9 / MOS Process Integration

In Chapter 9 we explore a number of relationships between process and device design and circuit layout. Processes are usually developed to provide devices with the highest possible performance in a specific circuit application, and one must understand the circuit environment and its relation to device parameters and device layout.

In this chapter we look at a number of basic concerns in MOS process design, including channel-length control; layout ground rules and ground-rule design; source/drain breakdown and punch-through voltages; and threshold-voltage adjustment. Metal-gate technology is discussed first, and then the important advantages of self-aligned silicon-gate technologies are presented. Discussions of CMOS technology and the application of anisotropic etching to MOS devices complete the chapter.

9.1 BASIC MOS DEVICE CONSIDERATIONS

To explore the relationship between MOS process design and basic device behavior, we begin by discussing the static current-voltage relationship for the MOS transistor, as developed in Volume IV of this series.[1] The cross section of two metal-gate NMOS transistors is shown in Fig. 9.1. In the linear region of operation, the drain current is given by

$$I_D = \overline{\mu}_n C_0 (Z/L)(V_{GS} - V_T - V_{DS}/2)V_{DS} \tag{9.1}$$

for $V_{GS} \geq V_T$ and $V_{DS} \leq V_{GS} - V_T$. $C_0 = K_s \varepsilon_0 / X_0$ is the oxide capacitance per unit area, $\overline{\mu}_n$ is the average majority-carrier mobility in the inversion layer, and V_T is the threshold voltage.

One of the first specifications required is the circuit-power-supply voltages, which set the maximum value of V_{GS} and V_{DS} that the devices must withstand. Once this choice is made, the only variables in eq. (9.1) which a circuit designer may adjust are the width and length of the transistor. Thus, the circuit designer varies the circuit topology and horizontal geometry to achieve the desired circuit function.

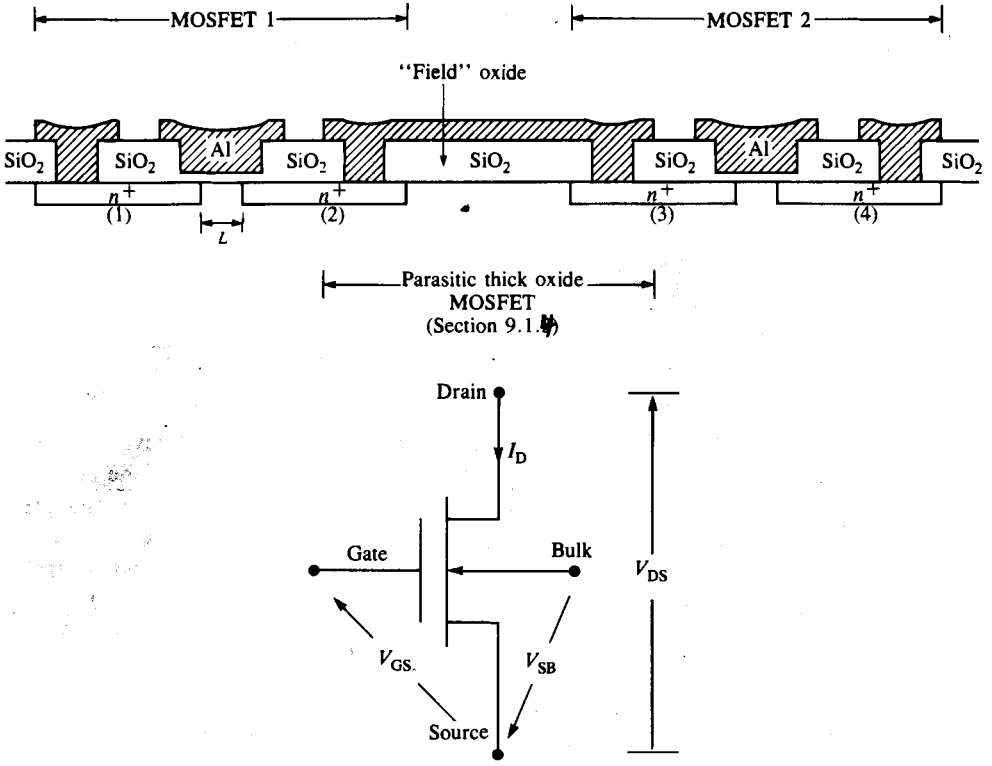


Fig. 9.1 (a) Cross section of an integrated circuit showing two adjacent NMOS transistors. A parasitic NMOS device is formed by the aluminum interconnection over the field oxide with diffused regions (2) and (3) acting as source and drain. (b) An NMOS transistor with gate-to-source (V_{GS}), drain-to-source (V_{DS}), and source-to-bulk (V_{SB}) voltages defined.

Other device parameters are fixed by the process designer, who must determine the process sequence, times, temperatures, etc., which ultimately determine the device structure and hence its characteristics. These include specifying the gate-oxide thickness, field-oxide thickness, substrate doping, and field and threshold-adjustment implantations. The process designer also supplies a set of "design rules" or "ground rules" which must be obeyed during circuit layout. These include minimum channel length and width, spacings between features on the same and different mask levels, and overlaps between features on different mask levels. A mask alignment sequence and tolerances must also be developed for the process.

9.1.1 Gate-Oxide Thickness

Current flow in the MOS transistor, for a given set of terminal voltages, is inversely proportional to the gate-oxide thickness. The gate oxide will generally be made as thin

as possible, commensurate with oxide breakdown and reliability considerations. High-quality silicon dioxide will typically break down at electric fields of 5 to 10 MV/cm, corresponding to 50 to 100 V across a 100-nm oxide. Present processes are using oxide thicknesses between 20 and 100 nm. Below 10 nm, current starts to flow by tunneling, and the oxide begins to lose its insulating qualities. The choice of oxide thickness is also related to hot electron injection into the oxide, a problem beyond the scope of this text.[2–4]

9.1.2 Substrate Doping and Threshold Voltage

Threshold voltage is an important parameter which determines the gate voltage necessary to initiate conduction in the MOS device. The threshold voltage[1] for a device with a uniformly doped substrate is given by:

$$\text{NMOS: } V_T = \Phi_M - \chi - \frac{E_g}{2q} + |\Phi_F| + [\sqrt{2K_s\varepsilon_0 qN_B(2|\Phi_F| + V_{SB})}]/C_0 - Q_{tot}/C_0 \tag{9.2}$$

$$\text{PMOS: } V_T = \Phi_M - \chi - \frac{E_g}{2q} - |\Phi_F| - [\sqrt{2K_s\varepsilon_0 qN_B(2|\Phi_F| - V_{BS})}]/C_0 - Q_{tot}/C_0$$

$$|\Phi_F| = (kT/q)\ln(N_B/n_i)$$

in which N_B is the substrate doping. $\Phi_M - \chi = -0.11$ for an aluminum gate, $\Phi_M - \chi = 0$ for an n^+-doped polysilicon gate, and $\Phi_M - \chi = +1.12$ for a p^+-doped polysilicon gate.

Q_{tot}[†] represents the total oxide and interface charge per cm^2 and adds a parallel shift of the curves in Fig. 9.2 to more negative values of V_T. This charge contribution to the threshold voltage had an extremely important influence on early MOS device fabrication. Q_{tot} tends to be positive, which makes the MOS transistor threshold more negative; n-channel transistors become depletion-mode devices ($V_T < 0$), whereas p-channel transistors remain enhancement-mode devices ($V_T < 0$). During early days of MOS technology, Q_{tot} was high, and the only successful MOS processing was done using PMOS technology. After the industry gained an understanding of the origin of oxide and interface charges, and following the advent of ion implantation, NMOS technology became dominant because of the mobility advantage of electrons over holes. Today, total charge levels have been reduced to less than 5×10^{10} charges/cm^2 in good MOS processes, and the oxide charge contribution to threshold voltage is minimal.

Substrate doping enters the threshold-voltage expression through both the $|\Phi_F|$ term and the square-root term. A plot of threshold voltage versus substrate doping for n- and p-channel, n^+ polysilicon-gate devices with 50-nm gate oxides is given in Fig. 9.2 for $Q_{tot} = 0$. The choice of substrate doping is complicated by other considerations including

[†] $Q_{tot} = Q_F + Q_{IT} + \gamma_M Q_M$

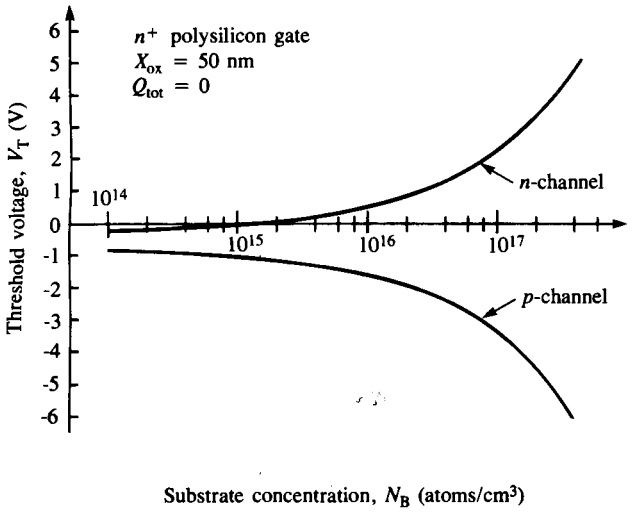


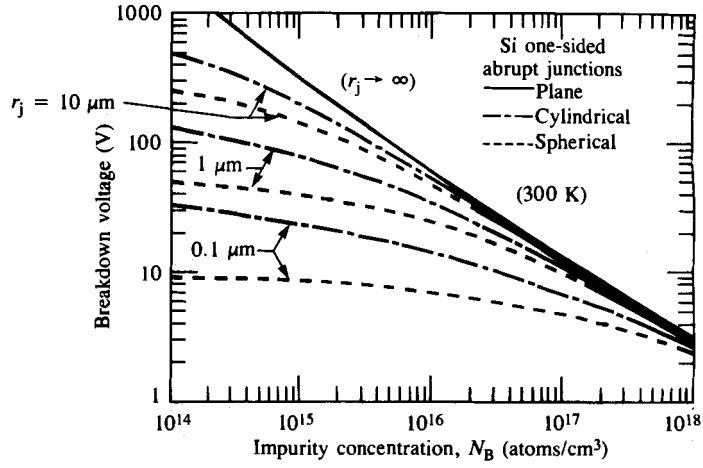
Fig. 9.2 Threshold voltages for *n*- and *p*-channel polysilicon-gate transistors with 50-nm gate oxides, calculated from eq. (9.2).

drain-to-substrate breakdown voltage, drain-to-source punch-through voltage, source-to-substrate and drain-to-substrate capacitances, and substrate sensitivity or body effect.

