contact vias. Although the buried layer contact can reduce area, the additional processing costs needed to provide this feature are often not justified; hence the buried contact feature is often not available in NMOS processes.

After stripping of any thin oxides that may be present at this stage in the moat region, a uniform thin layer of SiO_2, often termed *gate oxide* (200 to 1000 Å thick), is grown on the surface of the wafer. Stripping and regrowing provide better control of the critical gate oxide thickness. A layer of polysilicon (termed POLY I), which is about 2000 Å thick, is then deposited on the surface of the entire wafer. This is covered with photoresist, patterned with Mask #3, and etched to remove unwanted POLY I. The POLY I layer is used as gates for both enhancement and depletion transistors, as a plate for capacitors, as a conductor, and for resistors. The formation of a capacitor, a resistor, and enhancement and depletion transistors can be seen in Fig. 2A.1*c*. This corresponds to Step 27 in the process scenario of Table 2A.1. Note that the POLY I layer is over gate oxide in cross section AA′ and above field oxide in cross section BB′.

The remaining uncovered thin layers of SiO_2 are stripped and another thin layer (500–1000 Å) of SiO_2 is again grown over the entire surface. This serves as the dielectric for POLY I–POLY II capacitors, as an insulator for POLY I–POLY II crossovers, and as the gate oxide for transistors that use the POLY II layer as the gate. The thickness of this oxide layer is often ideally the same as that of the first gate oxide layer. By stripping the unexposed oxide and regrowing, a buildup in the depth of the oxide layers is prevented and more uniformity is attained. A second layer of polysilicon (termed POLY II) is then deposited, followed by photoresist and patterning with Mask #B. In the circuit shown in Fig. 2.2-5, the POLY II layer is used only for the upper plate of the capacitor, as shown in Fig. 2A.1*d* (Step B.9 of Table 2A.1). Note that it is slightly smaller than the underlying POLY I layer. This difference is standard practice when trying to accurately match capacitors to make one plate a little smaller than the other so that the smaller plate will effectively define the capacitor area independent of slight misalignments of the two plates. Additional practical considerations that further improve matching will be discussed later.

After stripping the thin SiO_2 layers, an n^+ diffusion is applied to the entire wafer. The field oxide and polysilicon layers serve as masks to the diffusion and prevent impurities from reaching the substrate in the protected areas. The n^+ diffusion creates the n-type drain and source regions of all transistors and makes any other unprotected moat areas n-type. The portion of the light depletion diffusion that is not protected by the polysilicon gates also becomes more heavily doped. Note that although no mask is used at this step, the n^+ diffusion mask, which was used as the first masking step, has essentially determined the n^+ diffusion regions fabricated at this step. The n^+ diffusion also penetrates into any exposed polysilicon layers, increasing the conductivity in these regions. The n^+ diffusion depth is about 5000 Å. It will be seen in the section discussing process parameters that the sheet resistance for POLY I and POLY II layers is about the same but that the sheet resistance of POLY I under POLY II is higher. This is due to the absence of the additional n^+ diffusion in the lower layer. Since the polysilicon gates serve as masks for the n^+ drain and source diffusions, the process is said to be *self-aligned*. In a self-aligned process, small misalignments of the gate (POLY I or POLY II) masks will not affect the gate geometry or dimensions, nor will they make the transistors nonfunctional.

Next, another insulating layer is deposited over the wafer surface. Doped deposited oxide, such as PSG, is often used for this purpose. This rather thick layer, ~ 6000 Å, serves as an insulator between the uppermost polysilicon layer and the subsequent metal layer. The field oxide depth is further increased with this deposited oxide layer. The entire wafer is again covered with photoresist, and Mask #4 (actually the fifth or sixth mask if POLY II and/or buried contact options are available) is used to pattern *contact openings* for the purpose of obtaining electrical contact from the top with the desired components. After the photoresist is developed, an etch that attacks the insulating layer but does not affect polysilicon or silicon makes the required openings. This etch is stopped in the vertical direction only by polysilicon or the single-crystal silicon of the substrate. The wafer takes the form shown in Fig. 2A.1*e* (Step 34 of Table 2A.1).

Metal (typically aluminum) is then deposited over the entire wafer, followed by another layer of photoresist. This metal layer is typically about 7000 Å thick. This photoresist is patterned by Mask #5, followed by an etch to remove unwanted metal. The metalization is used to interconnect components and provide external access to the integrated circuit. A metalization that interconnects the four basic components to form the circuit shown in Fig. 2.2-5 is shown in Fig. 2A.1*f* (Step 40 of Table 2A.1).

Large, square metal areas, called *bonding pads*, are needed to allow for contact with the IC package. Small bonding wires will later be connected from these pads to the pins on the IC package. These pads are also patterned with the metalization mask but are not shown in Fig. 2A.1*f* because of the large amount of area required for bonding pads relative to that needed for the components in Fig. 2.2-5.

A bonding pad is shown in Fig. 2A.1*g*. Four of these would be needed to interface the circuit of Fig. 2.2-5 with the IC package. The V_{BB} contact comes from the bottom side of the substrate. The bonding pad size has remained relatively constant for a long period of time even though considerable reductions in feature size of geometries on the die itself have been experienced. This is because the methods of physically mounting the die in packages and interconnecting the bonding pads to the pins in the package have not changed much. With bonding wires typically about 1 mil in diameter, it is difficult to reduce bonding pad size significantly.

The entire surface is finally covered with a passivation layer (often called glass or p-glass) to provide long-term stability of the IC by minimizing atmospheric contamination. A layer about 10,000 Å thick is often used for passivation. Since this layer is also an electrical insulator, it is necessary to again pattern it and make openings above the metal pads to allow for attaching the bonding wires. The final mask, Mask #6, is used for this purpose and is shown in Fig. 2A.1*g*.

For the simple circuit of Fig. 2.2-5, the area required for the bonding pads dominates that needed for the circuit itself. For simple circuits this is generally the case but as the complexity of the circuit increases, the percentage of the total area required for bonding pads becomes quite small.

Giving the information for each mask separately, as was done in Fig. 2A.1, makes it difficult to perceive the entire circuit and determine layer to layer

alignment. Sophisticated software packages termed layout editors are widely used, in which layers are color-coded and displayed simultaneously on high-resolution monitors. A single layout that simultaneously shows all mask information is shown in color in Plate 2. A color convention has been established for distinguishing separate layers. The color convention adopted in Plate 2 corresponds to that used in the MOSIS process and is discussed in more detail later in this chapter.

An interesting observation can be made from the layout of the MOS transistors of Fig. 2A.1. The MOS devices are totally geometrically symmetric with respect to drain and source and so must also be electrically symmetric. The designation of *drain* and *source* is thus arbitrary. In many applications a convention has evolved for convenience and consistency in device modeling in regard to drain and source designation. This convention will be discussed in Chapter 3. When appropriate, we will follow the established convention throughout this text.

