Chapter

2

The Well

To develop a fundamental understanding of CMOS integrated circuit layout and design, we begin with a study of the well. The well is the first layer fabricated when making a CMOS IC. The approach of studying the details of each fabrication (layout) layer will build a solid foundation for understanding the performance limitations and parasitics (the pn junctions, capacitances, and resistances inherent in a CMOS circuit) of the CMOS process.

The Substrate (The Unprocessed Wafer)

CMOS circuits are fabricated on and in a silicon wafer, as discussed in Ch. 1. This wafer is doped with donor atoms, such as phosphorus for an n-type wafer, or acceptor atoms, such as boron for a p-type wafer. Our discussion centers around a p-type wafer (the most common substrate used in CMOS IC processing). When designing CMOS integrated circuits with a p-type wafer, n-channel MOSFETs (NMOS for short) are fabricated directly in the p-type wafer, while p-channel transistors, PMOS, are fabricated in an "n-well." The substrate or well are sometimes referred to as the bulk or body of a MOSFET. CMOS processes that fabricate MOSFETs in the bulk are known as "bulk CMOS processes." The well and the substrate are illustrated in Fig. 2.1, though not to scale.

Often an epitaxial layer is grown on the wafer. In this book we will not make a distinction between this layer and the substrate. Some processes use a p-well or both n- and p-wells (sometimes called twin tub processes). A process that uses a p-type (n-type) substrate with an n-well (p-well) is called an "n-well process" ("p-well process"). *We will assume, throughout this book, that an n-well process is used for the layout and design discussions.*

A Parasitic Diode

Notice, in Fig. 2.1, that the n-well and the p-substrate form a diode. In CMOS circuits, the substrate is usually tied to the lowest voltage in the circuit (generally, the substrate is grounded) to keep this diode from forward biasing. Ideally, zero current flows in the substrate. We won't concern ourselves with how the substrate is connected to ground at this point (see Ch. 4).

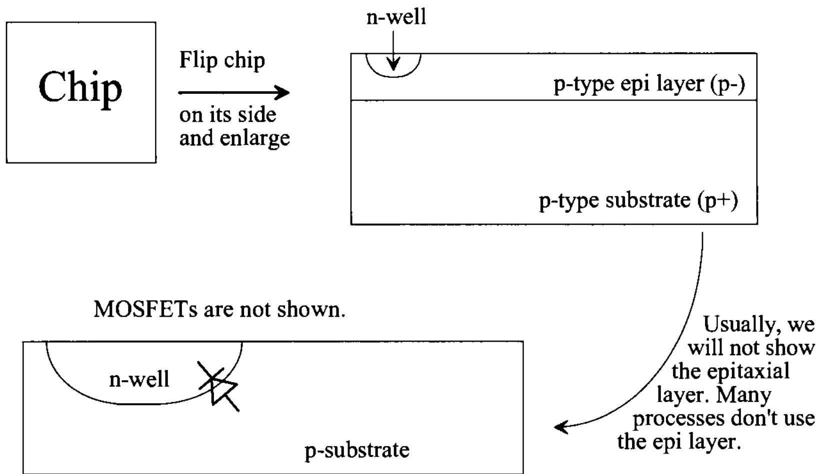


Figure 2.1 The top (layout) and side (cross-sectional) view of a die.

Using the N-well as a Resistor

In addition to being used as the body for p-channel transistors, the n-well can be used as a resistor, Fig. 2.2. The voltage on either side of the resistor must be large enough to keep the substrate/well diode from forward biasing.

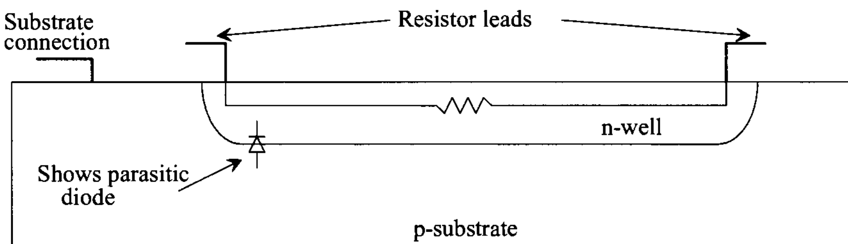


Figure 2.2 The n-well can be used as a resistor.

2.1 Patterning

CMOS integrated circuits are formed by patterning different layers on and in the silicon wafer. Consider the following sequence of events that apply, in a fundamental way, to any layer that we need to pattern. We start out with a clean, bare wafer, as shown in Fig. 2.3a. The distance given by the line A to B will be used as a reference in Figs. 2.3b–j. Figures 2.3b–j are cross-sectional views of the dashed line shown in (a). The small box in Fig. 2.3a is drawn with a layout program (and used for mask generation) to indicate where to put the patterned layer.

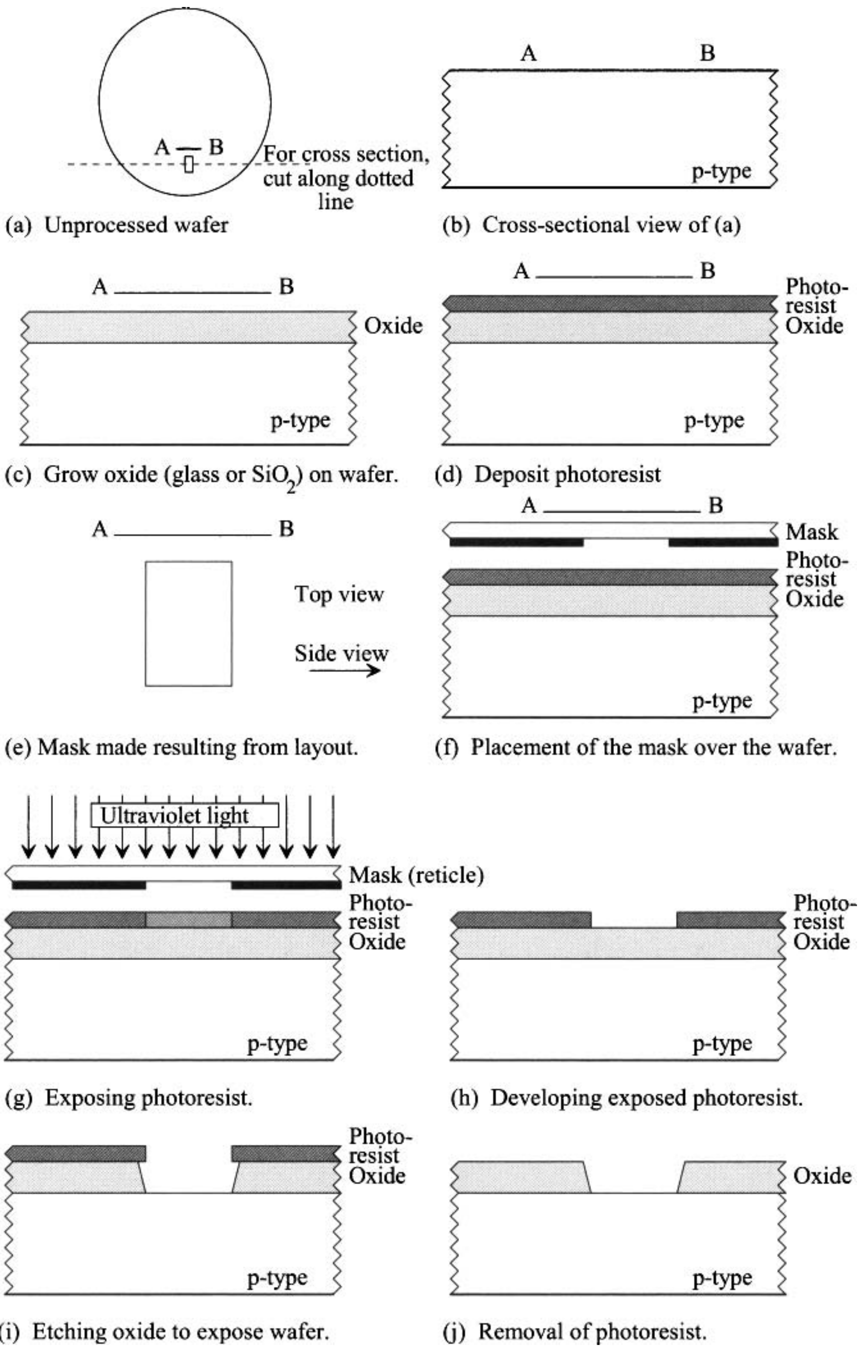


Figure 2.3 Generic sequence of events used in photo patterning.