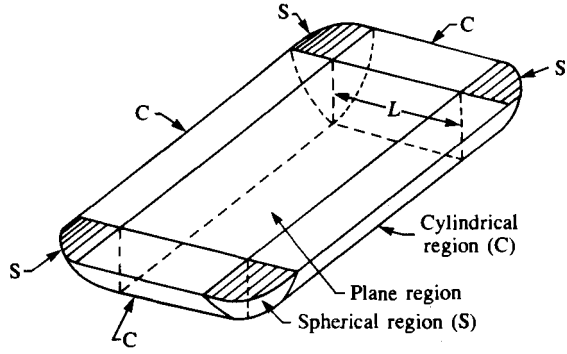
The source and drain regions are usually heavily doped to minimize their resistance and are essentially one-sided junctions in which the depletion region extends entirely into the substrate. Figure 9.3a gives the breakdown voltage of a one-sided *pn* junction as a function of the doping concentration on the lightly doped side of the junction.[5] Junction breakdown voltage decreases as doping level increases. Breakdown voltage is also a function of the radius of curvature of the junction space-charge region. Junction curvature enhances the electric field in the curved region of the depletion layer and reduces the breakdown voltage below that predicted by one-dimensional junction theory. A rectangular diffused area has regions with both cylindrical and spherical curvature, as shown in Fig. 9.3b.

Punch-through occurs when the drain depletion region contacts the source depletion region, and substrate doping must be chosen to prevent the merging of these depletion regions when the MOSFET is off. Punch-through will not occur if the channel length exceeds the sum of the depletion-layer widths of the source-to-substrate and drain-to-substrate junctions. For a transistor used as a load device in a logic circuit, the source-to-substrate and drain-to-substrate junctions must both support a voltage equal to the drain supply voltage plus the substrate supply voltage. The depletion-layer widths can be estimated using the formula for the width of a one-sided step junction:

$$W = \sqrt{(2K_s\varepsilon_0(|V_A| + \Phi_{bi}))/qN_B}$$
$$\Phi_{bi} = 0.56 + (kT/q)\ln(N_B/n_i) \tag{9.3}$$



(a)



(b)

Fig. 9.3 (a) Abrupt *pn* junction breakdown voltage versus impurity concentration on the lightly doped side of the side of the junction for both cylindrical and spherical structures. r_j is the radius of curvature. (b) Formation of cylindrical and spherical regions by diffusion through a rectangular window. Copyright, 1985, John Wiley & Sons, Inc. Reprinted with permission from ref. [5].

where V_A is the total applied voltage and Φ_{bi} is the built-in potential of the junction. If the channel length is greater than $2W$, punch-through should not occur. Figure 9.4 gives the depletion-layer width of *pn* junctions as a function of doping and applied voltage. Punch-through is not a limiting factor for most doping levels, except for very short-channel transistors.

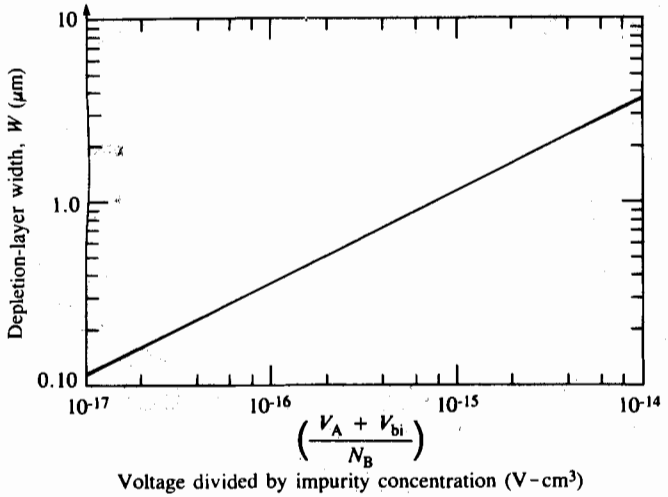


Fig. 9.4 Depletion-layer width of a one-sided step junction as a function of doping and applied voltage calculated from eq. (9.3).

The capacitance per unit area associated with a diffused junction is given by the parallel-plate capacitance formula with a plate spacing of W:

$$C_j = K_s\varepsilon_0/W$$

The larger the doping, the larger the capacitance. Zero bias and a doping concentration of $10^{16}/cm^3$ result in a junction capacitance of approximately 10 nf/cm^2.

Eq. (9.2) shows that the threshold voltage depends on the source-to-substrate voltage, V_{SB}. This variation is known as "substrate sensitivity" or "body effect," and it becomes worse as the substrate doping level increases.

From the above discussion, one can see that there are tradeoffs involved in the choice of substrate doping. Substrate doping is directly related to threshold voltage. It is desirable to reduce substrate doping to minimize junction capacitance and substrate sensitivity and to maximize breakdown voltage. Mobility also tends to be higher for lower doping levels. On the other hand, a heavily doped substrate will increase the punch-through voltage.

9.1.3 Threshold Adjustment

Ion implantation is routinely used to separate threshold-voltage design from the other factors involved in the choice of substrate doping. Substrate doping can be chosen based on a combination of breakdown, punch-through, capacitance, and substrate sensitivity considerations, and the threshold voltage is then adjusted to the desired value by adding

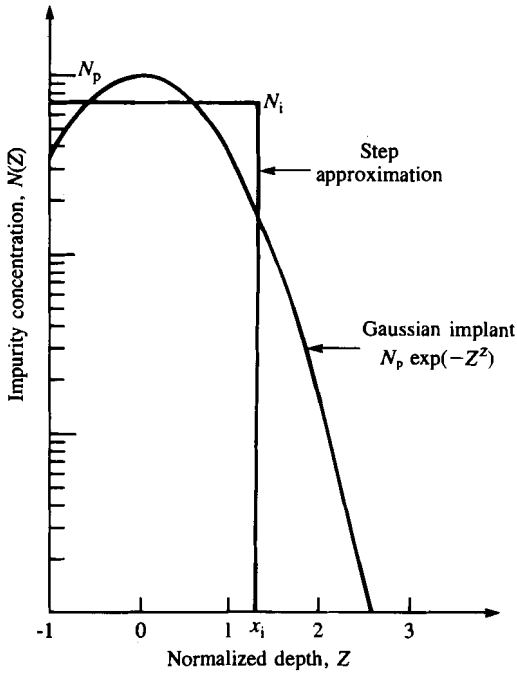


Fig. 9.5 Step approximation to a Gaussian impurity profile used to estimate the threshold-voltage shift achieved using ion implantation.

a shallow ion-implantation step to the process. Figure 9.5 shows a step approximation to an implanted profile used to adjust the impurity concentration near the surface. These additional impurities cause a shift in threshold voltage given approximately by

$$\Delta V_T = (1/C_0)(qQ_i)(1 - x_i/2x_d), \qquad x_i \ll x_d, \qquad x_d = \sqrt{qN_B/4K_s\epsilon_0|\Phi_F|} \tag{9.4}$$

where $Q_i = x_i N_i$ represents the implanted dose and x_d represents the depletion-layer width beneath the gate. For shallow implants, the threshold-voltage shift is approximately proportional to the implanted dose. The threshold-voltage shift is positive for acceptor impurities and negative for donor impurities.

Example 9.1: An NMOS transistor with an n^+ polysilicon gate is fabricated with a 25-nm gate oxide, a substrate doping of $3 \times 10^{15}/\text{cm}^3$, and source/drain junction depths of 3 μm. Determine the threshold voltage and drain-to-substrate breakdown voltages for this device. What is the punch-through voltage for a channel length of 4 μm if the substrate bias is -3 V? A shallow boron implantation is to be used to adjust the threshold to 1.0 V. What is the dose of this implant? Assume $V_{SB} = 0$ and $Q_{tot} = 0$.

Solution: For the n^+ polysilicon-gate transistor, $\Phi_M - \chi - E_g/2q = -0.56$ V and $|\Phi_F| = 0.33$ volts (for $n_i = 1 \times 10^{10}/cm^3$ and $kT/q = 0.026$ V). For $V_{SB} = 0$, the threshold voltage expression yields $V_T = -0.56 + 0.33 + 0.20$ V $= 0.03$ V. Interpolating Fig. 9.3 for spherical breakdown with a substrate doping of $3 \times 10^{15}/cm^3$ and a radius of curvature of 3 μm gives an estimated drain-to-substrate breakdown voltage of 60 V. To estimate the punch-through voltage, we use eq. (9.3) with $2W = 4$ μm and $V_A = V_D + 3$, where V_D is the drain voltage. Evaluating this expression yields $V_D = 89$ V.

For a shallow implant, the threshold-voltage shift is approximately $\Delta V_T = q\Delta Q/C_0$. A voltage shift of 0.97 V with an oxide thickness of 25 nm yields $\Delta Q = 8.4 \times 10^{11}/cm^2$.

NMOS depletion-mode ($V_T < 0$) transistors are routinely used in processes designed for high-performance logic applications. In order to reduce the NMOS threshold voltage, n-type impurities are implanted to form a built-in channel connecting the source and drain regions of the transistor, as in Fig. 9.6. The device characteristics of a depletion-mode transistor are similar, although not identical, to those of an enhancement-mode NMOS transistor, and the dose needed to shift the threshold voltage may be estimated using eq. (9.4).