In the process described, several alternative methods for constructing resistors and capacitors are available. For example, a region of moat with two contacts can be used as a resistor, and a capacitor can be made between POLY I and metal.

Processing step modifications such as omission of one polysilicon layer, omission of the depletion mask, substitution of metal gates for the polysilicon gates, and addition of another mask and implant to create enhancement transistors with two different threshold voltages are possible and common. Process procedure modifications such as using diffusions instead of implants; changing types of impurities; varying the thickness of oxide, polysilicon, or metal layers; including or excluding oxide stripping and regrowing steps; and changing types of photoresist are widespread and play major roles in yield, fabrication costs, and performance. The IC design engineer must be familiar with the process steps that will be used in fabrication when embarking on a new design.

For the process just described, it is the responsibility of the circuit designer to provide all information necessary to construct the seven masks shown in Fig. 2A.1. The size, shape, and spacing of the components are judiciously determined. The size and shape affect the performance of the circuit and are at the control of the circuit designer for optimizing performance, within constraints of minimum allowable size as determined by the capabilities of the process and maximum size as determined by economics. The spacing is also constrained by the capabilities of the process itself. The spacing and sizing specifications are obtainable from the design rules of the process, which are discussed later in this chapter. A process engineer will typically be responsible for providing design rules for a particular process.

2.2.1*b* CMOS Process

A discussion of a typical generic single-polysilicon silicon gate, p-well, n-substrate CMOS process follows. As in the NMOS case, variants in this process—such as a second metal layer, a second polysilicon layer, additional implants, oppositely doped substrate, or metal gates—are also well established. The devices available in the CMOS process under consideration are

1. n-channel MOSFETs.
2. p-channel MOSFETs.
3. Capacitors.
4. Resistors.
5. Diodes.
6. npn bipolar transistors.
7. pnp bipolar transistors.

The diodes and bipolar transistors are often considered parasitic components and are generally not extensively used as components in the circuit design itself. The process is tailored to maintain optimal characteristics in the n- and p-channel MOSFETs at the expense of poor characteristics for the bipolar transistors.

A method of physically constructing each of the first four components in the list will be considered. The approach followed here is similar to that followed for the NMOS process except that all mask details are included on the single layout of Fig. 2B.1 of Appendix 2B. A color version of this figure appears in Plate 3. Fewer details about oxide growth, photoresist application and patterning, and so on are provided since these steps are very similar to the corresponding steps for the NMOS process. Cross-sectional views along AA′ and BB′ in Fig. 2B.1*a* after each major step are shown in Fig. 2B.1. Interconnections follow the approach used in Section 2.2-1*a* for the NMOS process and are not discussed here. A summary of the major process steps appears in Table 2B.1. Additional information about this process relating to layout sizing rules, physical feature sizes, and electrical characterization parameters of the generic CMOS process can be found in Tables 2B.2–2B.5 of Appendix 2B. The process described here is very similar to the 3 μ CMOS/bulk process available through MOSIS.[7]

The starting point of this CMOS process is a polished n-type silicon disc. A layer of SiO_2 is first grown on the entire disc, followed by the application of a layer of photoresist. This photoresist is patterned with Mask #1 to provide openings for a p-tub (alternatively, p–well), which will serve as the substrate for the n-channel MOS devices. Either a deposition or implant is used to introduce the p-type impurities that form the tub. This diffusion is quite deep (about 30,000 Å). The remaining photoresist and SiO_2 are then stripped, a thin layer of SiO_2 regrown, and the entire surface covered with a layer of Si_3N_4.

Mask #2, termed the *moat mask* or the *active mask* by MOSIS, is used to pattern the Si_3N_4 layer. The Si_3N_4 layer is then etched away except above the regions that are to be the n^+ and p^+ diffusions or channel regions for the n-channel and p-channel MOSFETs. These diffusions will be added by subsequent processing steps to form drain and source regions for MOSFETs as well as to form guard rings. These protected regions are again termed moat. After the Si_3N_4 layer is opened, the remaining photoresist is stripped. Figure 2B.1*b* depicts the wafer after Step 15 of the process scenario of Table 2B.1.

An optional field threshold adjust step may be introduced at this point. This field threshold adjust would be used to raise the threshold voltage in the n-type

substrate in regions that will not contain devices. This will provide increased isolation between the p-channel transistors. Although an additional mask is required for this field adjust (Mask #A1 of Table 2B.1), the mask would be the complement of the union of the p-well mask and the active mask, Masks #1 and #2. As such, this mask information would be generated automatically and need not be separately provided by the designer.

A thick layer of field oxide (typically 10,000 Å) is grown in the regions not protected by the remaining Si_3N_4 that was patterned with Mask #2. The Si_3N_4, along with the remaining SiO_2 that was under this layer, are then stripped. The wafer at this stage is as depicted in Fig. 2B.1*c*. This corresponds to Step 18 in the process scenario of Table 2B.1. Note that along cross section AA′ several isolated areas are not protected by the field oxide. These areas will be used for fabricating transistors and guard rings. No breaks in the field oxide appear in the BB′ cross section. This corresponds to the region where the resistor and capacitor will appear; these devices are fabricated on top of the field oxide.

Next, a thin, uniform layer of SiO_2 (200 to 1000 Å), called in this case gate oxide, is regrown. A layer of polysilicon (typically 2000 Å) is then deposited, covered with photoresist, and patterned with Mask #3. This polysilicon layer, termed POLY or POLY I, is used for the gates of all transistors, as a plate on capacitors, for resistors, and for interconnects. Following etching and stripping, the wafer takes the form shown in Fig. 2B.1*d*. This corresponds to Step 25 in the process scenario of Table 2B.1.

An optional second polysilicon layer, termed POLY II, could be included here, as provided in the process scenario. The second polysilicon layer is not depicted in Fig. 2B.1. This second polysilicon layer would be separated from the first by a thin (500 to 1000 Å) insulating layer of SiO_2. The main purpose of the second polysilicon layer would be for the formation of capacitors with POLY I and POLY II as electrodes, although this layer would also find some use in interconnects and crossovers if available. The capacitance density and electrical characteristics of the poly–poly capacitors are more attractive than those obtainable with other capacitors available in this process. An additional mask, termed the POLY II or *electrode mask*, is needed to pattern this polysilicon layer. The etch of both the POLY I and POLY II layers produces an abrupt, sharp edge, making reliable coverage of this edge with thin material difficult. Since the oxide between POLY I and POLY II is thin, crossing of a POLY I boundary with POLY II may result in either a break in the POLY II or a shorting of POLY I and POLY II. To circumvent these problems, the crossing of a POLY I boundary with POLY II is often not permitted.