The first step in our generic patterning discussion is to grow an oxide, SiO_2 or glass, a very good insulator, on the wafer. Simply exposing the wafer to air yields the reaction $Si + O_2 \rightarrow SiO_2$. However, semiconductor processes must have tightly controlled conditions to precisely set the thickness and purity of the oxide. We can grow the oxide using a reaction with steam, H_2O, or with O_2 alone. The oxide resulting from the reaction with steam is called a wet oxide, while the reaction with O_2 is a dry oxide. Both oxides are called thermal oxides due to the increased temperature used during oxide growth. The growth rate increases with temperature. The main benefit of the wet oxide is fast growing time. The main drawback of the wet oxide is the hydrogen byproduct. In general terms, the oxide grown using the wet techniques is not as pure as the dry oxide. The dry oxide generally takes a considerably longer time to grow. Both methods of growing oxide are found in CMOS processes. An important observation we should make when looking at Fig. 2.3c is that the oxide growth actually consumes silicon. This is illustrated in Fig. 2.4. The overall thickness of the oxide is related to thickness of the consumed silicon by

$$x_{Si} = 0.45 \cdot x_{ox} \tag{2.1}$$

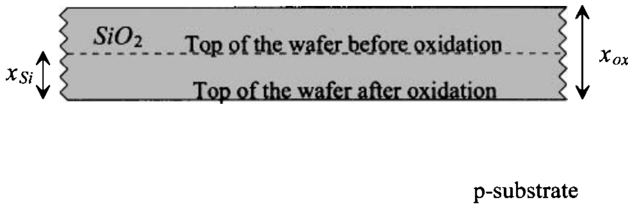


Figure 2.4 How growing oxide consumes silicon.

The next step of the generic CMOS patterning process is to deposit a photosensitive resist layer across the wafer (see Fig. 2.3d). Keep in mind that the dimensions of the layers, that is, oxide, resist, and the wafer, are not drawn to scale. The thickness of a wafer is typically 500 μm, while the thickness of a grown oxide or a deposited resist may be only a μm (10^{-6} m) or even less. After the resist is baked, the mask derived from the layout program, Figs. 2.3e and f, is used to selectively illuminate areas of the wafer, Fig. 2.3g. In practice, a single mask called a reticle, with openings several times larger than the final illuminated area on the wafer, is used to project the pattern and is stepped across the wafer with a machine called a stepper to generate the patterns needed to create multiple copies of a single chip. The light passing through the opening in the reticle is photographically reduced to illuminate the correct size area on the wafer.

The photoresist is developed (Fig. 2.3h), removing the areas that were illuminated. This process is called a positive resist process because the area that was illuminated was removed. A negative resist process removes the areas of resist that were not exposed to the light. Using both types of resist allows the process designer to cut down on the number of masks needed to define a CMOS process. Because creating the masks is expensive, lowering the number of masks is equated with lowering the cost of a process. This is also important in large manufacturing plants where fewer steps equal lower cost.

The next step in the patterning process is to remove the exposed oxide areas (Fig. 2.3i). Notice that the etchant etches under the resist, causing the opening in the oxide to be larger than what was specified by the mask. Some manufacturers intentionally bloat (make larger) or shrink (make smaller) the masks as specified by the layout program. Figure 2.3j shows the cross-sectional view of the opening after the resist has been removed.

2.1.1 Patterning the N-well

At this point we can make an n-well by diffusing donor atoms, those with five valence electrons, as compared to the 4 four found in silicon, into the wafer. Referring to our generic patterning discussion given in Fig. 2.3, we begin by depositing a layer of resist directly on the wafer, Fig. 2.3d (without oxide). This is followed by exposing the resist to light through a mask (Figs. 2.3f and g) and developing or removing the resist (Fig. 2.3h). The mask used is generated with a layout program. The next step in fabricating the n-well is to expose the wafer to donor atoms. The resist blocks the diffusion of the atoms, while the openings allow the donor atoms to penetrate into the wafer. This is shown in Fig. 2.5a. After a certain amount of time, depending on the depth of the n-well desired, the diffusion source is removed (Fig. 2.5b). Notice that the n-well "outdiffuses" under the resist; that is, the final n-well size is not the same as the mask size. Again, the foundry where the chips are fabricated may bloat or shrink the mask to compensate for this lateral diffusion. The final step in making the n-well is the removal of the resist (Fig. 2.5c).

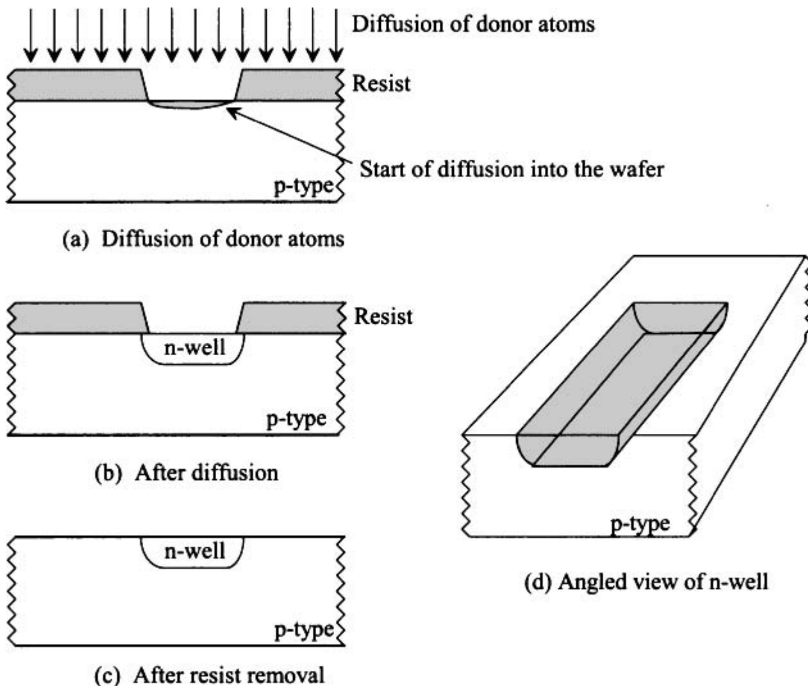


Figure 2.5 Formation of the n-well.

2.2 Laying Out the N-well

When we lay out the n-well, we are viewing the chip from the top. One of the key points in this discussion, as well as the discussions to follow, is that we do layout to a generic **scale factor**. If, for example, the minimum device dimensions are 50 nm (= 0.05 μm = 50×10^{-9} m), then an n-well box **drawn** 10 by 10, see Fig. 2.6, has an actual size after it is fabricated of $10 \cdot 50\, nm$ or half a micron (0.5 μm = 500 nm), neglecting lateral diffusion or other process imperfections. We scale the layout when we generate the GDS (calma stream format) or CIF (Caltech-intermediate-format) file from a layout program. (A GDS or CIF file is what the mask maker uses to make reticles.) Using integers to do layout simplifies things. We'll see in a moment that many electrical parameters are ratios (such as resistance) and so the scale factor cancels out of the ratio.

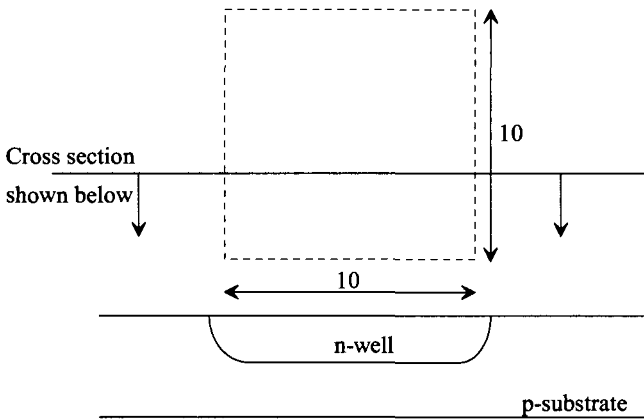


Figure 2.6 Layout and cross-sectional view of a 10 by 10 (drawn) n-well.

2.2.1 Design Rules for the N-well

Now that we've laid out the n-well (drawn a box in a layout program), we might ask the question, "Are there any limitations or constraints on the size and spacing of the n-wells?" That is to say, "Can we make the n-well 2 by 2?" Can we make the distance between the n-wells 1? As we might expect, there are minimum spacing and size requirements for all layers in a CMOS process. Process engineers, who design the integrated circuit process, specify the design rules. The design rules vary from one process technology (say a process with a scale factor of 1 μm) to another (say a process with a scale factor of 50 nm).