9.1.4 Field-Region Considerations

The region between the two transistors in Fig. 9.1 is called the *field* region and must be designed to provide isolation between adjacent MOS devices. Several factors must be considered. The metal line over the field region can act as the gate of a "parasitic NMOS transistor" with diffused regions (2) and (3) acting as its source and drain. In order to ensure that this parasitic device is never turned on, the magnitude of the threshold voltage in this region must be much higher than that in the normal gate region. Referring to eq. (9.2), the threshold voltage may be made higher by increasing the oxide thickness

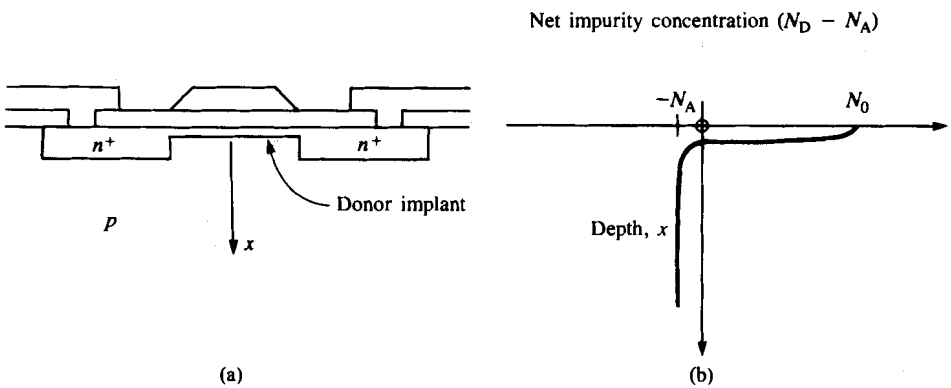


Fig. 9.6 (a) Formation of a depletion-mode NMOS transistor using a shallow ion-implanted layer; (b) net impurity profile under the gate of the depletion-mode MOSFET.

in the field region and by increasing the doping below the field oxide. The "field oxide" is typically made three to ten times thicker than the gate oxide of the transistors.

Another problem occurs for NMOS transistors. The substrate for NMOS transistors is p-type, usually doped with boron. We know that thermal oxidation results in depletion of boron from the surface of the silicon, and looking at eq. (9.2) we see that boron depletion will lower the threshold voltage of the transistors in the field region. A field implant step is often added to modern processes to increase the threshold voltage and compensate for the boron depletion during field-oxide growth.

For PMOS devices, the substrate is typically doped with phosphorus. During oxidation, phosphorus pileup at the surface tends to increase the threshold voltage in the field region. Thus phosphorus pileup helps to keep the parasitic field devices turned off.

One must also ensure that parasitic conduction does not occur between two adjacent devices due to punch-through. The source and drain diffusions of each transistor must be spaced far enough from the source and drain diffusions of the other transistors to ensure that the depletion regions do not merge together. The spacing between adjacent transistors must be greater than twice the maximum depletion-layer width.

9.2 MOS TRANSISTOR LAYOUT AND DESIGN RULES

Design of the layout for transistors and circuits is constrained by a set of rules called the "design rules" or "ground rules." These rules are technology-specific and specify minimum sizes, spacings, and overlaps for the various shapes that define transistors. Processes are designed around a "minimum feature size," which is the width of the smallest line or space that can be reliably transferred to the surface of the wafer using a given generation of lithography.

To produce a basic set of ground rules, we must also know the maximum misalignment which can occur between two mask levels. Figure 9.7a shows the nominal

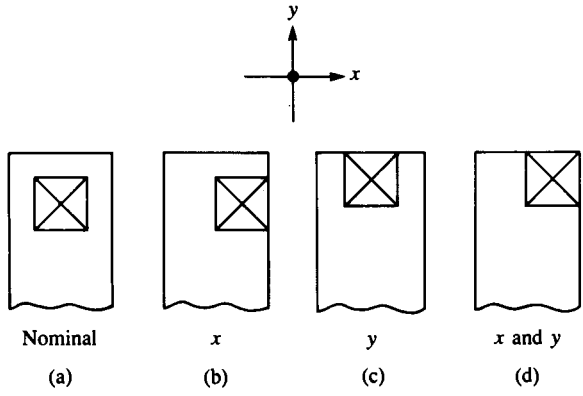


Fig. 9.7 (a) Nominal alignment of the contact and metal masks; (b) worst-case misalignment in the x-direction, (c) in the y-direction, and (d) in both directions.

position of a metal line aligned over a contact window. The metal overlaps the contact window by at least one "alignment tolerance" in all directions. During the fabrication process, the alignment will not be perfect, and the actual structure may have misalignment in both the x and y directions. Figures 9.7b through d show the result of worst-case misalignment of the patterns in the x, y, and both directions simultaneously. Our set of design rules will assume that this "alignment tolerance" is the same in both directions.

9.2.1 Metal-Gate Transistor Layout

Figure 9.8 shows the process sequence for a basic metal-gate process. The first mask defines the position of the source and drain diffusions. Following diffusion, the second mask is used to define a window for growth of the thin gate oxide. The third and fourth masks delineate the contact openings and metal pattern. The metal-gate mask sequence, omitting the final passivation layer mask, is as follows:

1. Source/drain diffusion mask	First mask
2. Thin oxide mask	Align to level 1
3. Contact window mask	Align to level 1
4. Metal mask	Align to level 2

An alignment sequence must be specified in order to properly account for alignment tolerances in the ground rules. In this metal-gate example, mask levels two and three are aligned to the first level, and level four is aligned to level two.

We will first look at a set of design rules for metal-gate transistors similar in concept to the rules developed by Mead and Conway.[6] These ground rules were designed to permit easy movement of a design from one generation of technology to another by simply changing the size of a single parameter, λ. In order to achieve this goal, the rules are quite loose in terms of level-to-level alignment tolerance. We will explore tighter ground rules later in this chapter.

A set of metal-gate rules is shown in Fig. 9.9. The minimum feature size $F = 2\lambda$, and the alignment tolerance $T = \lambda$. The parameter λ could be 5 μm, 2 μm, or 1 μm, for example. Transistors designed using our ground rules will fail to operate properly if the misalignment exceeds the specified alignment tolerance T.

On the metal level, minimum line widths and spaces are equal to 2λ. In some processes, the metal widths are made larger because the metal level encounters the most mountainous topology of any level.

On the diffusion level, the minimum linewidth is 2λ. The minimum space between diffusions is increased to 3λ to ensure that the depletion layers of adjacent lines do not merge together. However, the spacing between the source/drain diffusions of a transistor may be 2λ.

In this set of rules, the alignment tolerance between two mask levels is assumed to be 1λ, which represents the maximum shift of one level away from its nominal position, relative to the level to which it is being aligned. A 1λ shift can occur in both the x and y directions.

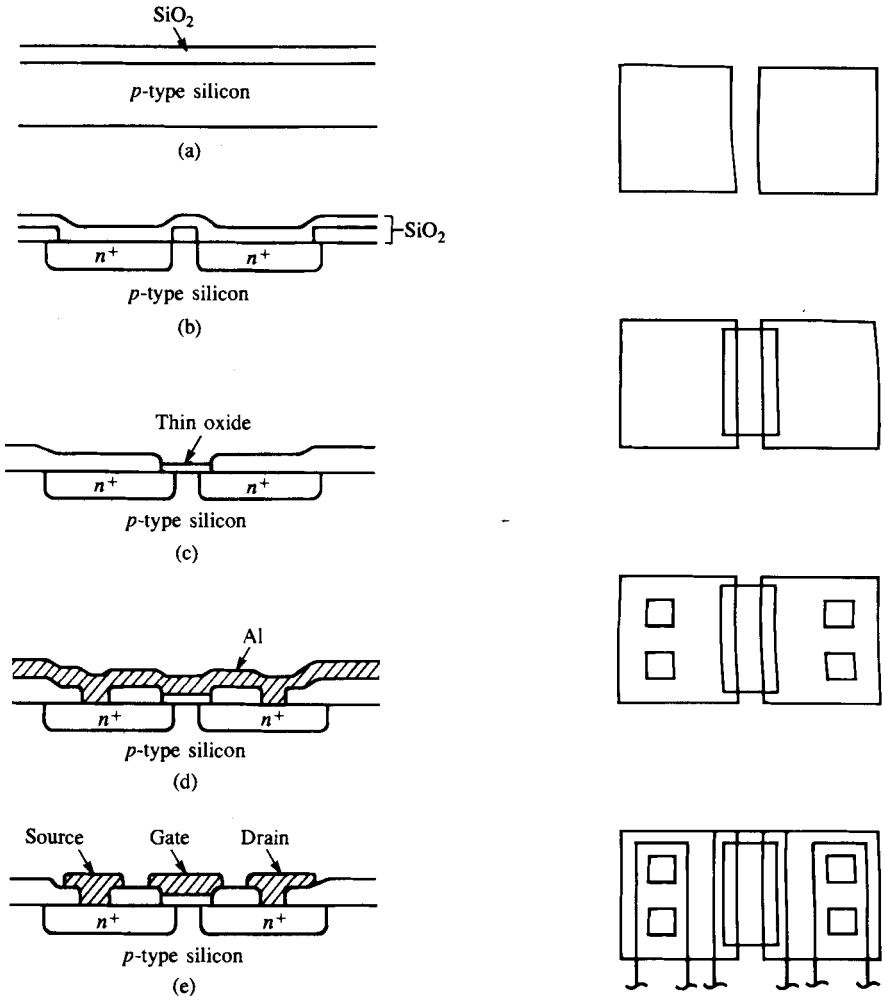


Fig. 9.8 Mask steps and device cross sections in a metal-gate process. (a) Substrate ready for first mask step; (b) substrate following source/drain diffusion and oxide regrowth, (c) following gate-oxide growth, (d) following contact window mask and aluminum deposition, and (e) following metal delineation.