At this stage the drain and source diffusions for both the n-channel and p-channel transistors are added. Although two different types of diffusions and hence two separate masks are required, the designer need specify only one of the two masks. In this process, only those areas not protected by field oxide are capable of accepting any diffusion impurities. This is termed the moat, or active, region. It is further provided in this process that any moat area that is not exposed to n-type impurities will be exposed to p-type impurities. Consequently, the designer selects those moat regions that are to become p-type with Mask #4,

which is termed the p^+ select mask. The n^+ select mask (Mask #5), which is used to pattern those regions of moat that are to become n-type, is automatically generated from the complement of the p^+ select mask intersected with the moat (active) mask. p^+ select is used in the substrate to form p-channel transistors and interconnects and is used in the p-well to provide ohmic contact to the p-well as well as for guard rings.

Correspondingly, n^+ select is used in the p-well to form n-channel transistors and interconnects, and in the substrate to make top ohmic contacts as well as additional guard rings. Further comments about guard rings and their role in latch-up protection appear in Section 2.4. As was the case in the NMOS process, the polysilicon layer or layers are patterned prior to the p^+ and n^+ diffusions. The polysilicon that lies in the moat serves as a diffusion mask for these diffusions and provides self-alignment of the gate with the drain and source regions.

The n^+ and p^+ diffusions are much shallower than the p-well diffusion and are typically in the 5000 Å and 7000 Å ranges, respectively. Following the p^+ and n^+ diffusions, which occur prior to Step 36 in the process scenario, the cross-sectional profile is as shown in Fig. 2B.1*e*. The n-channel and p-channel transistors, along with the ohmic contacts and guard rings, are clearly visible at this stage. A thick insulating layer, which is a deposited oxide (often PSG), is then placed over the entire wafer. This insulating layer is about 6000 Å thick and serves as an insulator between the uppermost polysilicon layer and the subsequent metal layer. This causes a further thickening of the field oxide and is depicted above the dashed interface of the field oxide layer shown in Fig. 2B.1*f*.

Mask #6 is used to pattern contact openings. Areas unprotected by photoresist after patterning are etched away. This etch will consume insulating layers but is stopped by either polysilicon or the silicon substrate. This provides for metal contact of either a polysilicon layer or a p^+ or n^+ diffusion depending on which is the uppermost layer present.

After the contacts are opened, metal is applied uniformly to the wafer and patterned with Mask #7. This is termed the *metal mask* (or, if subsequent metal layers are to be added, the *metal 1 mask*). This corresponds to Step 48 in the process scenario and the cross section of Fig. 2B.1*f*.

An optional second metal can be added at this stage. This requires two additional mask steps, one for making contact with underlying metal 1 and the other for patterning the second metal layer. The mask used to pattern the contact openings, or vias, between the two metal layers is termed the *via mask*. Polyimide is often used as the insulating layer between the two metals because it offers advantages in step coverage over other commonly used insulating layers.

Following application of a passivation layer, often termed p-glass, the passivation is opened above the bonding pads to provide for electrical contact from the top with Mask #8. This is often termed the *glass mask*. The pad layout is similar to that discussed for the NMOS process and depicted in Fig. 2A.1*g*.

This completes the CMOS processing steps for the generic CMOS process scenario of Table 2B.1. The resistor, capacitor, n-channel MOSFET, and p-channel MOSFET should be apparent from Figs. 2B.1*a* and 2B.1*f*.

2.2.1*c* Practical Process Considerations

The equipment needed for the CMOS process is basically the same as is needed for the NMOS process previously described. With this equipment the minimum feature size for the CMOS process is comparable to that for the NMOS process. It should be noted that eight masks and considerably more processing steps than are required for the basic six-mask NMOS process are needed for this CMOS process. In addition, it will be seen later that considerably more area is required for the same number of devices in a CMOS process than in an NMOS process with the same feature size. The increase in size is due largely to the required size of the large p-tubs and the n- and p-type guard rings. These increases in area are, however, often offset by less complicated designs and/or the superior performance that is attainable with the CMOS process.

Several physical and processing-dependent material characteristics cause the physical MOSFET to differ from the ideal. The processing-dependent material characteristics will be considered first.

WIDTH AND LENGTH REDUCTION. A typical cross section of the n-channel MOSFET along EE′ and FF′ of Fig. 2B.1*a* is compared with the ideal in Fig. 2.2-6. These cross sections are intentionally not to scale so that they will better illustrate the actual characteristics.

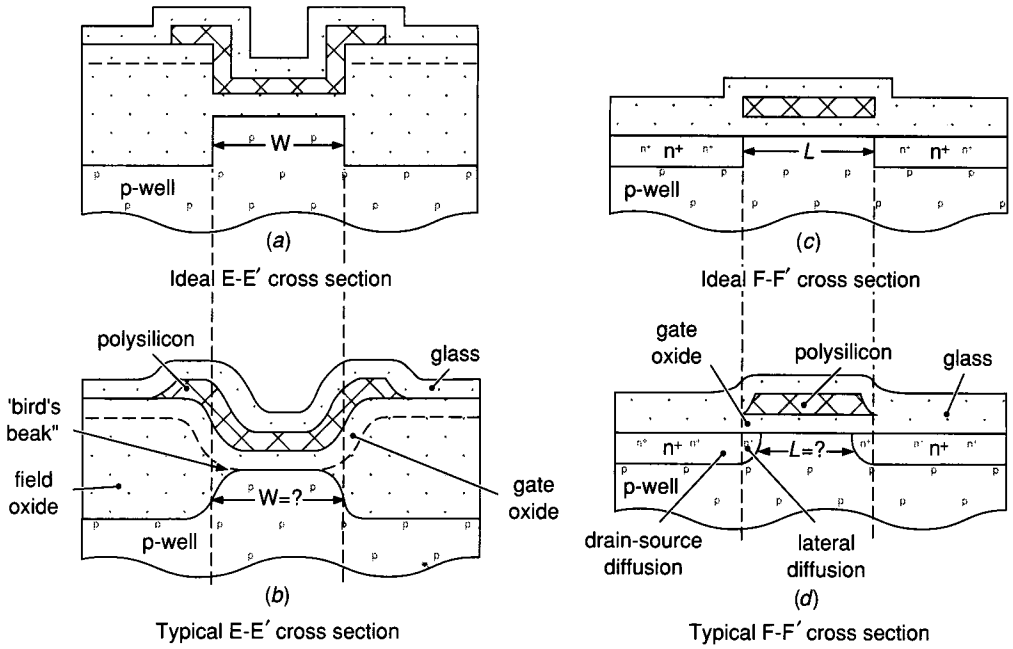

FIGURE 2.2-6
Width and length reduction in MOSFETS.