Figure 2.7 shows sample design rules for the n-well. The minimum size (width or length) of any n-well is 6, while the minimum spacing between different n-wells is 9. As the layout becomes complicated, the need for a program that ensures that the design rules are not violated is needed. This program is called a *design rule checker* program (DRC program). Note that the minimum size may be set by the quality of patterning the resist (as seen in Fig. 2.5), while the spacing is set by the parasitic npn transistor seen in Fig. 2.7. (We don't want the n-wells interacting to prevent the parasitic npn from turning on.)

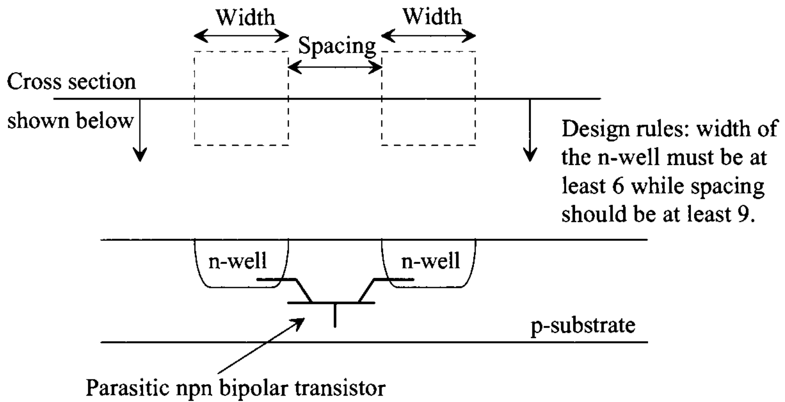


Figure 2.7 Sample design rules for the n-well.

2.3 Resistance Calculation

In addition to serving as a region in which to build PMOS transistors (called the body or bulk of the PMOS devices), n-wells are often used to create resistors. The resistance of a material is a function of the material's resistivity, ρ, and the material's dimensions. For example, the slab of material in Fig. 2.8 between the two leads has a resistance given by

$$R = \frac{\rho}{t} \cdot \frac{L \cdot scale}{W \cdot scale} = \frac{\rho}{t} \cdot \frac{L}{W} \tag{2.2}$$

In semiconductor processing, all of the fabricated thicknesses, t, seen in a cross-sectional view, such as the n-well's, are fixed in depth (**this is important**). When doing layout, we only have control over W (width) and L (length) of the material. The W and L are what we see from the top view, that is, the layout view. We can rewrite Eq. (2.2) as

$$R = R_{square} \cdot \frac{L}{W} \rightarrow R_{square} = \frac{\rho}{t} \tag{2.3}$$

R_{square} is the *sheet resistance* of the material in Ω/square (noting that when $L = W$ the layout is square and $R = R_{square}$).

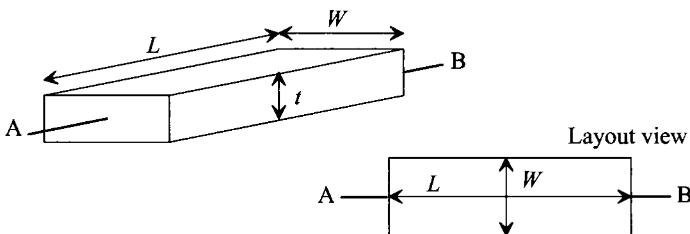


Figure 2.8 Calculation of the resistance of a rectangular block of material.

Example 2.1
Calculate the resistance of an n-well that is 10 wide and 100 long. Assume that the n-well's sheet resistance is typically 2 kΩ/square; however, it can vary with process shifts from 1.6 to 2.4 kΩ/square.

The typical resistance, using Eq. (2.3), between the ends of the n-well is

$$R = 2{,}000 \cdot \frac{100}{10} = 20\ k\Omega$$

The maximum value of the resistor is 24k, while the minimum value is 16k. ■

Layout of Corners

Often, to minimize space, resistors are laid out in a serpentine pattern. The corners, that is, where the layer bends, are not rectangular. This is shown in Fig. 2.9a. All sections in Fig 2.9a are square, so the resistance of sections 1 and 3 is R_{square}. The equivalent resistance of section 2 between the adjacent sides, however, is approximately 0.6 R_{square}. The overall resistance between points A and B is therefore $2.6 \cdot R_{square}$. As seen in Ex. 2.1 the actual resistance value varies with process shifts. The layout shown in Fig. 2.9b uses wires to connect separate sections of unit resistors to avoid corners. Avoiding corners in a resistor is the (generally) preferred method of layout in analog circuit design where the ratio of two resistors is important. For example, the gain of an op-amp circuit may be R_F/R_I.

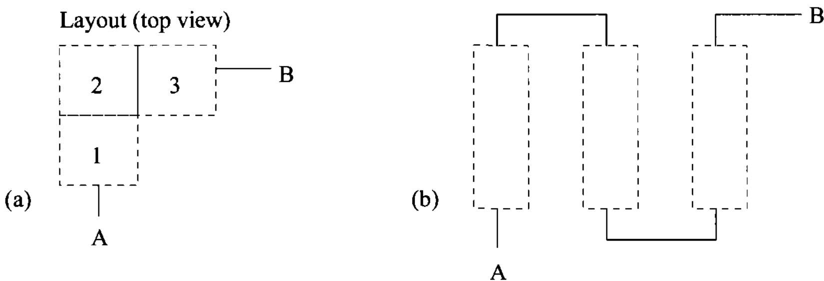


Figure 2.9 (a) Calculating the resistance of a corner section and (b) layout to avoid corners.

2.3.1 The N-well Resistor

At this point, it is appropriate to show the actual cross-sectional view of the n-well after all processing steps are completed (Fig. 2.10). The n+ and p+ implants are used to increase the threshold voltage of the field devices; more will be said on this in Ch. 7. In all practical situations, the sheet resistance of the n-well is measured with the field implant in place, that is, with the n+ implant between the two metal connections in Fig. 2.10. Not shown in Fig. 2.10 is the connection to substrate. The field oxide (FOX; also known as ROX or recessed oxide) are discussed in Chs. 4 and 7 when we discuss the active and poly layers. The reader shouldn't, at this point, feel they should understand any of the cross-sectional layers in Fig. 2.10 except the n-well. Note that the field implants *aren't drawn* in the layout and so their existence is transparent to the designer.

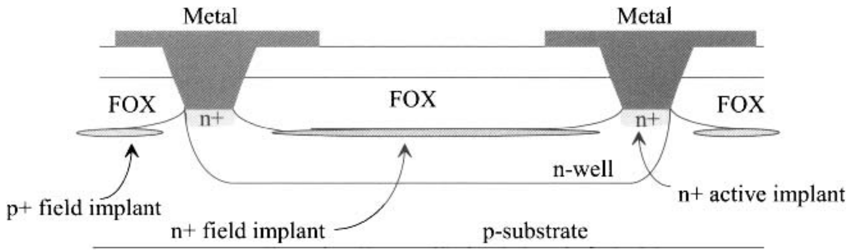


Figure 2.10 Cross-sectional view of n-well showing field implant. The field implantation is sometimes called the "channel stop implant."

2.4 The N-well/Substrate Diode

As seen in Fig. 2.1, placing an n-well in the p-substrate forms a diode. It is important to understand how to model a diode for hand calculations and in SPICE simulations. In particular, let's discuss general diodes using the n-well/substrate pn junction as an example. The DC characteristics of the diode are given by the Shockley diode equation, or

$$I_D = I_S\left(e^{\frac{V_d}{nV_T}} - 1\right) \tag{2.4}$$

The current I_D is the diode current; I_S is the scale (saturation) current; V_d is the voltage across the diode where the anode, A, (p-type material) is assumed positive with respect to the cathode, K, (n-type); and V_T is the *thermal voltage*, which is given by $\frac{kT}{q}$ where k = Boltzmann's constant (1.3806×10^{-23} Joules per degree Kelvin), T is temperature in Kelvin, n is the emission coefficient (a term that is related to the doping profile and affects both the exponential behavior of the diode and the diode's turn-on voltage), and q is the electron charge of 1.6022×10^{-19} coulombs. The scale current and thus the overall diode current are related in SPICE by an area factor (not associated with or to be confused with the scale term we use in layouts, Eq. (2.2)). The SPICE (Simulation Program with Integrated Circuit Emphasis) circuit simulation program assumes that the value of I_s supplied in the model statement was measured for a device with a reference area of 1. If an area factor of 2 is supplied for a diode, then I_S is doubled in Eq. (2.4).