Square contacts are a minimum feature size of 2λ in each dimension. It is normal practice to ensure that the contact is completely covered by metal even for worst-case alignment. Depending on the alignment sequence, a 1λ or 2λ metal border will be required around the contact window. Likewise, a contact window must be completely surrounded by a 1λ or 2λ border of the diffused region beneath the contact.

For our metal-gate transistors, the thin oxide region will be aligned to diffusion, so it requires a 1λ overlap over the source/drain diffusions in the length direction. The

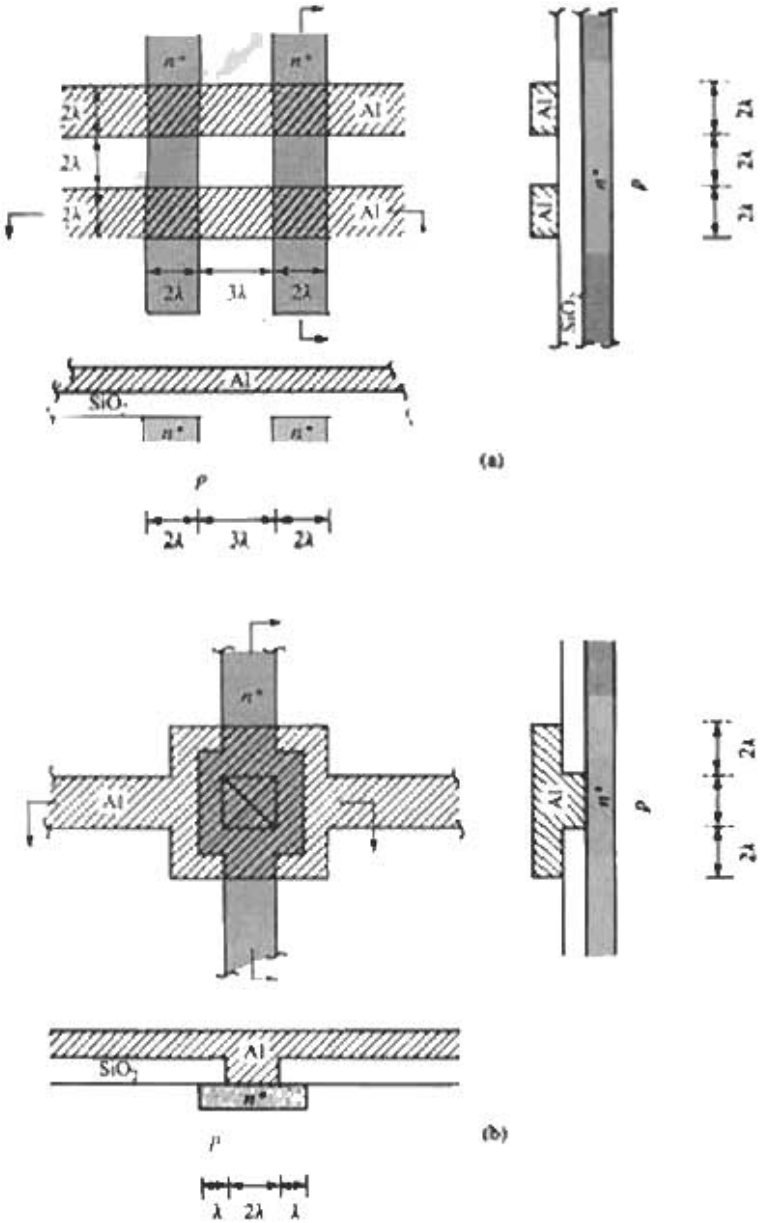


Fig. 9.9 A simple "λ-based" set of "design rules" or "ground rules" based on an alignment sequence in which levels 2 and 3 are aligned to level 1 and level 4 is aligned to level 2. (a) Rules for metal and diffused interconnection lines; (b) rules for contacts between metal and diffusion.

source/drain regions must also extend past the thin oxide by at least 1λ in the width direction. Contacts must be inside the diffusions by 1λ. The metal level is aligned to the thin oxide level, whereas the contacts are aligned to the diffusion level. A worst-case layout therefore requires a 2λ border of metal around contact windows but only a 1λ border around the thin oxide regions.

Figure 9.10 shows the horizontal layout and vertical cross section of a minimum-size NMOS metal-gate transistor with $Z/L = 10\lambda/2\lambda = 5/1$ at the mask level. The two diffusions are spaced by a minimum feature size of 2λ. Thin oxide must overlap the diffusions by 1λ in the length direction and underlap the diffusions by 1λ in the width direction. Metal must overlap thin oxide by 1λ. Accumulated alignment tolerances cause the minimum width of the gate metal to be 6λ. The spacing between metal lines must be 2λ. The metal over the contact holes must be 8λ wide because of the alignment sequence used, and the contact hole must be 1λ inside the edge of the diffusion. The resulting minimum transistor is 26λ in the length direction and 16λ in the width direction.

A new design rule has been introduced into this layout. The gate metal is spaced 1λ from the diffusion to prevent the edge of a metal line from falling directly on top of the edge of the diffusion in the nominal layout.

Several observations can be made by looking at this structure. First, note that the transistor is $416\lambda^2$ in total area, whereas the active channel area of the device is $20\lambda^2$! The rest of the area is required in order to make contacts to the various regions, within the constraints of the minimum feature size and alignment tolerance rules. Second, there is a substantial area of thin and thick oxide in which the gate metal overlaps the source and drain regions of the transistor. This increases the gate-to-source and gate-to-drain capacitance of the transistor. In this metal-gate transistor layout, the channel is defined by the junction edges in the length direction and by the thin oxide region in the width direction.

It should also be noted that there are several small contact windows in the source and drain regions. The usual practice is to make all the contact windows the same size throughout the wafer. From a processing point of view, equal-size contact windows will all tend to open at the same time during the etching process.

9.2.2 Polysilicon-Gate Transistor Layout

Transistors fabricated using polysilicon-gate technology have a number of important advantages over those built using metal-gate processes. We will discover some of these advantages by looking at the layout and structure of the polysilicon-gate transistor.

The mask sequence for the basic polysilicon-gate process from Chapter 1 is (again without passivation layer) as follows:

1. Active region (thin oxide) mask — First mask
2. Polysilicon mask — Align to level 1
3. Contact window mask — Align to level 2
4. Metal mask — Align to level 3

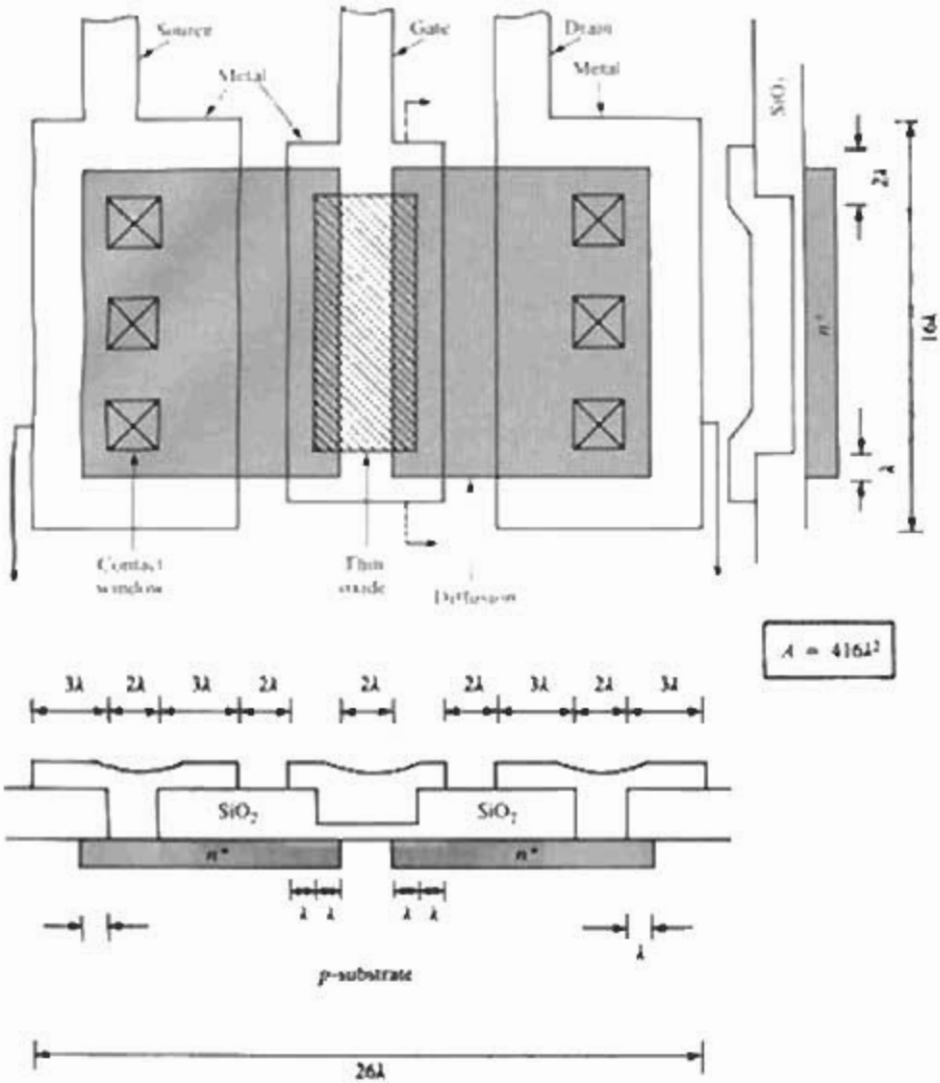


Fig. 9.10 Minimum-size metal-gate transistor with a Z/L ratio of 5/1 using the design rules of Fig. 9.9. The active gate region is less than 5% of the total device area

Some new design rules must be introduced for this process. Polysilicon lines and spaces will both be a minimum feature size of 2λ. The polysilicon gate must overlap the thin oxide region by an alignment tolerance λ. The above alignment sequence requires 1λ polysilicon and 1λ metal borders around contacts. However, contact holes should have a 2λ border of thin oxide due to tolerance accumulation.