It will be seen later that the width and length of the MOSFET are key parameters at the control of the designer that play a major role in device performance. The width, W, is the width of the moat, or active, region as depicted in Fig. 2.2-6*a*, which corresponds to the EE′ cross section of the MOSFET, and the length L is the distance between the drain and source diffusions, as indicated in the FF′ cross section of Fig. 2.2-6*c*. It should be emphasized that the device dimensions are determined by the size of the *intersection* of the poly mask and the active mask and not by the dimensions of the poly pattern that forms the gate.

In the typical cross section of Fig. 2.2-6*b*, it can be seen that during the field oxide growth, encroachment into the active region effectively reduced the width of the transistor. This oxide encroachment is termed *bird's beaking* due to the distinctive shape of the encroachment. This is particularly troublesome because the width of the transistor is no longer precisely defined and because the exact amount of width reduction is not easily controllable. A second factor that affects the effective width is the accuracy with which the protective Si_3N_4 layer used to pattern the field oxide can be controlled. The effective width of this layer is affected by both the patterning of the photoresist and the problems associated with etching that were discussed in Section 2.1. In addition, since a thin SiO_2 layer (200–800 Å) is applied prior to the Si_3N_4 to minimize mechanical stress at the Si_3N_4 interface, the encroachment of the SiO_2 growth also limits accuracy.

Similar problems in controlling the length of the transistor exist, as indicated in Figs. 2.2-6*c* and *d*. The major source of length reduction is associated with the lateral diffusion of the drain and source diffusions, which are difficult to precisely control. Assuming the lateral and vertical diffusion rates are equal and that the diffusion depth is 5000 Å, the total length reduction due to lateral diffusion, since it diffuses in from both ends, would be around 1 μ. This is very significant and problematic in short channel transistors. Other factors that affect the effective length are the accuracy in patterning the photoresist that defines the polysilicon gate length and the accuracy in controlling the polysilicon gate etch itself.

In summary, both length and width reduction are inherent with existing processing technologies. Although they can be partially compensated for by considering these reductions during design or automatically adjusting (termed *size-adjust*) the geometrical database to over- or undersize the appropriate mask geometries, these effects are difficult to precisely control, and the exact width and length of the device are difficult to define. These effects are particularly troublesome for small geometries with device dimensions in the 1 μ or smaller range. Partial compensation with the mask size-adjust is often provided, thus allowing the designer to assume that the nominal value of the actual dimensions on silicon agree with those specified on the design.

LATERAL WELL DIFFUSION. Lateral diffusion associated with the creation of the p-well also deserves mention. The depth of the p-well is about 3 μ and, the lateral diffusion associated with the well formation is comparable. Although not a major factor limiting device performance, this lateral diffusion consumes considerable surface area and forces the designer to leave a large distance between isolated p-wells and between any p-well and p^+ diffusion in the substrate, thus

increasing chip cost by increasing die area. One way to partially minimize the impact of the large amount of area associated with well boundaries is to group large numbers of n-channel transistors into a single p-well when the wells for these transistors are to be tied to the same potential. Tradeoffs between these area savings and the corresponding increase in interconnect area must be made.

LATCH-UP. The physics of layered doped silicon is also problematic. It is well known that a four-layer sandwich of doped material, npnp or pnpn, forms a Silicon Controlled Rectifier (SCR). Once an SCR is "fired" (switched to on conducting state), it continues to conduct until the gate signal is removed and current flow is interrupted. Several parasitic bipolar transistors and an SCR are identified on the cross section of Fig. 2.2-7 which is based on Fig. 2B.1*f*. Several diodes can also be identified. Although this CMOS process has not been optimized for obtaining good performance of these bipolar devices, there are limited practical applications of some of the diodes and bipolar transistors. The SCR, however, is very undesirable and if it is caused to fire, excessive current will usually flow, causing destructive failure of the integrated circuit. The firing of the SCR is termed *latch-up* in CMOS circuits. The CMOS designer must make certain that latch-up cannot occur in any design.

Latch-up problems are strongly layout dependent. A theoretical treatment of the latch-up problem is beyond the scope of this text, but a thorough understanding of this problem is not needed for successfully designing CMOS circuits provided the designer is familiar with layout techniques that circumvent the problem. Guard rings are widely used to prevent latch-up. Exact requirements for guard ring

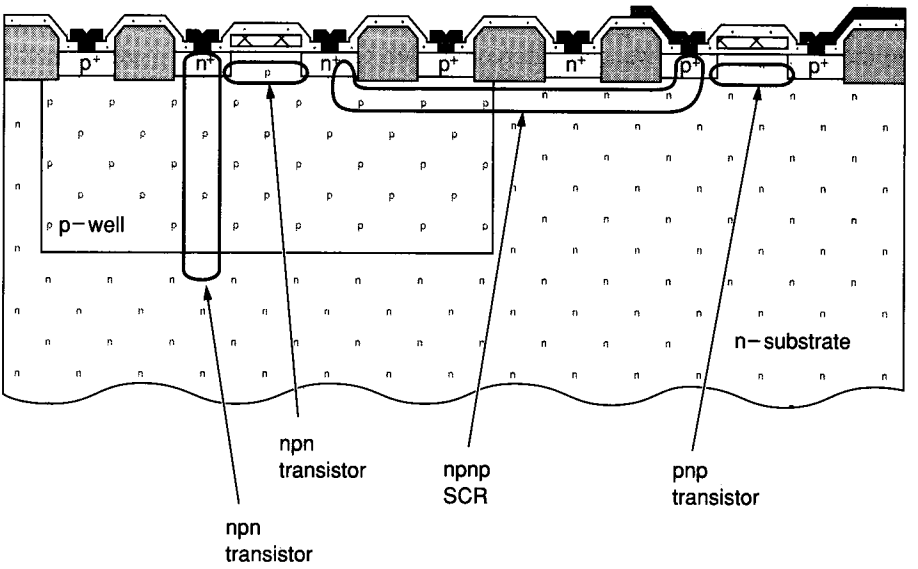

FIGURE 2.2-7
Parasitic transistors in a p-well CMOS process.

placement will be determined once a particular CMOS process is defined. In some processes, separate and additional n^+ and/or p^+ diffusions are included specifically for guard ring formation. This requires additional masks and additional processing steps. In the CMOS process discussed in this section, no additional masking or processing steps are required since the normal drain and source diffusions are also used to fabricate guard rings.

One way to obtain latch-up protection in the generic CMOS process of this section is to completely encircle every p-well with a p^+ guard ring. Such a guard ring is shown in Fig. 2B.1*a* around the periphery of the p-well. Metal contact is made as often as possible to this guard ring to further reduce resistance. The guard ring will then typically be connected via metallization to the lowest dc potential in the circuit—V_{SS} or ground, for example. Although not shown in Fig. 2B.1*a*, encircling the p-well with an n^+ guard ring provides additional protection and is also desirable. A partial n^+ guard ring separating the p-well from the p-channel substrate transistor can be seen in the same figure. As before, numerous metal contacts are made to this guard and it is subsequently tied to the highest potential in the circuit.