2.4.1 A Brief Introduction to PN Junction Physics

A conducting material is made up of atoms that have easily shared orbiting electrons. As a simple example, copper is a better conductor than aluminum because the copper atom's electrons aren't as tightly coupled to its nucleus allowing its electrons to move around more easily. An insulator has, for example, eight valence electrons tightly coupled to the atom's nucleus. A significant electric field is required to break these electrons away from their nucleus (and thus for current conduction). A semiconductor, like silicon, has four valence electrons. Silicon's conductivity falls between an insulator and a conductor (and thus the name "semiconductor"). As silicon atoms are brought together, they form both a periodic crystal structure and bands of energy that restrict the allowable energies an electron can occupy. At absolute zero temperature, (T = 0 K), all of the valence electrons in the semiconductor crystal reside in the valence energy band, E_v. As temperature

increases, the electrons gain energy (heat is absorbed by the silicon crystal), which causes some of the valence electrons to break free and move to a conducting energy level, E_c. Figure 2.11 shows the movement of an electron from the valence band to the conduction band. Note that there aren't any allowable energies between E_v and E_c in the silicon crystal structure (if the atom were by itself, that is, not in a crystal structure this exact limitation isn't present). Further note that when the **electron** moves from the valence energy band to the conduction energy band, a **hole** is left in the valence band. Having an electron in the conduction band increases the material's conductivity (the electron can move around easily in the semiconductor material because it's not tightly coupled to an atom's nucleus). At the same time a hole in the valence band increases the material's conductivity (electrons in the valence band can move around more easily by simply falling into the open hole). The *key point* is that increasing the number of electrons or holes increases the materials conductivity. Since the hole is more tightly coupled to the atom's nucleus (actually the electrons in the valence band), its mobility (ability to move around) is lower than the electron's mobility in the conduction band. This point is **fundamentally important**. The fact that the mobility of a hole is lower than the mobility of an electron (in silicon) results in, among other things, the size of PMOS devices being larger than the size of NMOS devices (when designing circuits) in order for each device to have the same drive strength.

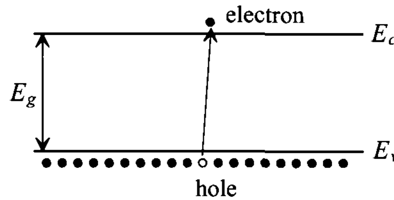


Figure 2.11 An electron moving to the conduction band, leaving behind a hole in the valence band.

Carrier Concentrations

Pure silicon is often called *intrinsic* silicon. As the temperature of the silicon crystal is increased it absorbs heat. Some of the electrons in the valence band gain enough energy to jump the bandgap energy of silicon, E_g (see also Eq. (23.21)), as seen in Fig. 2.11. This movement of an electron from the valence band to the conduction band is called *generation*. When the electron loses energy and falls back into the valence band, it is called *recombination*. The time the electron spends in the conduction band, before it recombines (drops back to the valence energy band), is random and often characterized by the carrier lifetime, τ_T(a root-mean-square, RMS, value of the random times the electrons spend in the conduction band of the silicon crystal). While discussing the actual processes involved with generation-recombination (GR) is outside the scope of this book, the carrier lifetime is a practical important parameter for circuit design. Another important parameter is the number of electrons in the conduction band (and thus the number of holes in the valence band) at a given time (again a random number). These carriers are called intrinsic carriers, n_i. At room temperature

$$n_i \approx 14.5 \times 10^9 \; carriers/cm^3 \qquad (2.5)$$

noting that cm^3 indicates a volume. If we call the number of free electrons (meaning electrons excited up in the conduction band of silicon) n and the number of holes p, then for *intrinsic silicon*,

$$n = p = n_i \approx 14.5 \times 10^9 \ carriers/cm^3 \quad (2.6)$$

This may seem like a lot of carriers. However, the number of silicon atoms, N_{Si}, in a given volume of crystalline silicon is

$$N_{Si} = 50 \times 10^{21} \ atoms/cm^3 \quad (2.7)$$

so there is only one excited electron/hole pair for (roughly) every 10^{12} silicon atoms.

Next let's add different materials to intrinsic silicon (called *doping* the silicon) to change silicon's electrical properties. If we add a small amount of a material containing atoms with five valence electrons like phosphorous (silicon has four), then the added atom would bond with the silicon atoms and the *donated* electron would be free to move around (and easily excited to the conduction band). If we call the density of this added donor material N_D with units of atoms/cm^3 and we assume the number of atoms added to the silicon is much larger than the intrinsic carrier concentration, then we can write the number of free electrons (the electron concentration n) in the material as

$$n \approx N_D \text{ when } N_{Si} >> N_D >> n_i \quad (2.8)$$

A material with added donor atoms is said to be an "n-type" material. Similarly, if we were to add a small material to silicon with atoms having three valence electrons (like boron), the added material would bond with the silicon resulting in a hole in the valence band. Again, this increases the conductivity of silicon because, now, the electrons in the valence band can move into the hole (having the effect of making it look like the hole is moving). The added material in this situation is said to be an *acceptor* material. The added material accepts an electron from the silicon crystal. If the density of the added acceptor material is labeled N_A, then the hole concentration, p, in the material is

$$p \approx N_A \text{ when } N_{Si} >> N_A >> n_i \quad (2.9)$$

A material with added acceptor atoms is said to be a "p-type" material.

If we dope a material with donor atoms, the number of free electrons in the material, n, goes up, as indicated by Eq. (2.8). We would expect, then, the number of free holes in the material to go down (some of those free electrons fall easily into the available holes reducing the number of holes in the material). The relationship between the number of holes, electrons, and intrinsic carrier concentration, is governed by the *mass-action law*

$$pn = n_i^2 \quad (2.10)$$

Consider the following example.

Example 2.2
Suppose silicon is doped with phosphorous having a density, N_D, of 10^{18} atoms/cm^3. Estimate the doped silicon's hole and electron concentration.

The electron concentration, from Eq. (2.8) is, $n = 10^{18}$ electrons/cm^3 (one electron for each donor atom). The hole concentration is found using the mass-action law as

$$p = \frac{n_i^2}{n} = \frac{(14.5 \times 10^9)^2}{10^{18}} = 210 \text{ holes/cm}^3$$

Basically, all of the holes are filled. Note that with a doping density of 10^{18} there is one dopant atom for every 50,000 silicon atoms. If we continue to increase the doping concentration, our assumption that $N_{Si} >> N_D$ isn't valid and the material is said to be *degenerate* (no longer mainly silicon). A degenerate semiconductor doesn't follow the mass-action law (or any of the equations for silicon we present). ■

Fermi Energy Level

To describe the carrier concentration in a semiconductor, the Fermi energy level is often used. The Fermi energy level is useful when determining the contact potentials in materials. For example, the potential that you have to apply across a diode before it turns on is set by the p-type and n-type material contact potential difference. Also, the threshold voltage is determined, in part, by contact potentials.

The Fermi energy level simply indicates the energy level where the probability of occupation by a free electron is 50%. Figure 2.12a shows that for intrinsic silicon the (intrinsic) Fermi level, E_i is close to the middle of the bandgap. In p-type silicon, the Fermi level, E_f, moves towards the valence band, Fig. 2.12b, since the number of free electrons, n , is reduced with the abundance of holes. Figure 2.12c shows the location of the Fermi energy level in n-type silicon. The Fermi level moves towards E_c with the abundance of electrons in the conduction band.