Figure 9.11 shows the layout of the polysilicon-gate device with $Z/L = 5/1$ using these design rules. The total area is $168\lambda^2$. The active channel region now represents 12% of the total area, compared with less than 5% for the metal-gate device. The polysilicon gate acts as a barrier material during source/drain implantation and results in "self-alignment" of the edge of the gate to the edge of the source/drain regions. Self-alignment of the gate to the channel reduces the size of the transistor and eliminates the overlap region between the gate and the source/drain regions. In addition, the size of

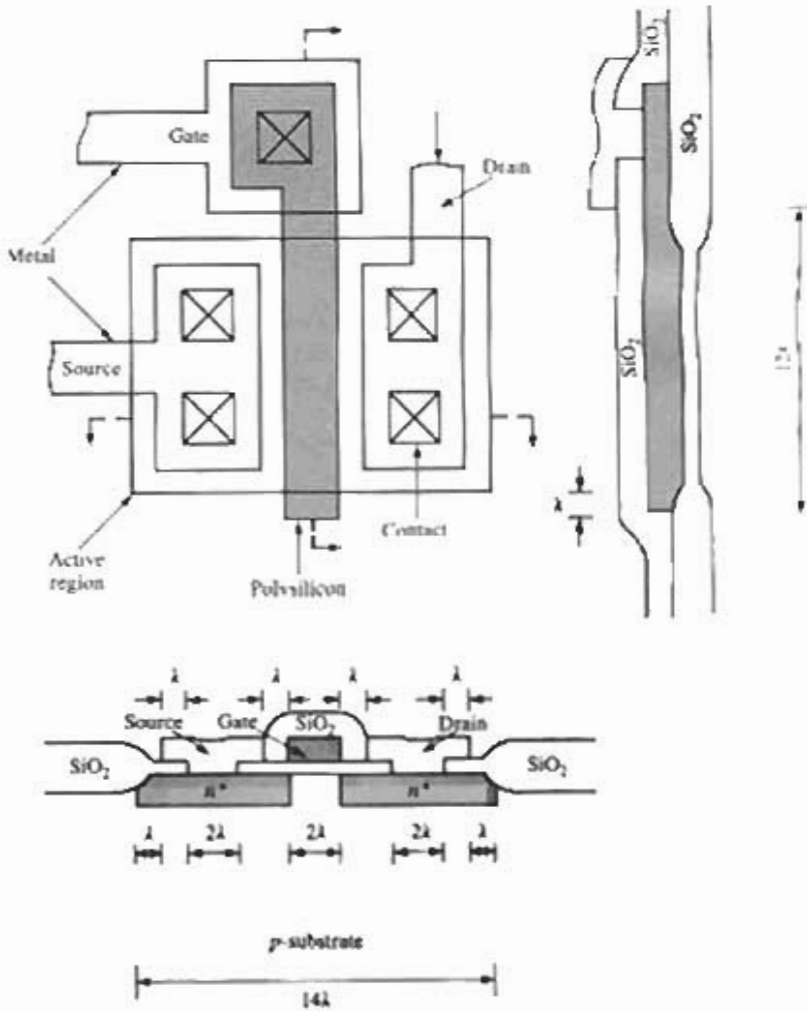


Fig. 9.11 Minimum-size polysilicon-gate transistor layout for $Z/L = 5/1$. The active gate region occupies 12% of the transistor area, and parasitic gate capacitance is minimized.

the transistor is reduced because the source/drain metallization can be placed nearer to the gate. In the polysilicon-gate layout, the channel is defined by the polysilicon gate in the length direction and by the thin oxide in the width direction.

A very important side benefit resulting from this process is the third level of interconnection provided by the polysilicon. Circuit wiring may be accomplished on the diffusion, metal, and polysilicon levels in the polysilicon-gate technology.

A design rule concerning edges has again been introduced into this layout. Metal lines are spaced 1λ from the polysilicon gate to prevent the edge of the metal line from falling directly on top of the edge of the polysilicon line in the nominal layout.

9.2.3 More-Aggressive Design Rules

The design rules discussed so far have focused on minimum feature size and alignment tolerance. F and T are determined primarily by the type of lithography being practiced. However, linewidth expansion and shrinkage throughout the process also strongly affect the ground rules. Expansion or shrinkage may occur during mask fabrication, resist exposure, resist development, etching, or diffusion. These linewidth changes are normally factored into the design rules.

In addition, alignment variation is a statistical process. Worst-case misalignments occur only a very small percentage of the time. (For a Gaussian distribution, a 3σ misalignment occurs only 2% of the time.) Our set of rules based on worst-case alignment tolerances is very pessimistic. For example, assuming that contacts are misaligned by λ in one direction, at the same time that the metal level is misaligned in the opposite direction by λ, results in an accumulated tolerance of 2λ. However, this situation would most probably never occur.

Let us consider the impact of tightening two design rules in the polysilicon-gate process. First, we will let the edge of one layer align with the edge of another layer. Second, a contact window will be allowed to run over onto the field oxide by 1λ. The resulting layout using our polysilicon-gate alignment sequence is shown in Fig. 9.12. The total area of the device has been reduced 25% to $120\lambda^2$, and the active channel region now represents 17% of the total transistor area. We see how ground rule changes can have a substantial effect on device area.

9.2.4 Channel Length and Width Biases

Figure 9.13 presents another example of the interaction of the process with design-rule definitions. Here we will assume a metal-gate process in which the source/drain junction depth is equal to λ and lateral diffusion equals vertical diffusion. Since we know that the source/drain diffusions will move laterally under the edge of the oxide openings, the contact windows can be aligned with the edge of the diffusions at the mask level but will still be 1λ within the border of the diffusion in the final structure.

However, lateral diffusion requires the length of channel at the mask level to be doubled to achieve the same electrical channel length in the device. The actual channel

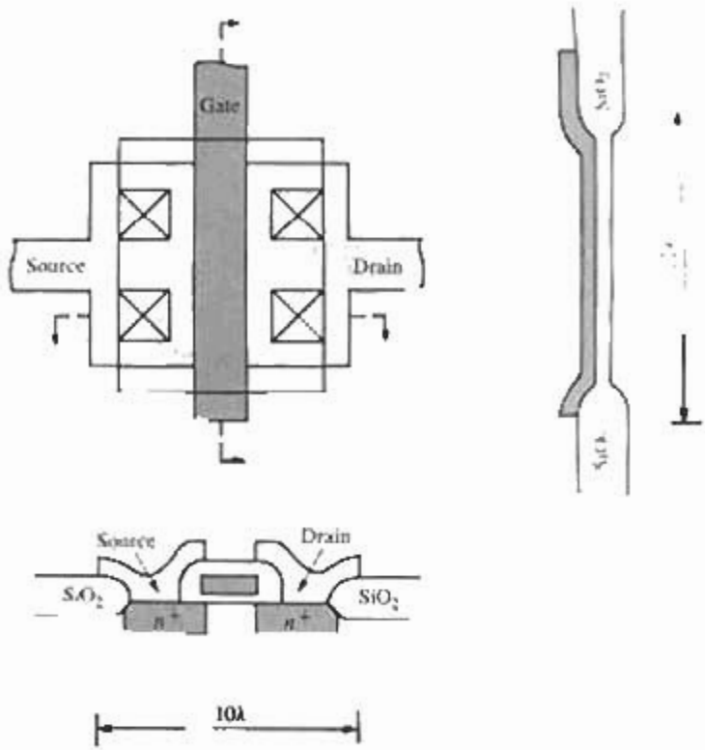


Fig. 9.12 More aggressive layout of the polysilicon-gate transistor in which two ground rules have been relaxed. Active gate area is now 17% of total device area.

length $L = L_m - \Delta L$, where L_m is the channel length as originally drawn on the mask and ΔL is the channel-length shrinkage which occurs during processing. This is an important area where the process must be controlled. For devices with short channel lengths, ΔL may be so severe that the devices become unusable. For the metal-gate layout of Fig. 9.13, $Z_m/L_m = 4\lambda/4\lambda = 1/1$ at the mask level, and $Z/L = 4\lambda/2\lambda = 2/1$ in the fabricated transistor.

The development of self-aligned polysilicon-gate technology with ion-implanted source/drain regions was a major improvement. The polysilicon-gate process eliminates most, but not all, of both the channel shrinkage caused by lateral diffusion and the overlap capacitance resulting from alignment tolerances in the metal-gate process.

In Fig. 9.11, one can see another source of channel bias. The "bird's beak" reduces the size of the active region to below that defined by the active region mask, and it introduces a process bias into the channel width of the polysilicon-gate transistor. $Z = Z_m - \Delta Z$, where Z_m is the width at the mask level and ΔZ is the channel-width shrinkage during processing.