Breaks in a guard ring must be avoided since these breaks could provide a path for breakdown. Consider first the p^+ guard ring in the p-well. Breaks can occur one of two ways. The most obvious is to exclude a segment from either the moat mask or the p^+ select mask. The other way is to cross the guard ring *anywhere* with polysilicon in the process described in Table 2B.1. Such a crossing will cause a break because the polysilicon is patterned prior to the p^+ diffusion and serves as a mask to this diffusion. The break would thus occur under any polysilicon crossing of the intended guard ring. The exclusion of polysilicon crossing of the guard ring is undesirable from a circuit designer's viewpoint because it complicates interconnection between devices in the p-well and those outside the p-well; all interconnection crossings must be made of metal to avoid breaking the guard ring. Polysilicon crossing of guard rings in other process scenarios where a separate p^+ guard ring diffusion is available may be permissible. Correspondingly, breaks in the n^+ guard ring will occur if a segment is omitted from the active mask, if it is crossed with p^+ select, or if it is crossed with polysilicon.

Although complete enclosure of the p-well with the n^+ guard ring is desirable, some designers using the generic CMOS process described in this section use only the p^+ guard ring or have the p^+ guard ring and include the n^+ guard material only between the p-channel transistors and the p-well, as depicted in Fig. 2B.1*a*.

INPUT PROTECTION. Static breakdown is also of concern, and protection of inputs must be provided to prevent destructive breakdown when handling the devices. The major sources of concern are inputs that have a direct connection to a region separated from the rest of the circuit only by thin oxide, such as gate oxide or poly–poly oxide, with no direct connection to any diffused region. Such inputs would include the gates of any transistors or any connection to a floating polysilicon capacitor electrode. The breakdown is due to a destructive breakdown

of this thin oxide due to electric fields that exceed the oxide breakdown voltage. As stated in Chapter 1, silicon dioxide will break down when electric fields are in the 5 MV/cm to 10 MV/cm range. With 1000 Å gate oxides, this would occur for voltage inputs in the 50 V to 100 V range. Although these are beyond the maximum allowable input voltages specified for a typical 3 μ CMOS process, these voltages are much less than the static voltages experienced when handling these chips. The problem is even worse for thinner gate oxides. Such breakdown is destructive and must be prevented. Input protection circuitry is used for this purpose. This input protection must not interfere with the normal operation of the circuit. A single simple protection circuit is typically developed and is used repeatedly by connecting it to each pad that requires protection.

One common protection scheme involves connecting the input pads through a small polysilicon resistor to a reverse-biased diode that nondestructively breaks down at voltages below the critical gate oxide breakdown voltage. The node that is to be protected then becomes an internal node coincident with the node corresponding to the interconnection between the protection resistor and the diode. The resistor is used to safely limit peak current flow in the protection diode. In the CMOS process described in this section, this diode would be constructed by putting an n^+ diffusion in a p-well with a p^+ select guard ring around the periphery of the well. This guard ring would be connected to the lowest potential in the circuit, typically ground or V_{SS}, and the n^+ diffusion would be connected to the pad that is to be protected through the polysilicon resistor. This protection circuit provides protection through the reverse breakdown voltage of the diode if the input is positive and through normal forward-biased diode conduction if the input is negative. For the NMOS process, single diode protection can be attained by connecting the critical node through a polysilicon resistor to an n^+ diffusion. No guard ring is available or required in an NMOS circuit.

An alternative that provides all protection through normal forward-biased diode conduction is obtained if a second diode of opposite polarity shunts the diode just described. This diode is constructed from a p^+ diffusion in the substrate, with the n-substrate connected to the highest potential in the circuit and the p^+ diffusion connected to the intersection node of the first diode and the polysilicon resistor. It is recommended that this p^+ diffusion be encircled by an n^+ guard ring, which also would be connected to the substrate. Under normal operation the diodes in the input circuitry do not conduct, so the input protection is ideally transparent to the user. Actually, the diodes do contribute to a small amount of leakage current. They also contribute to a small parasitic capacitance connected to an ac ground, which may be of limited concern in some applications.

Although the input protection schemes discussed could be used on any input or output pad, such circuitry is generally not required on pads that are already directly connected to a diffusion region, even if they are also connected to layers separated by thin oxide from other nodes in the circuit, because the diffused region itself forms part of the diode and thus provides inherent self-protection. Nevertheless, care should always be exercised when handling any MOS devices, even if good circuit-level protection has been included, to reduce the chance of destroying the integrated circuit by static breakdown.

2.2.2 Bipolar Process

The basic active devices in the bipolar process are the npn and pnp transistors. These names are descriptive since the devices are constructed with three layers of n- or p-type semiconductor material, with the middle layer different from the other two. These layers can be fabricated either laterally or vertically. A simplified pictorial description of these transistors, including the established symbols for the devices, appears in Fig. 2.2-8. Several excellent references discuss the basic operation of the BJT.[3,12–17] The modeling of the BJT is discussed in Chapter 3. As will be seen later, the characteristics of the collector and emitter regions as well as their geometries are intentionally different and as such the designation of the collector and emitter contacts is not arbitrary. The convention that has been established for designating the collector and emitter contacts will be discussed in Chapter 3.