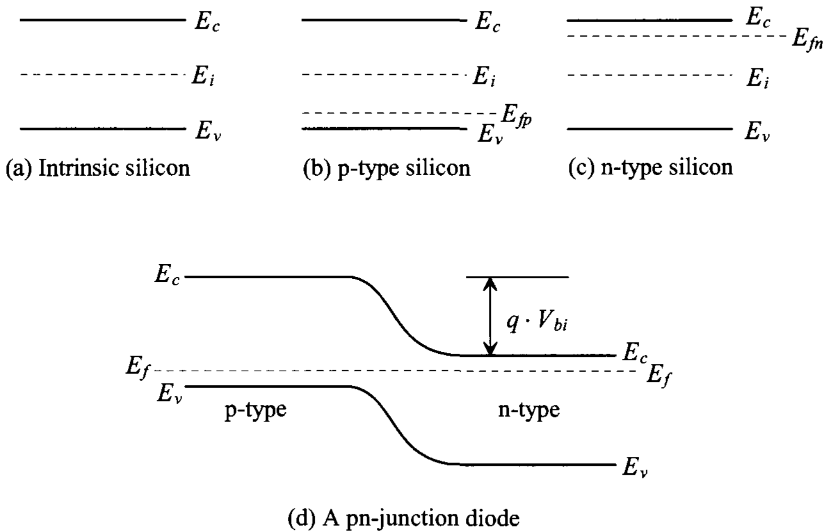


Figure 2.12 The Fermi energy levels in various structures.

The energy difference between the E_i and E_f is given, for a p-type semiconductor, by

$$E_i - E_{fp} = kT \cdot \ln \frac{N_A}{n_i} \tag{2.11}$$

and for an n-type semiconductor by

$$E_{fn} - E_i = kT \cdot \ln \frac{N_D}{n_i} \tag{2.12}$$

The band diagram of a pn junction (a diode) is seen in Fig. 2.12d. Note how the Fermi energy level is constant throughout the diode. A variation in E_f would indicate a nonequilibrium situation (the diode has an external voltage applied across it). To get current to flow in a diode, we must apply an external potential that approaches the diode's contact potential (its built-in potential, V_{bi}). By applying a potential to forward bias the diode, the conduction energy levels in each side of the diode move closer to the same level. The voltage applied to the diode when the conduction energy levels are exactly at the same level is given by

$$V_{bi} = \frac{E_{fn} - E_{fp}}{q} = \frac{kT}{q} \cdot \ln \frac{N_A N_D}{n_i^2} \tag{2.13}$$

noting that $kT/q = V_T$ is the thermal voltage.

2.4.2 Depletion Layer Capacitance

We know that n-type silicon has a number of mobile electrons, while p-type silicon has a number of mobile holes (a vacancy of electrons in the valence band). Formation of a pn junction results in a depleted region at the p-n interface (Fig. 2.13). A depletion region is an area depleted of mobile holes or electrons. The mobile electrons move across the junction, leaving behind fixed donor atoms and thus a positive charge. The movement of holes across the junction, to the right in Fig. 2.13, occurs for the p-type semiconductor as well with a resulting negative charge. The fixed atoms on each side of the junction within the depleted region exert a force on the electrons or holes that have crossed the junction. This equalizes the charge distribution in the diode, preventing further charges from crossing the diode junction and also gives rise to a parasitic capacitance. This parasitic capacitance is called a *depletion* or *junction* capacitance.

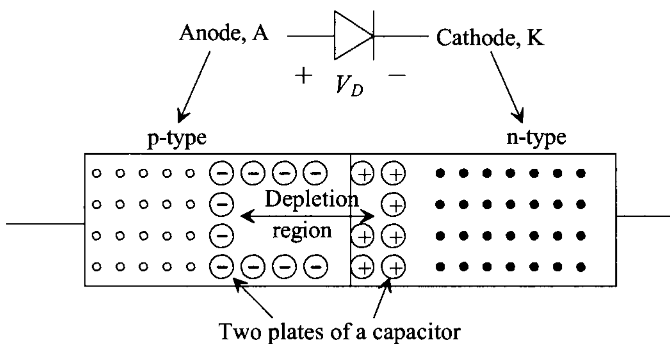


Figure 2.13 Depletion region formation in a pn junction.

The depletion capacitance, C_j, of a pn junction is modeled using

$$C_j = \frac{C_{j0}}{\left[1 - \left(\frac{V_D}{V_{bi}}\right)\right]^m} \tag{2.14}$$

C_{j0} is the zero-bias capacitance of the pn junction, that is, the capacitance when the voltage across the diode is zero. V_D is the voltage across the diode, m is the grading coefficient (showing how the silicon changes from n- to p-type), and V_{bi} is the built-in potential given by Eq. (2.13).

Example 2.3

Schematically sketch the depletion capacitance of an n-well/p-substrate diode 100 × 100 square (with a scale factor of 1 μm), given that the substrate doping is 10^{16} atoms/cm^3 and the well doping is 10^{17} atoms/cm^3. The measured zero-bias depletion capacitance of the junction is 100 aF/μm^2 (= $100 \times 10^{-18}\ F/\mu m^2$), and the grading coefficient is 0.333. Assume the depth of the n-well is 3 μm.

The n-well doping (n-type side of the diode) is $N_D = 10^{17}$, while the substrate doping is N_A is 10^{16}. We can calculate the built-in potential using Eq. (2.13)

$$V_{bi} = 26mV \cdot \ln \frac{10^{16} \cdot 10^{17}}{(14.5 \times 10^9)^2} = 759\ mV$$

The depletion capacitance is made up of a *bottom* component and a *sidewall* component, as shown in Fig. 2.14 (see Eq. (5.17) for the more general form of Eq. (2.14)).

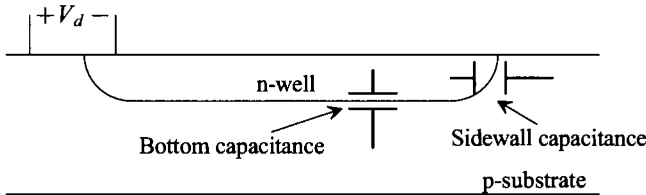


Figure 2.14 A pn junction on the bottom and sides of the junction.

The bottom zero-bias depletion capacitance, C_{j0b}, is given by

$$C_{j0b} = (capacitance/area) \cdot (scale)^2 \cdot (bottom\ area) \tag{2.15}$$

which, for this example, is

$$C_{j0b} = (100\ aF/\mu m^2) \cdot (1\ \mu m)^2 \cdot (100)^2 = 1\ pF$$

The sidewall zero-bias depletion capacitance, C_{j0s}, is given by

$$C_{j0s} = (capacitance/area) \cdot (depth\ of\ the\ well) \cdot (perimeter\ of\ the\ well) \cdot (scale)^2$$

or (2.16)

$$C_{j0s} = (100\ aF/\mu m^2) \cdot (3) \cdot (400) \cdot (1\ \mu m)^2 = 120\ fF$$

The total diode depletion capacitance between the n-well and the p-substrate is the parallel combination of the bottom and sidewall capacitances, or

$$C_j = \frac{C_{j0b}}{\left[1 - \left(\frac{V_D}{V_{bi}}\right)\right]^m} + \frac{C_{j0s}}{\left[1 - \left(\frac{V_D}{V_{bi}}\right)\right]^m} = \frac{C_{j0b} + C_{j0s}}{\left[1 - \left(\frac{V_D}{V_{bi}}\right)\right]^m} \tag{2.17}$$

Substituting in the numbers, we get

$$C_j = \frac{1\,pF + 0.120\,pF}{\left(1 - \left(\frac{V_D}{0.759}\right)\right)^{0.33}} = \frac{1.120\,pF}{\left(1 - \left(\frac{V_D}{0.759}\right)\right)^{0.33}}$$

A sketch of how this capacitance changes with reverse potential is given in Fig. 2.15. Notice that when we discuss the depletion capacitance of a diode, it is usually with regard to a reverse bias (V_D is negative). When the diode becomes forward-biased minority carriers, electrons in the p material and holes in the n material, injected across the junction, form a *stored* or *diffusion* charge in and around the junction and give rise to a storage or diffusion capacitance. This capacitance is usually much larger than the depletion capacitance. Furthermore, the time it takes to remove this stored charge can be significant. ■

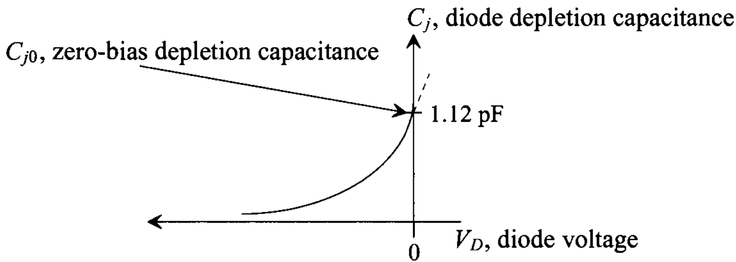


Figure 2.15 Diode depletion capacitance against diode reverse voltage.