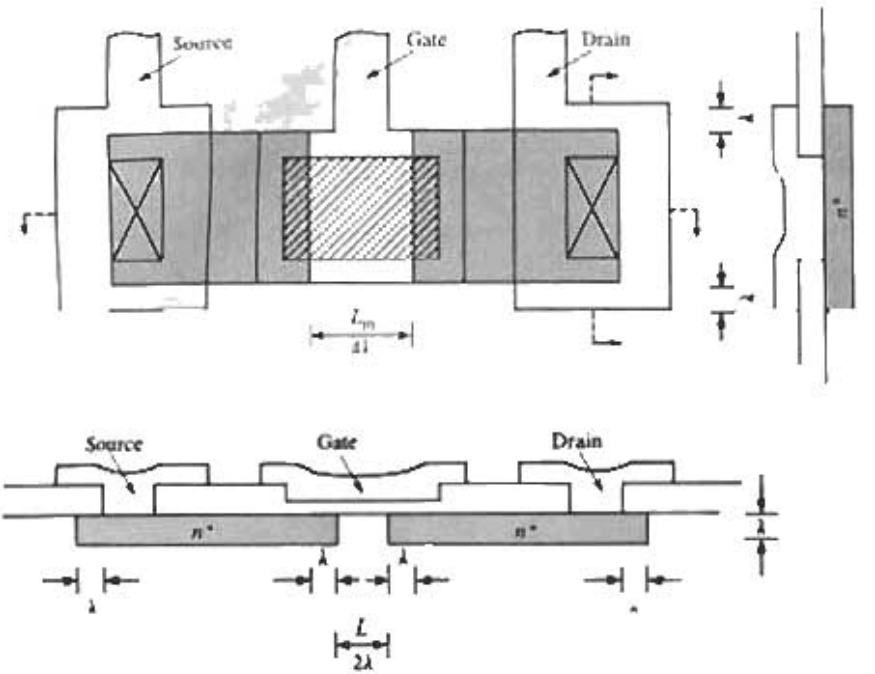


Fig. 9.13 Channel-length bias in a metal-gate NMOS device caused by lateral diffusion under the edge of the diffusion window. The transistor has $Z/L = 1/1$ at the mask level but ends up with an actual $Z/L = 2/1$ after the device is fabricated. Layout of the contact position is based on knowledge of the lateral diffusion which occurs during processing.

In sets of very tight design rules developed for high-volume-production ICs, all critical dimensions are adjusted to account for the processing and alignment sequences. This often results in a layout which must conform to a set of 50 to 100 design rules.[7] Such a set of design rules is highly technology-specific and cannot be transferred from one generation of lithography to the next. The Mead-and-Conway-style rules[7] reach a compromise between a set of rules which is overly pessimistic and wastes a lot of silicon area, and one that is extremely complex but squeezes out all excess area. The Mead-and-Conway-style design rules are being used for low-volume ICs in which design time, and not silicon area, is of dominant importance.

9.3 COMPLEMENTARY MOS (CMOS) TECHNOLOGY

The basic CMOS process of Fig. 1.5 requires a *p*-well diffusion and formation of both NMOS and PMOS transistors. Substrate resistivity is chosen to give the desired PMOS characteristics, and an additional implant step may be introduced to adjust the PMOS

threshold separately. The *p*-well-to-substrate junction may range from a few microns to as much as twenty microns in depth. The net surface concentration of the *p*-well must be high enough above the substrate concentration to provide adequate process control without severely degrading the mobility and threshold voltage of the NMOS transistors. The surface concentration of the *p*-well typically ranges between three and ten times the substrate impurity concentration. An additional implant step is often introduced to adjust the NMOS threshold voltage.

Parasitic bipolar devices are formed in the CMOS process in which merged *pnp* and *npn* transistors form a four-layer (*pnpn*) lateral SCR, as shown in Fig. 9.14. If this SCR is turned on, the device may destroy itself via a condition called *latchup*.[8,9] The *p*-well depth and the spacings between the source/drain regions and the edge of the *p*-well must be carefully chosen to minimize the current gain of the bipolar transistors and the size of the shunting resistors R_s and R_w. A CMOS process will have a number of additional ground rules which are not present in an NMOS or PMOS process. A more detailed discussion of the design of bipolar transistors will be given in Chapter 10.

In order to reduce the resistance of the two shunting resistors, "guard ring" diffusions are sometimes added to the process, as in Fig. 9.14. Guard rings can be formed using the source/drain diffusions of the PMOS and NMOS transistors or can be added as

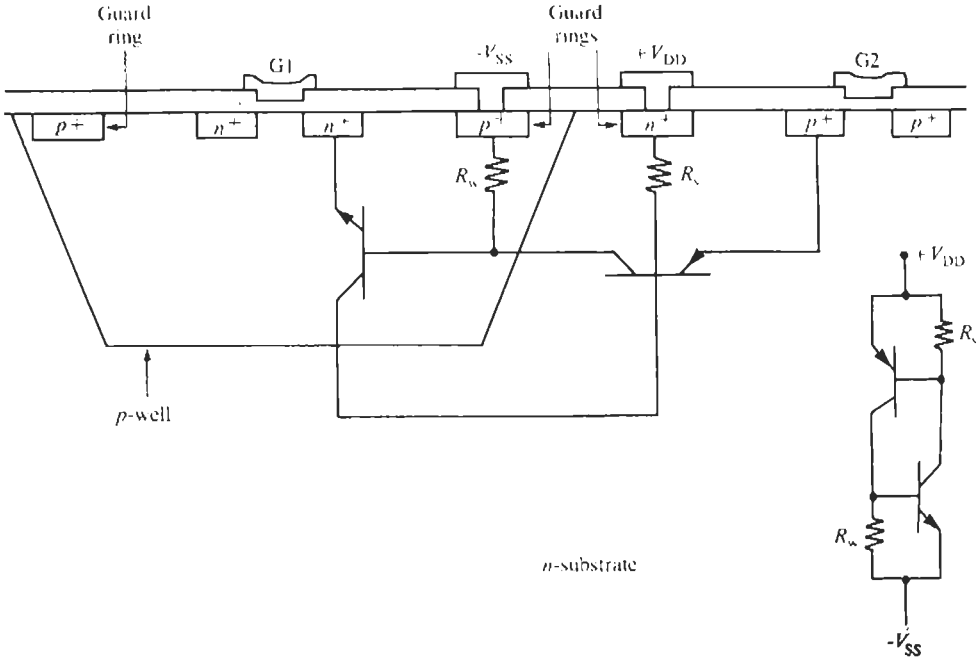


Fig. 9.14 Cross-section of a CMOS structure, showing the existence of a parasitic lateral *pnpn* SCR and the use of guard rings to reduce the value of R_s and R_w.

separate diffusion steps. Recent CMOS processes have used an *n*-well version of this technology, which permits optimization of the NMOS devices fabricated in the original substrate.

Twin-well processes, such as in Fig. 9.15, permit separate optimization of both NMOS and PMOS devices.[10] A lightly doped *n*- or *p*-type epitaxial layer is grown on a heavily doped *n*- or *p*-type substrate. (Lightly doped *n*- and *p*-type regions are often referred to as ν and π regions, respectively.) Separate implantations and diffusions are used to form *wells* for both the NMOS and PMOS transistors. The low-resistivity substrate substantially reduces the substrate resistance R_s and improves latchup resistance.

Example 9.2: A CMOS process uses an *n*-type substrate with a doping of $10^{15}/\text{cm}^3$. An implant/drive-in schedule will be used to form a *p*-well with a net surface concentration of $4 \times 10^{15}/\text{cm}^3$ and a junction depth of 7.5 μm. (**a**) What is the drive-in time at 1150 °C? (**b**) Solve for the implanted dose in silicon. (**c**) What are the threshold voltages of the *n*- and *p*-channel transistors if the oxide thickness is 50 nm?

Solution: The 7.5-μm junction depth and low surface concentration suggest that the well has a Gaussian profile resulting from a two-step diffusion or implant/diffusion process. A final surface concentration of $5 \times 10^{15}/\text{cm}^3$ is required to produce a net concentration of $4 \times 10^{15}/\text{cm}^3$ at the surface. Solving for the Dt product yields

$$Dt = x_j^2/2 \ln(N_0/N_B) = 8.74 \times 10^{-8} \text{ cm}^2$$

At 1150 °C, $D = 8.87 \times 10^{-13}$ cm^2/sec, which gives $t = 27.5$ h. The dose in silicon is given by $Q = N_0\sqrt{\pi Dt} = 2.62 \times 10^{12}/\text{cm}^2$. The *p*-channel devices reside in the *n*-type substrate with a doping concentration of $10^{15}/\text{cm}^3$. From Fig. 9.2, the threshold voltage will be -0.95 V. The deep well diffusion will be almost constant near the surface with a value of $4 \times 10^{15}/\text{cm}^3$. Figure 9.2 yields an *n*-channel threshold of 0.2 V. A threshold adjustment implant would be needed in this process to increase the *n*-channel threshold voltage.

9.4 OTHER MOS STRUCTURES

Chemical etching techniques may be used to crystallographically etch silicon. A solution of KOH, water, and alcohol[11] etches the ⟨100⟩, ⟨110⟩, and ⟨111⟩ crystal planes at relative rates of 40:30:1. This etch may be masked by silicon dioxide or silicon nitride and can be used to etch cavities and V-shaped grooves in ⟨100⟩ silicon (Fig. 9.16).