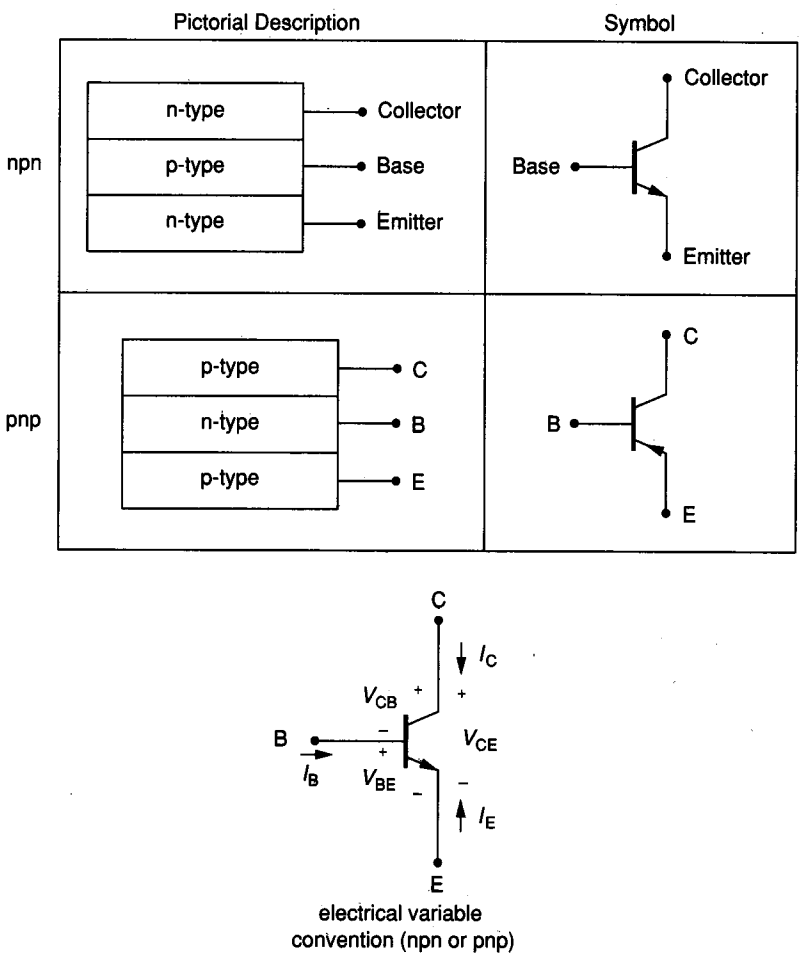

FIGURE 2.2-8
Bipolar transistors.

The components available in the bipolar process are

1. npn bipolar transistors.
2. pnp bipolar transistors.
3. Resistors.
4. Capacitors.
5. Diodes.
6. Zener diodes.
7. Junction Field Effect Transistors (JFETs)—not available in all bipolar processes.

A familiarity with the process is crucial to utilizing this wide variety of components in the design of an integrated circuit. Unlike discrete component circuit design with these same devices, whose characteristics can be specified over a wide range and which can be connected in any manner, the basic characteristics of the devices available for bipolar integrated circuits are determined by the process and the range of practical values and parameters is severely limited. In addition, the methods of interconnection are strictly limited, the basic devices have characteristics that are quite temperature dependent, and the passive component values are typically somewhat dependent on the signal applied. In spite of these restrictions (to be discussed later), very clever analog and digital bipolar integrated circuits have evolved. The bipolar process is used for the popular TTL, ECL, and I^2L digital logic families as well as a host of linear integrated circuits, including the 741 operational amplifier, the 723 voltage regulator, and the 565 phase locked loop. Although minor variances in the processing steps are common, the major differences are in device sizes and impurity concentrations and profiles. The discussion of a typical seven-mask bipolar process follows. The major process steps are outlined in Table 2C.1 of Appendix 2C.

The construction of npn transistors, pnp transistors, resistors, and capacitors will be considered. The location of these components can be seen in the top view containing mask information shown in Fig. 2C.1*a* of Appendix 2C.

The starting point in this bipolar process is a clean, polished p-type silicon wafer. A layer of SiO_2 is first grown over the wafer. Following application of a layer of photoresist, Mask #1 is used to pattern the n^+ buried layer. The n^+ buried layer serves the purpose of decreasing collector resistance and minimizing the parasitic current flow from collector to substrate in npn transistors. It also helps decrease the base resistance in lateral pnp transistors. Either a deposition or implant, followed by a drive in diffusion, can be used to introduce the n-type impurities into the substrate through the openings provided by Mask #1. A layer of oxide, which grows during the diffusion, is then stripped. After the n^+ diffusion the wafer takes the form shown in Fig. 2C.1*b*.

An n-type epitaxial (crystalline) layer is then grown over the entire wafer. The thickness of this layer typically varies between 2 μ and 15 μ, with the thinner layers used for digital circuits and the thicker layers for analog circuits. This layer will be used for the collector region in npn transistors. The epitaxial

layer is shown in Fig. 2C.1*c*. Note that some of the impurities in the buried layer have migrated (or *out-diffused*) into the epitaxial layer during its growth. A thick layer of SiO_2 (typically 5000 Å) is then grown over the entire surface.

Mask #2 is used to pattern the SiO_2 layer for the p^+ isolation diffusion. SiO_2 is etched from the areas not photographically protected by Mask #2 to allow for this drive in diffusion following a p^+ deposition. The p^+ isolation diffusion is used to electrically separate adjacent transistors. It is wide and deep since it must completely penetrate the epitaxial layer to provide the required isolation. The wafer at this stage is as shown in Fig. 2C.1*d*. Although the isolation diffusion is shown with vertical edges in the figure, lateral diffusion, typically comparable to the vertical diffusion, causes significant out-diffusion laterally under the oxide layer, thus making the top of the channel stop considerably wider than the bottom. Following this diffusion, another thick layer of SiO_2 is grown over the entire wafer.

An optional shallow, high-resistance p-diffusion could be added at this step. This is not depicted in Fig. 2C.1 but is listed as an option in Table 2C.1. This step would provide a mechanism for making practical diffused resistors in the 1 kΩ to 20 kΩ range. A typical sheet resistance of this region would be 1 kΩ/□ to 2 kΩ/□ . Mask #A in Table 2C.1 is used to pattern these regions.

Mask #3 is used to pattern the SiO_2 layer and define the base regions for the npn transistors as well as the collector and emitter regions for lateral pnp devices. A p-type deposition and a subsequent drive in diffusion create these regions in the unprotected areas. This diffusion is much shallower than was the isolation diffusion and must not penetrate the epitaxial layer. The isolation mask openings provided by Mask #2 are typically reopened with Mask #3 to provide a few additional p-type impurities. The wafer at this stage is shown in Fig. 2C.1*e*.

Following growth of another layer of SiO_2, Mask #4 is used to pattern the emitter regions for the npn transistors. An n^+ deposition followed by a drive in diffusion creates the emitter regions. Openings are also made in the oxide above the collector to add small n^+ wells in the lightly doped collector region to provide for better electrical contact from the surface. The integrated circuit at this stage is as depicted in Fig. 2C.1*f*. The emitter diffusions must be shallow so as not to penetrate the relatively shallow p-type base regions already created. The amount and profile of the impurities in the n^+ emitter regions and the thickness of the p-type base region, which is now sandwiched between the n^+ emitter and the n-collector, strongly influence the gain of the transistor.

Mask #5 is used to pattern contact openings to allow for top contact of the circuit with the metallization. The entire circuit is then covered with a thin layer of metal. Mask #6 is used to pattern the metal, followed by the addition of a passivation layer. The completed cross-sectional view of the four components under consideration is shown in Fig. 2C.1*g*. Mask #7 patterns pad openings to allow for electrical contact to the bonding pads.