2.4.3 Storage or Diffusion Capacitance

Consider the charge distribution of the forward-biased diode shown in Fig. 2.16. When the diode becomes forward biased, electrons from the n-type side of the junction are attracted to the p-type side (and vice versa for the holes). After an electron drifts across the junction, it starts to diffuse toward the metal contact. If the electron recombines, that is, falls into a hole, before it hits the metal contact, the diode is called a *long base diode.* The time it takes an electron to diffuse from the junction to the point where it recombines is called the carrier lifetime, τ_T (see Sec. 2.4.1). For silicon this lifetime is on the order of 10 µs. If the distance between the junction and the metal contact is short, such that the electrons make it to the metal contact before recombining, the diode is said to be a *short base diode.* In either case, the time between crossing the junction and recombining will be labeled τ_T (transit time). A capacitance is formed between the electrons diffusing into the p-side and the holes diffusing into the n-side, that is, formed between the minority carriers. (Electrons are the minority carriers in the p-type semiconductor.) This capacitance is called a **diffusion** capacitance or **storage** capacitance due to the presence of the stored, or diffusing, minority carriers around the forward-biased pn junction.

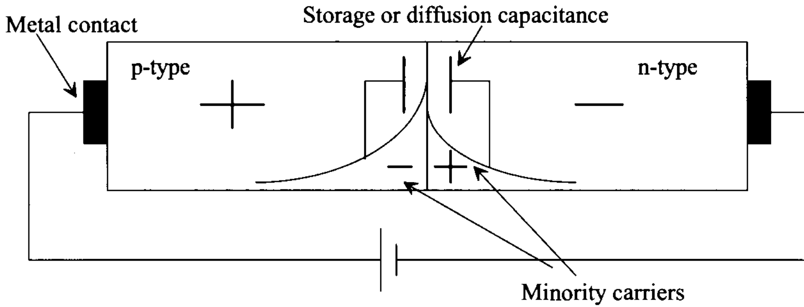


Figure 2.16 Charge distribution in a forward-biased diode.

We can characterize the storage capacitance, C_S, in terms of the minority carrier lifetime. Under DC operating conditions, the storage capacitance is given by

$$C_S = \frac{I_D}{nV_T} \cdot \tau_T \tag{2.18}$$

I_D is the DC current flowing through the forward-biased junction given by Eq. (2.4). Looking at the diode capacitance in this way is very useful for analog AC small-signal analysis. However, for digital applications, we are more interested in the large-signal switching behavior of the diode. It should be pointed out that, in general, for a CMOS process, it is undesirable to have a forward-biased pn junction. If we do have a forward-biased junction, it usually means that there is a problem. For example, electrostatic protection diodes are turning on (Fig. 4.17), capacitive feedthrough is possibly causing latch-up, or such. These topics appear in more detail later in the book.

Consider Fig. 2.17. In the following diode switching analysis, we assume that V_F >> 0.7, V_R < 0 and that the voltage source has been at V_F long enough to reach steady-state condition; that is, the minority carriers have diffused out to an equilibrium condition. At the time t_1, the input voltage source makes an abrupt transition from a forward voltage of V_F to a reverse voltage of V_R, causing the current to change from $\frac{V_F - 0.7}{R}$ to $\frac{V_R - 0.7}{R}$. The diode voltage remains at 0.7 V, because the diode contains a stored charge that must be removed. At time t_2, the stored charge is removed. At this point, the diode basically looks like a voltage-dependent capacitor that follows Eq. (2.14). In other

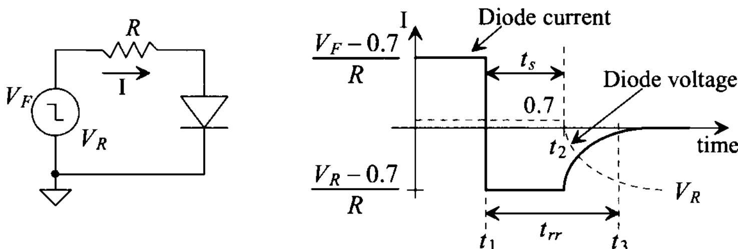


Figure 2.17 Diode reverse recovery test circuit.

words for $t > t_2$, the diode-depletion capacitance is charged through R until the current in the circuit goes to zero and the voltage across the diode is V_R. This accounts for the exponential decay of the current and voltage shown in Fig. 2.17.

The diode storage time, the time it takes to remove the stored charge, t_s, is simply the difference in t_2 and t_1, or

$$t_s = t_2 - t_1 \tag{2.19}$$

This time is also given by

$$t_s = \tau_T \cdot \ln \frac{i_F - i_R}{-i_R} \tag{2.20}$$

where $\frac{V_F - 0.7}{R} = i_F$ and $\frac{V_R - 0.7}{R} = i_R$ = a negative number in this discussion. Note that it is quite easy to determine the minority carrier lifetime using this test setup.

Defining a time t_3, where $t_3 > t_2$, when the current in the diode becomes 10% of $\frac{V_R - 0.7}{R}$, we can define the diode reverse recovery time, or

$$t_{rr} = t_3 - t_1 \tag{2.21}$$

Note that the reverse recovery time (the time it takes to shut off a forward-biased diode) is one of the big reasons that digital circuits made using silicon bipolar transistors don't perform as well, in general, as their CMOS counterparts.

Table 2.1 SPICE parameters related to diode.

Name	SPICE	
I_S	IS	Saturation current
R_S	RS	Series resistance
n	N	Emission coefficient
V_{bd}	BV	Breakdown voltage
I_{bd}	IBV	Current which flows during V_{bd}
C_{j0}	CJ0	Zero-bias pn junction capacitance
V_{bi}	VJ	Built-in potential
m	M	Grading coefficient
τ_T	TT	Carrier transit time

2.4.4 SPICE Modeling

The SPICE diode model parameters are listed in Table 2.1. The series resistance, R_s, results from the finite resistance of the semiconductor used in making the diode and the contact resistance, the resistance resulting from a metal contact to the semiconductor. At this point, we are only concerned with the resistance of the semiconductor. For a reverse-biased diode, the depletion layer width changes, increasing for larger reverse voltages (decreasing both the capacitance and series resistance, of the diode). However, when we model the series resistance, we use a constant value. In other words, SPICE will not show us the effects of a varying R_S.

Example 2.4
Using SPICE, explain what happens when a diode with a carrier lifetime of 10 ns is taken from the forward-biased region to the reverse-biased region. Use the circuit shown in Fig. 2.18 to illustrate your understanding.

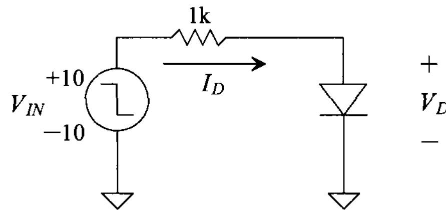


Figure 2.18 Circuit used in Ex. 2.4 to demonstrate simulation of a diode's reverse recovery time.

We assume a zero-bias depletion capacitance of 1 pF. The SPICE netlist used to simulate the circuit in Fig. 2.18 is shown below.

```
*** Figure 2.19 CMOS: Circuit Design, Layout, and Simulation ***

.control
destroy all
run
let id=-i(vin)*1k
plot vd vin id
.endc

D1    vd    0    Dtrr
R1    vin   vd   1k
Vin   vin   0    DC    0    pulse 10 -10 10n .1n .1n 20n 40n

.Model Dtrr D  is=1.0E-15 tt=10E-9 cj0=1E-12 vj=.7 m=0.33
.tran 100p 25n
.end
```

Figure 2.19 shows the current through the diode (I_D), the input voltage step (V_{IN}), and the voltage across the diode (V_D).