VMOS technology[12] makes use of the grooves to reduce the channel length and increase the Z/L ratio of the MOS transistor. A basic VMOS process is shown in Fig. 9.17. The channel is formed along the four sides of the groove, and the channel length is determined by the thickness of the epitaxial layer and the diffusions. At the time VMOS was invented, channel lengths achievable with this technology were much shorter than those that could be achieved with normal planar technology, because the epitaxial

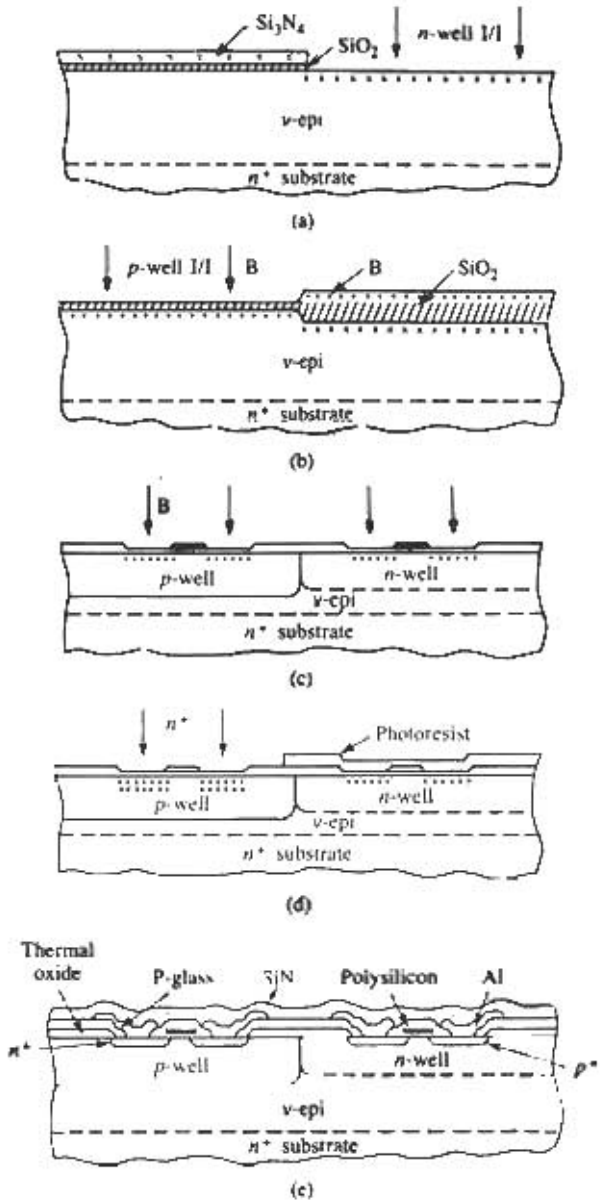


Fig. 9.15 Twin-well CMOS structure at several stages of the process. (a) n-well ion implant; (b) p-well implant; (c) nonselective p^+ source/drain implant; (d) selective n^+ source/drain implant using photoresist mask; (e) final structure. Copyright 1980 IEEE. Reprinted with permission from ref. [10].

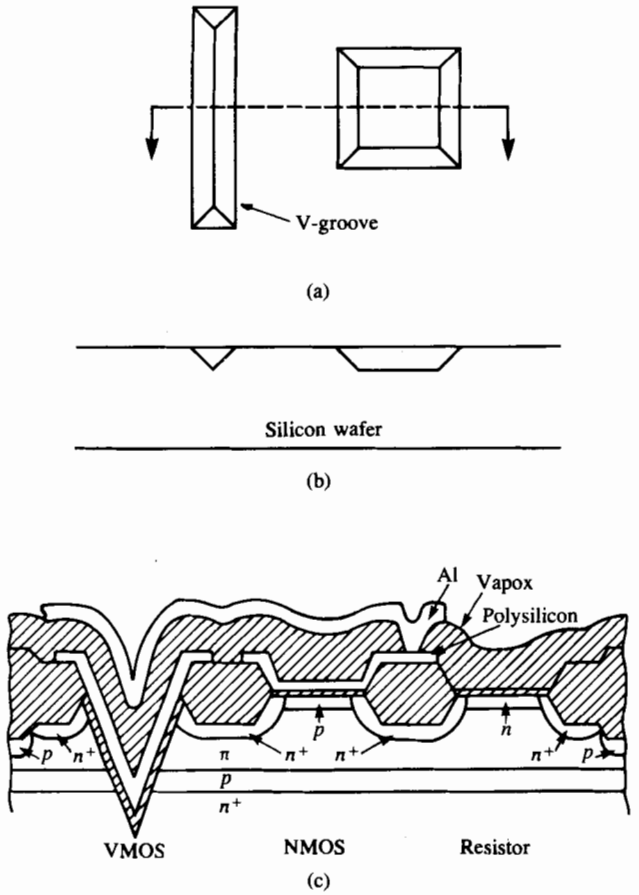


Fig. 9.16 Groove formation in the surface of ⟨100⟩ silicon using anisotropic etching of silicon. (a) Top view; (b) cross section; (c) use of a groove in the formation of a VMOS transistor. Copyright 1977 IEEE. Reprinted from ref. [12] with permission.

layer thickness was not limited by lithographic dimensions. Present-day MOS power transistors have improved and expanded the use of the ideas of the original VMOS structure.[13]

9.5 SUMMARY

In this chapter we have explored the interaction of process design with MOS device characteristics and transistor layout, including the relationships between processing parameters and breakdown voltage, punch-through voltage, threshold voltage, and junction capacitance. A low value of substrate doping is desired to minimize junction

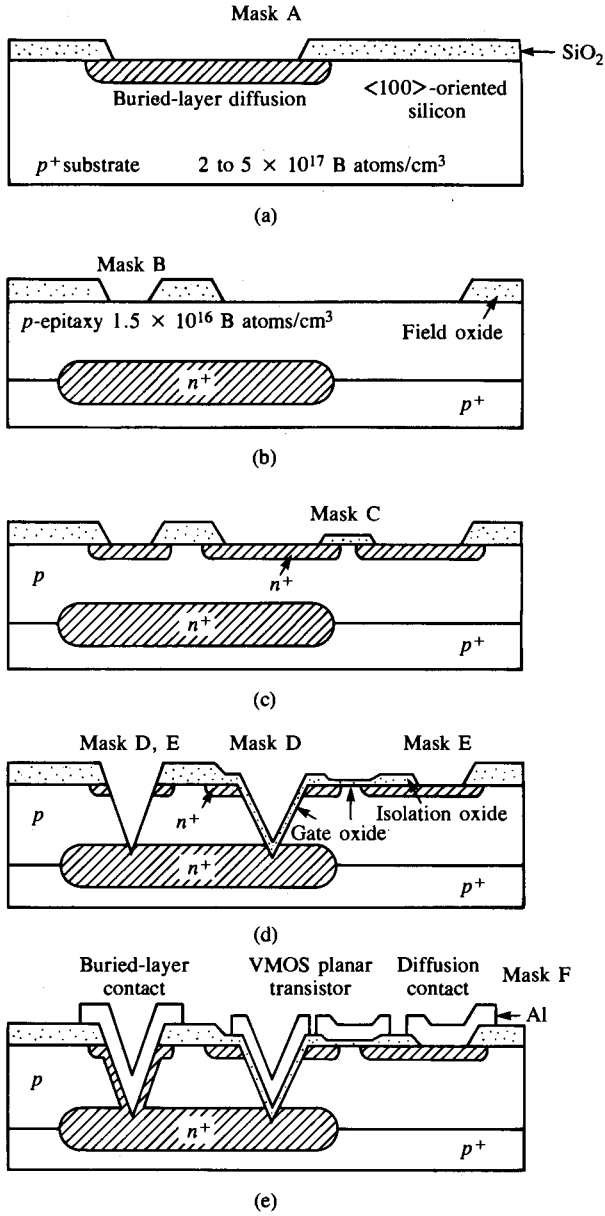


Fig. 9.17 Fabrication sequence for the formation of VMOS transistors for use in memory circuits. (a) Buried-layer diffusion; (b) epitaxial growth; (c) source/drain diffusion; (d) anisotropic etch and oxide growth; (e) metallization and pattern definition. Copyright 1978 IEEE. Reprinted with permission from ref. [17].

capacitance, substrate sensitivity, and junction breakdown voltage, whereas a high substrate doping is needed to maximize punch-through voltage. The use of ion implantation permits the designer to separately tailor the threshold voltage of the transistor.

We have developed basic ideas relating minimum feature size and alignment tolerances and have discussed simple sets of layout design rules. The strong relation between layout design rules and the size of transistors has been demonstrated. Polysilicon-gate technology has been shown to result in a much smaller device area than metal-gate technology for a given transistor Z/L ratio as well as to minimize the parasitic gate capacitance of the device. In addition, the polysilicon-gate process substantially reduces channel-length bias caused by lateral diffusion.

A combination of ion implantation and diffusion is commonly used to form the p- or n-well required for CMOS technology. VLSI CMOS often uses twin-well processes which permit separate optimization of both the n- and p-channel devices.

REFERENCES

[1] R. F. Pierret, *Field Effect Devices,* Volume IV in the Modular Series on Solid State Devices, Addison-Wesley, Reading, MA, 1983.

[2] S. A Abbas and R. C. Dockerty, "N-channel Design Limitations due to Hot Electron Trapping," IEEE IEDM Digest, p. 35–38 (1975).

[3] T. H. Ning, C. M. Osburn, and H. N. Yu, "Threshold Instability in IGFETs due to Emission of Leakage Electrons from Silicon Substrate into Silicon Dioxide," Applied Physics Letters, *29,* 198–199 (1976).

[4] P. E. Cottrell and E. M. Buturla, "Steady State Analysis of Field Effect Transistors via the Finite Element Method," IEEE IEDM Digest, p. 51–54 (1975).

[5] S. M. Sze, *Semiconductor Devices—Physics and Technology,* John Wiley & Sons, New York, 1985.

[6] C. A. Mead and L. Conway, *VLSI Design,* Addison-Wesley, Reading, MA, 1980.

[7] Brian Spinks, *Introduction to Integrated Circuit Layout,* Chapter 7, Prentice-Hall, Englewood Cliffs, NJ, 1985.

[8] A. Ochoa, W. Dawes, and D. Estreich, "Latchup Control in CMOS Integrated Circuits," IEEE Transactions on Nuclear Science, *NS-26,* 5065–5068 (December, 1979).

[9] R. S. Payne, W. N. Grant, and W. J. Bertram, "The Elimination of Latchup in Bulk CMOS," IEEE IEDM Digest, p. 248–251 (December, 1980).

[10] L. C. Parrillo, R. S. Payne, R. E. Davis, G. W. Reutlinger, and R. L. Field, "Twin-Tub CMOS — A Technology for VLSI Circuits," IEEE IEDM Digest, p. 752–755 (December, 1980).

[11] E. Bassous, "Fabrication of Novel Three-Dimensional Microstructures by the Anisotropic Etching of ⟨100⟩ and ⟨110⟩ Silicon," IEEE Transactions on Electron Devices, *ED-25,* 1178–1185 (October, 1978).