Two modifications of this process deserve mention. One involves adding a *deep collector* diffusion. This requires an additional masking step and is used to diffuse impurities under the area where the collector contacts are to be made. This step would occur either before or after the isolation diffusion and is used to

extend n^+ impurities all the way from the surface to the buried layer. Since this is such a deep diffusion, an area penalty in the collector is experienced. The deep diffusion is used to reduce collector resistance in high-current applications. The second modification involves adding an additional p-diffusion in the p-channel stops. This also requires an additional mask step and is used to avoid surface inversion in high-voltage parts. With the exception of open collector circuits, this step is not common in basic logic parts.

The npn and pnp transistors in this process are depicted in Fig. 2.2-9. The npn transistor is called a vertical npn device since the emitter, base, and collector regions are stacked vertically. It can now be seen that the n^+ buried layer decreases the collector resistance that must be modeled in series with the collector.

The pnp transistor is called a lateral device since it is stacked laterally (horizontally). The base width cannot practically be made as narrow and the base area is not as accurately controllable as for the npn device. In addition, the emitter and collector regions must have the same impurity profile. The characteristics of the lateral transistors are generally considered poorer than those of the vertical devices. Other pnp transistors, not shown in Fig. 2C.1, can be constructed by using the p-type base diffusion as the emitter, the n-type epitaxial layer as the base, and the p-type substrate as the collector. These devices are called substrate transistors. The performance of these devices is also mediocre, and applications are restricted since all collectors of substrate transistors are common.

The capacitor that was constructed in Fig. 2C.1 may at the outset appear to be merely a diode. It would serve the purpose of a diode, even though the area is considerably more than may be required in most applications. When reverse biased, however, the depletion layer forms the dielectric, and the p and n regions on either side form the capacitor plates. Capacitors made like this, with total

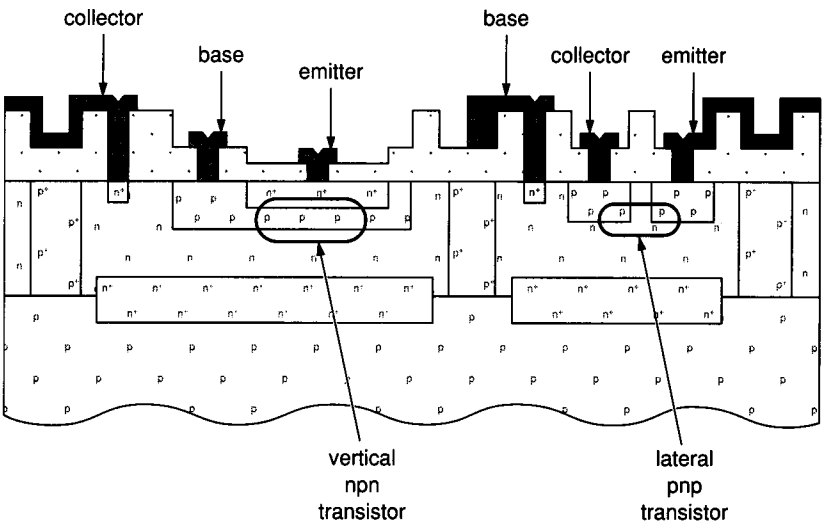

FIGURE 2.2-9
Vertical and lateral transistors in a bipolar process.

capacitances from the sub-picofarad to the 100 picofarad range, have proven practical. The capacitors are not without limitations, however. The requirement that the junction must always be reverse biased severely limits the interconnection flexibility of this device. The width of the depletion layer is voltage dependent, making the capacitance nonlinear. Temperature also affects the capacitance value. Finally, the base–emitter junction typically breaks down with a reverse bias of about 7 V, limiting the maximum voltage that can be applied to the capacitor. The base–collector junction can also be used as a capacitor. Its characteristics are very similar to the base-emitter capacitor with the exception that it offers an increased reverse breakdown at the expense of a lower capacitance density. Junction capacitors will be discussed in more detail in Chapter 3. An alternative to the junction capacitor would be a metal–oxide–semiconductor capacitor formed between the metal and the n^+ emitter diffusion. Although the characteristics of this capacitor would be better than those of the junction capacitors, an extra mask step is generally required to provide a means of selectively stripping the thick oxide above the emitter region so that a thin oxide can be regrown. The thin oxide is needed to get the capacitance density up to a practical level.

The resistor of Fig. 2C.1 is actually just a serpentined strip of the lightly doped p-type base diffusion. The underlying n-type epitaxial layer is generally contacted and taken to the highest potential in the circuit to prevent current flow into this region. Several other techniques for fabricating resistors in this process are available. They will be discussed in Chapter 3.

Although minimum feature sizes are comparable for the bipolar and MOS processes, standard bipolar processes require more area per device than do the NMOS processes. A major reason for this increased area is the deep and wide p^+ channel stops that are required for device isolation in standard bipolar processes. An alternative bipolar process using trench isolation[23] is available which offers a significant improvement in component density over the standard bipolar process.

2.2.3 Hybrid Technology

The hybrid approach to integrated circuit design involves attaching two or more integrated circuit dies (typically of different types), along with some discrete components in some cases, in a single package to form what is called a hybrid integrated circuit. It is often, and desirably, transparent to the consumer whether the circuit is monolithic or hybrid; in some cases, however, the hybrid packages are considerably larger. The hybrid integrated circuit is typically more costly than the monolithic structures. The extra cost and size of hybrid integrated circuits is offset, in some demanding applications, by improved performance capabilities.

Hybrid circuits containing discrete components occupy considerably less area than the conventional PC board/discrete component approach. They have played a major role in demanding analog signal processing applications such as high-resolution A/D and D/A converters and precision active filters. Tolerances, temperature dependence, and area-induced component value limits for resistors and capacitors in standard MOS and bipolar processes have limited the development of monolithic integrated circuits for precision continuous-time signal processing. Thick film and thin film passive components have reasonable toler-

ances, are easily trimmable, have acceptable temperature coefficients that can be tailored for tracking, and offer reasonable tradeoffs between area required and component values. These thick film and thin film networks are commonly used for the passive components in hybrid integrated circuits. A discussion of thick film and thin film processing technologies follows.

THICK FILM CIRCUITS. The thick film technology is relatively old, requires considerable area compared to monolithic circuits, can be used for relatively high-power applications, and can be applied at relatively high frequencies (up to 1 GHz) although it is typically limited to a few MHz. The increased area required by the thick film circuits is offset by the reduced cost in equipment and processing materials required for the thick film process, the latter being a small fraction of that required for either bipolar or MOS processes.

The components available in a thick film process are resistors and capacitors along with conducting interconnects. Layers of different material are successively screened onto an insulating substrate. These materials are used for resistors and conductors as well as for the dielectrics of capacitors.

The number of resistive layers varies but practical limitations generally restrict this to at most three. Typical thickness of these layers (called pastes or inks) is about 20 μ, but they may vary considerably by design. The actual thickness of these layers is not accurately controllable ($\pm$ 30%) due to limitations in the screening process itself.