What's interesting to notice about this circuit is that current actually flows through the diode in the negative direction, even though the diode is forward biased (has a forward voltage drop of 0.7 V). During this time, the stored minority carrier charge (the diffusion charge) is removed from the junction. The storage time is estimated using Eq. (2.20)

$$t_s = 10ns \cdot \ln \frac{9.3 + 10.7}{10.7} = 6.25\ ns$$

which is close to the simulation results. Note that the input pulse doesn't change until 10 ns after the simulation starts. This ensures a steady-state condition when the input changes from 10 to −10 V. ■

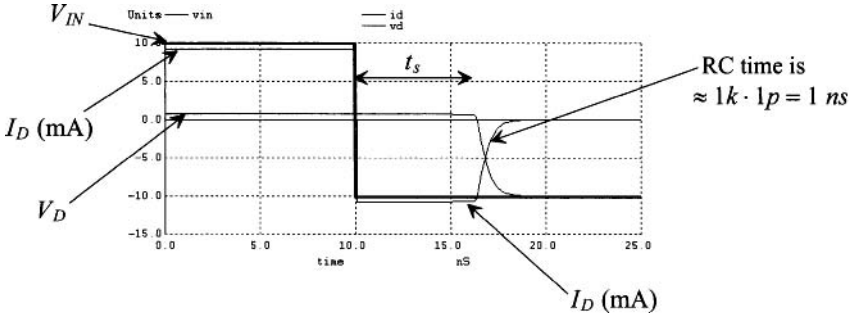


Figure 2.19 The simulation results for Ex. 2.4.

2.5 The RC Delay through an N-well

At this point, we know that the n-well can function as a resistor and as a diode when used with the substrate. Figure 2.20a shows the parasitic capacitance and resistance associated with the n-well. Since there is a depletion capacitance from the n-well to the substrate, we could sketch the equivalent symbol for the n-well resistor, as shown in Fig. 2.20b. This is the basic form of an RC transmission line. If we put a voltage pulse into one side of the n-well resistor, then a finite time later, called the delay time and measured at the 50% points of the pulses, the pulse will appear.

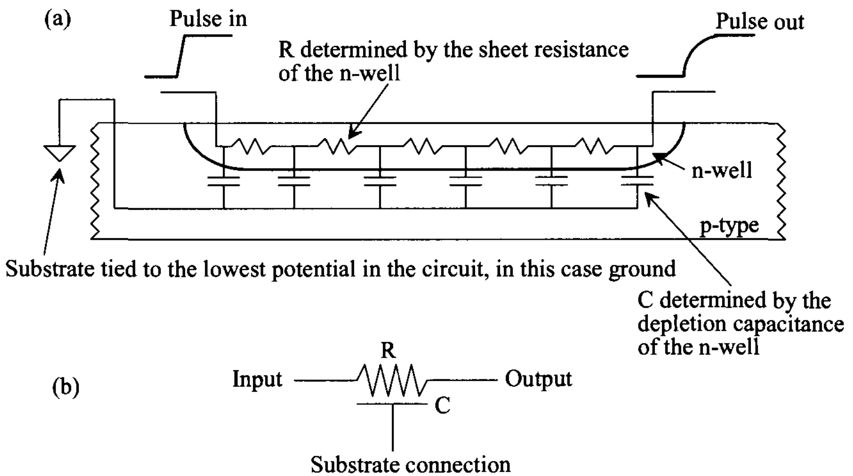


Figure 2.20 (a) Parasitic resistance and capacitance of the n-well and (b) schematic symbol.

RC Circuit Review

Figure 2.21 shows a simple RC circuit driven from a voltage pulse. If the input pulse transitions from 0 to V_{pulse} at a time which we'll call zero, then the voltage across the capacitor (the output voltage) is given by

$$V_{out}(t) = V_{pulse}(1 - e^{-t/RC}) \tag{2.22}$$

The time it takes the output of the RC circuit to reach 50% of V_{pulse} (defined as the circuit's delay time) is determined using

$$\frac{V_{pulse}}{2} = V_{pulse}(1 - e^{-t_d/RC}) \rightarrow t_d \approx 0.7RC \tag{2.23}$$

The rise time of the output pulse is defined as the time it takes the output to go from 10% of the final voltage to 90% of the final voltage (V_{pulse}). To determine the rise time in terms of the RC time constant, we can write

$$0.1V_{pulse} = V_{pulse}(1 - e^{-t_{10\%}/RC}) \tag{2.24}$$

and

$$0.9V_{pulse} = V_{pulse}(1 - e^{-t_{90\%}/RC}) \tag{2.25}$$

Solving these two equations for the rise time gives

$$t_r = t_{90\%} - t_{10\%} \approx 2.2RC \tag{2.26}$$

We will use these results *often* when designing digital circuits. **It's important** that any electrical engineer be able to derive Eqs. (2.23) and (2.26).

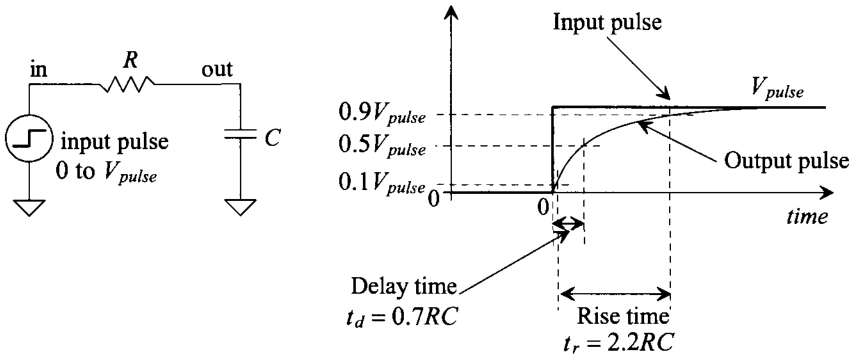


Figure 2.21 Rise and delay times in an RC circuit.

Distributed RC Delay

The n-well resistor seen in Fig. 2.20 is an example of a distributed RC circuit (not a single, RC like the one seen in Fig. 2.21). In order to estimate the delay through a distributed RC, consider the circuit seen in Fig. 2.22. The delay to node A is estimated using

$$t_{dA} = 0.7R_{square}C_{square} \tag{2.27}$$

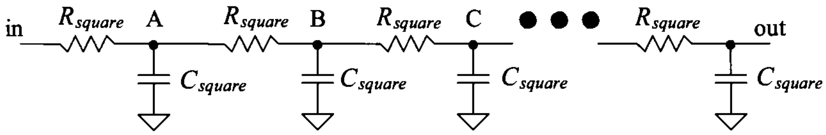


Figure 2.22 Calculating the delay through a distributed RC delay.

The delay to node B is the sum of the delay to point A plus the delay associated with charging the capacitance at node B through $2R_{square}$ or

$$t_{dB} = 0.7(R_{square}C_{square} + 2R_{square}C_{square}) \tag{2.28}$$

Similarly, the delay to node C is

$$t_{dC} = 0.7(R_{square}C_{square} + 2R_{square}C_{square} + 3R_{square}C_{square}) \tag{2.29}$$

For a large number of sections, l, we can write the overall delay through the distributed RC delay as

$$t_d = 0.7R_{square}C_{square} \cdot (1 + 2 + 3 + 4 + ... + l) \tag{2.30}$$

The term in parentheses can be written as

$$(1 + 2 + 3 + 4 + ... + l) = \frac{l(l+1)}{2} \tag{2.31}$$

and so for a large number of sections l

$$t_d \approx 0.35 \cdot R_{square}C_{square} \cdot l^2 \tag{2.32}$$

where the R_{square} and C_{square} are the resistance and capacitance of each square of the distributed RC line.

Example 2.5

Estimate the delay through a 250 kΩ resistor made using an n-well with a width of 10 and a length of 500. Assume that the capacitance of a 10 by 10 square of n-well to substrate is 5 fF. Verify your answer with SPICE.

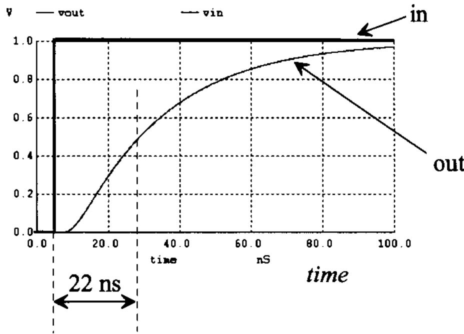
We can divide the n-well up into 50 squares each having a size of 10 by 10 and a resistance of 5 kΩ. The delay through the resistor is then (remembering 1 femto = 10^{-15})

$$t_d = 0.35 \cdot (5k) \cdot (5f) \cdot (50)^2 \approx 22\ ns$$

Note that the total resistance value is $R_{square} \cdot l$ while the total capacitance of the resistor to substrate is $C_{square} \cdot l$. We can use this result to quickly estimate the delay of a distributed RC line (sometimes called an RC transmission line) as

$$t_d = 0.35 \cdot (\text{total resistance}) \cdot (\text{total capacitance}) \tag{2.33}$$

The simulation results are seen in Fig. 2.23. A SPICE lossy transmission line was used to model the distributed effects of the resistor. ■




```
*** Figure 2.23 CMOS: Circuit Design, Layout, and Simulation ***
.control
destroy all
run
plot vin vout
.endc
.tran 100p 100n

O1 Vin 0 Vout 0 TRC
Rload Vout 0 1G
Vin vin 0 DC 0 pulse 0 1 5n 0
.model TRC ltra R=5k C=5f len=50
.end
```

Figure 2.23 SPICE simulations showing the delay through an n-well resistor.