[12] T. J. Rodgers, F. B. Jenne, B. Frederick, J. J. Barnes, W. R. Hiltpold, and J. D. Trotter, "VMOS Memory Technology," IEEE ISSCC Digest, p. 74–75 (February, 1977).

[13] B. J. Baliga and D. Y. Chen, *Power Transistors: Device Design and Applications*, IEEE Press, New York, 1984.

[14] K. P. Roenker and L. W. Linholm, "An NMOS Test Chip for a Course in Semiconductor Parameter Measurements," National Bureau of Standards Internal Report 84–2822 (April, 1984).

[15] T. J. Russell, T. F. Leedy, and R. L. Mattis, "A Comparison of Electrical and Visual Alignment Test Structures for Evaluating Photomask Alignment in Integrated Circuit Manufacturing," IEEE IEDM Digest, p. 7A–7F (December, 1977).

[16] D. S. Perloff, "A Four-Point Electrical Measurement Technique for Characterizing Mask Superposition Errors on Semiconductor Wafers," IEEE Journal of Solid-State Circuits, *SC–13*, 436–444 (August, 1978).

[17] K. Hoffman and R. Losehand, "VMOS Technology Applied to Dynamic RAMs," IEEE Journal of Solid-State Circuits, *SC–13*, 617–622 (October, 1978).

PROBLEMS

9.1 What is the maximum gate-to-source voltage that a MOSFET with a 10-nm gate oxide can withstand. Assume that the oxide breaks down at 5 MV/cm and that the substrate voltage is zero.

9.2 Two n^+ diffused lines are running parallel in a substrate doped with 10^{15} boron atoms/cm^3. The substrate is biased to -5 V, and both lines are connected to $+5$ V. Using one-dimensional junction theory, calculate the minimum spacing needed between the lines to prevent their depletion regions from merging.

9.3 Use one-dimensional junction theory to estimate the punch-through voltage of a MOSFET with a channel length of 1 μm. Assume a substrate doping of 3×10^{16}/cm^3 and a substrate bias of 0 V.

9.4 Calculate the threshold voltage for the NMOS transistor with the doping profile shown in Fig. P9.4. Assume an n^+ polysilicon-gate transistor with a gate-oxide thickness of 50 nm.

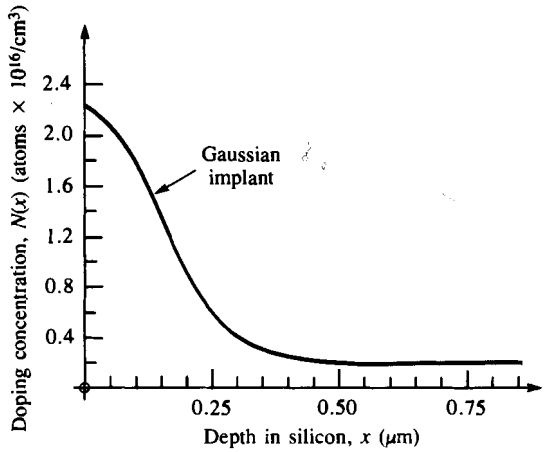


Fig. P9.4

9.5 An n-well CMOS process starts with a substrate doping of $3 \times 10^{15}/\text{cm}^3$. The well doping near the surface is approximately constant at a level of $3 \times 10^{16}/\text{cm}^3$. The gate-oxide thicknesses are both 40 nm.

(a) Calculate the thresholds of the n- and p-channel transistors using eqs. (9.2).

(b) Calculate the boron doses needed to shift the NMOS threshold to $+1$ V and the PMOS threshold to -1 V. Assume that the threshold shifts are achieved through shallow ion implantations. Neglect oxide charge.

9.6 High-performance NMOS logic processes use depletion-mode NMOS transistors for load devices. This requires a negative threshold which can be obtained by implanting a shallow arsenic or phosphorus dose into the channel region. Calculate the arsenic dose needed to achieve a -3-V threshold in an n^+ polysilicon-gate NMOS transistor which has a substrate doping of $3 \times 10^{16}/\text{cm}^3$ and a gate-oxide thickness of 50 nm.

9.7 Draw a composite view of the situation resulting from a worst-case misalignment of the masks for the MOSFET layout shown in Fig. 9.10. Assume metal aligns to thin oxide, and thin oxide and contacts align to the diffusion.

9.8 Develop a new set of ground rules for the metal-gate transistor of Section 9.2, assuming that levels 2, 3, and 4 are all aligned to level 1. Redraw the transistor of Fig. 9.10 using your new rules. In what ways is this layout better or worse than that originally given in Fig. 9.10?

9.9 Draw a cross section of a metal-gate NMOS transistor and a composite view of its mask set, assuming an aggressive layout which takes into account all lateral diffusion. Assume a source/drain junction depth of 2.5 μm and assume that lateral diffusion equals 80% of vertical diffusion. Assume λ is 2 μm and $Z/L = 10/1$.

9.10 Draw the layout of a three-input NMOS NOR-gate with the dimensions given on the circuit schematic in Fig. P9.10. Be sure to merge diffusions wherever possible. Use the more aggressive ground rules developed for polysilicon-gate devices.

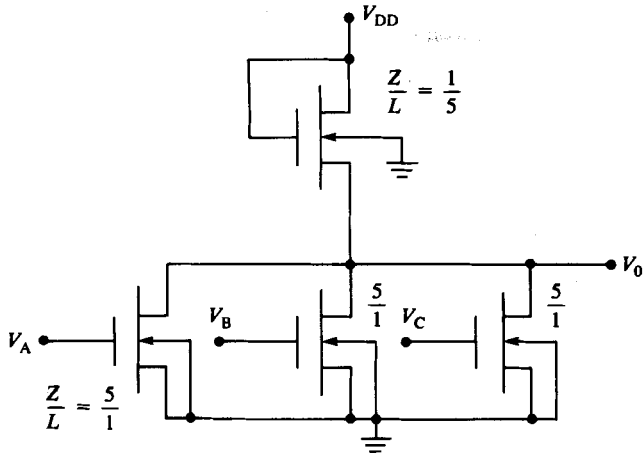


Fig. P9.10

9.11 Our design rule examples used an alignment tolerance which was one-half the feature size. This ratio represents a very loose alignment capability. Develop a new set of design rules similar

to those of Fig. 9.11 for $T = \alpha$ and $F = 4\alpha$. Draw the new minimum-size polysilicon-gate transistor using your rules. Compare the area of your transistor with the area of the transistor of Fig. 9.11 if $\lambda = 2\alpha$.

9.12 An implant with its peak concentration at the silicon surface is used to adjust the threshold of an NMOS transistor. We desire to model this implant by a rectangular approximation similar to that of Figure 9.5. Show that $N_i = N_p \pi/4$ and $x_i = \Delta R_p \sqrt{8/\pi}$ by matching the first two moments of the two impurity distributions.

9.13 A number of types of alignment test structures have been developed.[14, 15] Figure P9.13 shows a simple test structure which can be used to measure the misregistration of the contact

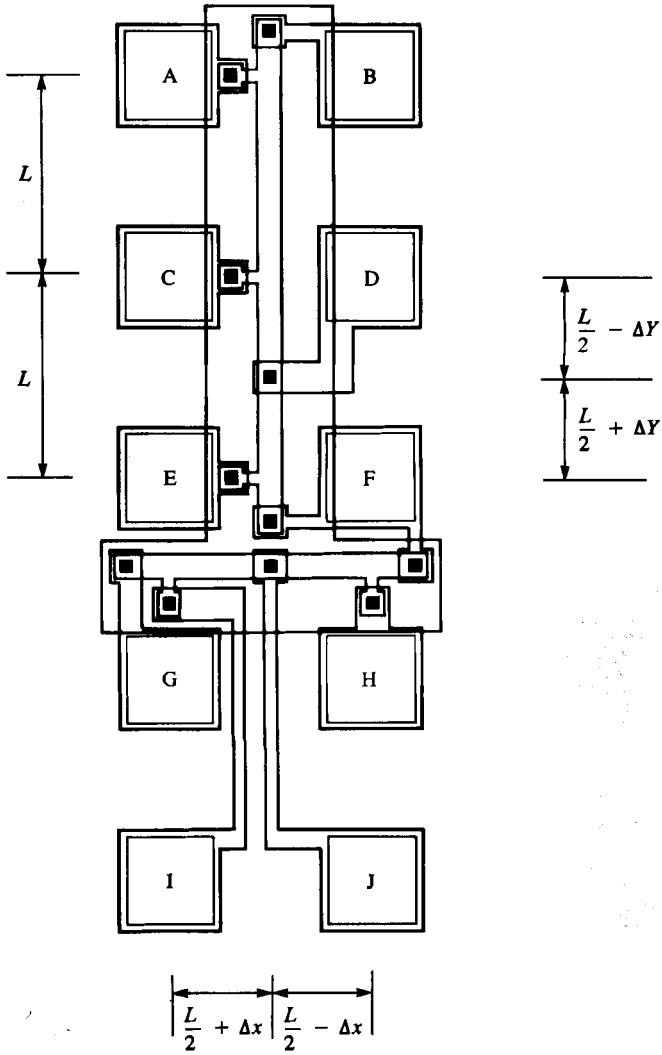


Fig. P9.13

window mask relative to the diffusion mask.[16] Two linear potentiometers, one in the horizontal direction and one in the vertical direction, are fabricated using diffused resistors. The distance between contacts A and C is the same as that between C and E, and the contact from pad D is nominally one-half the distance between pads C and E. A current is injected between pads B and F, and the voltages between pads C-D and D-E are measured.

(a) Show that the misregistration in the y direction is given by $\Delta Y = \frac{1}{2} L\ (V_{DE} - V_{CD})/V_{AC}$.

(b) Derive a similar relationship for misregistration in the x direction.