The thick film process offers the most advantages for resistor fabrication. Although capacitors are often included, the electrical characteristics of thick film capacitors are not outstanding and the capacitance density is quite low. Discrete chip capacitors, which have much better characteristics than their thick film counterparts, are often bonded to thick film resistive networks in hybrid circuit applications.

The minimum conductor width in a typical thick film process is about 250 μ, and minimum resistor widths are about 1250 μ. It can be seen that these are orders of magnitude larger than the corresponding minimum feature sizes for the MOS and bipolar processes (1–5 μ).

Screening involves forcing the paste through small holes in a tightly stretched piece of fabric called a screen, typically constructed of stainless steel. The grid is quite regular. The spacing of the holes can be specified, but practical physical limitations relating to both the mechanical characteristics of the steel and the physical characteristics of the inks prevent the use of extremely fine meshes. Screens with a grid spacing ranging from 100 to 300 filaments/inch are typical. This spacing restricts thick film resolution to somewhere around 500 μ. Where paste is not desired, holes in the screen are plugged by a mask. A squeegee is used to force the ink through the unrestricted areas. Following screening, each layer is fired to harden it.

Inks are available with sheet resistances that satisfy the equation

$$1\ \Omega/\square < R_{\square} < 10\ \text{M}\Omega/\square \tag{2.2-9}$$

for fired layers 20 μ thick. This large latitude in ink characteristics allows for a wide range of resistor values. Since only one type of ink can be used for each

resistor layer, the tradeoffs between area and sheet resistance must be considered when specifying the ink sheet resistances. Even though adjusting the length is a convenient means of establishing resistor values, long thick film resistors are to be avoided because they develop "hot spots" and are difficult to trim. The "hot spots" are caused at regions where the resistive layer is a little thinner and/or narrower than surrounding regions. This causes an increased resistance in this small region, which under constant current causes increased local power dissipation. This power dissipation causes heating, which typically further increases the resistance and power dissipation. Heating causes deterioration of the film layer at these points. Deterioration in these regions can eventually result in device failure. Short, wide resistors are also to be avoided. The main problem with short, wide resistors is the inability to accurately specify the size of the resistor since the contacts will overlap with a considerable portion of the device. A reasonable rule of thumb for the allowable width (W) / length (L) ratio for a rectangular resistor is

$$\frac{1}{10} < \frac{W}{L} < 3 \tag{2.2-10}$$

Although the W/L ratio is constrained, the values for W and L remain to be specified within the design rules of the process. It is a good practice to make the resistors large if the area is available to minimize edge roughness effects, increase power dissipation capability, and make trimming easier.

Serpentined patterns such as shown in Problem 2.11 should also be avoided with thick film technology. This is due to the increased current density that will result from current crowding at the inner corners of serpentined structures.

A capacitor is constructed by screening a conductive layer, followed by a dielectric, followed by another conductor. The dielectric is generally applied in two coats to minimize pinholes, which would short the capacitor plates together. With the two layers of dielectric, the chances of a pinhole coincident with both layers are greatly reduced. Since a thick film capacitor is actually a parallel plate capacitor, the capacitance is given by

$$C = \epsilon_R \epsilon_0 \frac{A}{t} \tag{2.2-11}$$

where $\epsilon_0 = 8.854$ pF/m, A is the area of the capacitor plates, t is the dielectric thickness, and ϵ_R is the relative dielectric constant. Inks with relative dielectric constants from 10 to 1000 are available. The high dielectric constants offer a reasonable capacitance density at the expense of large and nonlinear temperature coefficients. The lower dielectric materials offer improved performance but are restricted to applications requiring small capacitors due to a low capacitance density. As in the MOS and bipolar processes, the upper plate of thick film capacitors is typically a little smaller than the lower to minimize capacitance changes caused by minor misalignments. The two conductive layers used for the capacitor plates also serve as interconnects. If crossovers of two conductors are required, the dielectric layer can be used as an insulator at the expense of creating a small parasitic capacitor at the crossover.

The screen geometries, along with a cross-sectional view of a typical thick film process, are depicted in Fig. 2D.1 of Appendix 2D. This process has two

resistive screenings as well as two conductive layers and a dielectric for capacitor fabrication. The major process steps are listed in Table 2D.1. Process parameters and characteristics, along with design rules for a typical thick film process, are also given in Appendix 2D.

Thick film resistors can be trimmed with a laser or by abrasion. These trims, which can be very accurate, remove part of the thick film layer and thus increase the resistance. For this reason, resistors that are to be trimmed are typically targeted to be undersized in value by about 40% to guarantee trimmability in spite of process variations.

Thick film capacitors can also be trimmed by abrasively removing part of the upper plate (along with some dielectric). Since this decreases the capacitance, the thick film capacitors are typically targeted to be oversized by about 40%.

THIN FILM CIRCUITS. The components available in thin film processes are resistors and capacitors, although often only resistors are included due to both the specific applications which naturally benefit from thin film technology and the practical limitations of thin film capacitors. Thin film circuits are much smaller than thick film circuits. They are similar to thick film circuits in that successive layers are applied to an insulating substrate as contrasted to the MOS and bipolar processes, where some of the processing steps involve diffusions that actually penetrate the substrate. For conductors, thin film thicknesses are typically from 100 to 500 Å although thicknesses of several thousand angstroms are occasionally used if a high conductivity is needed. Film thicknesses from 100 to 2000 Å for resistors and film thicknesses in the 3000 Å region for dielectrics of capacitors are common. Note that these film thicknesses are comparable to the thicknesses of layers applied in the MOS and bipolar processes but are orders of magnitude thinner than the 20 μ (200,000 Å) typical of thick films. The sheet resistance range for thin film resistors is typically from 50 Ω/□ to 250 Ω/□, which is considerably less than is available in thick film processes. The thin film layers are applied by uniformly coating the entire wafer with the film. Then unwanted areas are selectively patterned and etched with a photolithographic process similar to that used in the MOS and bipolar cases.

The minimum feature sizes for the thin film components are comparable to those of the MOS and bipolar processes. The temperature characteristics and performance of thin film components are quite good with the exception of the dielectrics for capacitors, which are quite lossy. "Hot spots," which were a problem with thick film circuits for long resistors, are not a major problem with thin film resistors since the thin films are typically more uniform and since thin film applications generally require smaller current flow.

Thin film circuits are much more expensive to produce than their thick film counterparts because of the sophisticated equipment that is needed for both the photolithographic process and the film depositions and etching. They are used extensively in telecommunication circuits at low frequencies but also find applications at higher frequencies (up to 30 GHz) as well.

Thin film resistors can be accurately trimmed by a laser. Thin film capacitors are not well adapted to a continuous trim although binarily weighted capacitors connected with laser-fusible links are trimmable in quantized decrements.