Distributed RC Rise Time

A similar analysis to what was used to arrive at Eq. (2.32) can be used to determine the rise time through a distributed RC line. The result is

$$t_r = 1.1 \cdot R_{square} C_{square} \cdot l^2 \tag{2.34}$$

Using this equation in Ex. 2.5 results in an output rise time of approximately 69 ns. Comparing this estimate to the simulation results in Fig. 2.23, we see good agreement.

2.6 Twin Well Processes

Before going too much further, let's summarize some of the layout discussions presented in this chapter and discuss some concerns. Examine the cross-sectional views seen in Fig. 2.24. We know that the body of an NMOS transistor is p-type, while the body of a PMOS device is n-type. In the n-well process, Fig. 2.24a, the NMOS are fabricated directly in the p-type substrate and the PMOS are made in the n-well. For a p-well process, Fig. 2.24b, the NMOS are made in the p-well while the PMOS are fabricated in the n-type substrate. Note that sometimes the term **tub** is used in place of well (e.g., an n-tub process) because the resulting semiconductor area has a cross-sectional view like a bathtub's.

When implanting the n-well, in Fig. 2.24a for example, the substrate must be *counter-doped.* This means that the p-substrate must have n-type dopants (such as phosphorous or arsenic) added until its concentration changes from p- to n-type. The problem with counter-doping the p-substrate to make an n-well is that the quality of the resulting semiconductor isn't as good as it would be by simply taking intrinsic silicon and adding donor atoms. The acceptor atoms in the p-substrate become ionized (e.g., electrons fall into the holes) increasing scattering and reducing mobility. Sometimes the effects of counter doping are called *excessive doping effects*. In an n-well process, the PMOS devices suffer from excessive doping and so the quality of the device isn't as good as the quality of a PMOS device in a p-well process. For example, a PMOS device fabricated in an n-well is slower than a PMOS device fabricated in an n-substrate.

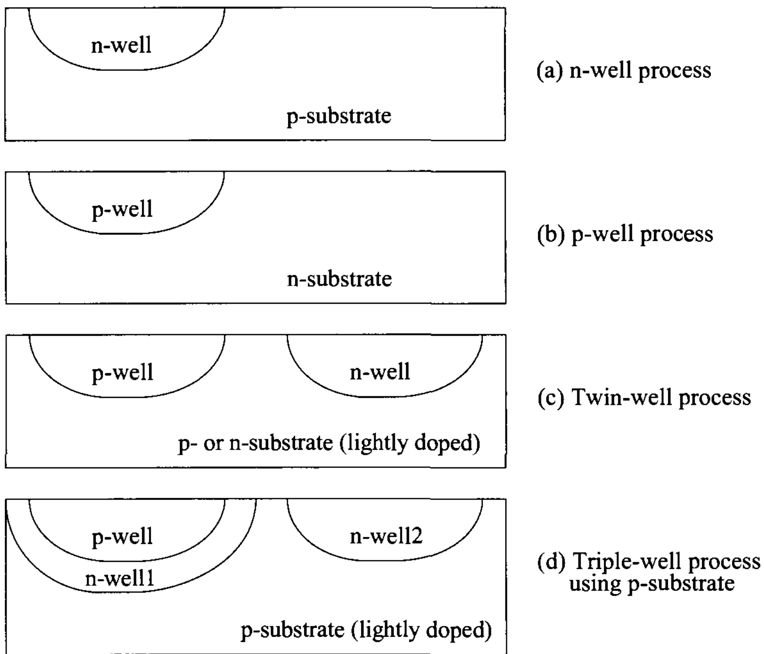


Figure 2.24 The different possible wells used in a bulk CMOS process.

In an attempt to reduce excessive doping effects, a twin-well process, Fig. 2.24c, can be used. When using a lighter doped substrate, the amount of counter doping isn't as significant. We don't use an intrinsic silicon substrate because it is difficult to control the doping at such low levels. If a p-substrate is used, then the p-well is electrically connected to the substrate. The bodies of the NMOS are then all tied to the same potential, usually ground. To allow the bodies of the MOSFETs to be at different potentials a triple-well process can be used, Fig. 2.24d. The added n-well isolates the p-well from the substrate (the n-well1 and p-substrate form a diode that electrically isolates the p-well from the substrate). The p-well can exist directly in the substrate too.

Design Rules for the Well

Figure 2.25 shows the design rules, from MOSIS, for the well. Notice that there are four different sets of rules that the layout designer can follow. In this book **we will use the CMOSedu rules** to illustrate design examples. Before saying why, let's provide a little history and background. We know from Ch. 1 that MOSIS collects chip designs from various sources (education, private, not-for-profit, etc). These designs are put together to form the masks used for making the chips (on multiproject wafers). The actual vendors used by MOSIS to fabricate chips has changed throughout the years. To make the layouts transferable as well as scalable between different CMOS processes, MOSIS came up with the so-called SCMOS rules (scalable CMOS design rules). A parameter, λ, is used in the rules. All of the layouts are drawn on a λ grid. When making the GDS file (or CIF file), the layout is scaled by this factor. For example, if an n-well box is drawn 10λ by 10λ with

Rule	Description	Lambda			Scale
		SCMOS	SUBM	DEEP	CMOSedu
1.1	Minimum width	10	12	12	6
1.2	Minimum spacing between wells at different potential	9	18	18	9
1.3	Minimum spacing between wells at same potential	6	6	6	3
1.4	Minimum spacing between wells of different type (if both are drawn)	0	0	0	0

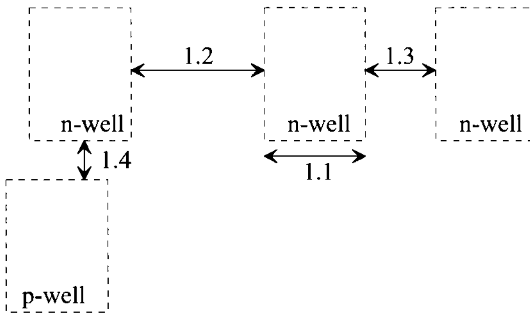


Figure 2.25 Layouts showing the MOSIS design rules for the n-well.

a lambda of 0.3 μm, then when the GDS file is exported (called streaming the layout out) the actual size of the layout is 3 μm square. If the layout were used in a different process, only the value of λ would need to be changed. In other words, the exact same layout can be used in a different technology. Being able to use the same layout and simply *scale it* is a significant benefit of CMOS.

When the SCMOS rules were first introduced, the minimum size of a dimension in CMOS was approximately I μm. The "fabs" or vendors (the factory where the wafers are actually processed) also have a set of design rules. In general, the fab's design rules are tighter than the SCMOS rules. For example, one fab may specify a minimum n-well width of 3 μm, while another fab, in the same process technology, may specify 4 μm. The SCMOS rules may specify 5 μm to cover all possible situations (the price of using the SCMOS rules over the vendor's rules is larger layout areas). Unfortunately, as the process dimensions have shrunk over time, the MOSIS SCMOS rules weren't relaxed enough, making modifications necessary. This has led to the MOSIS submicron rules (SUBM) and the MOSIS deep-submicron (DEEP) rules seen in Fig. 2.25 (there are three sets of MOSIS scalable design rules, SCMOS, DEEP, and SUBM). Older processes still use the SCMOS rules, while the smaller technologies use the modified rules. Note that if a layout passes the DEEP rules, it will also pass the SCMOS rules (except for the exact via size).

Why use the CMOSedu rules in this book? Why not use one of the three sets of design rules from MOSIS? The answer comes from how we lay out MOSFETs. The **minimum length of a MOSFET using the MOSIS rules is 2** (2λ, keeping in mind that some scale factor is used when generating the GDS or CIF file). In the CMOSedu rules

we took the MOSIS DEEP rules and divided by two. This means that layouts in the CMOSedu rules are *exactly the same* as the MOSIS DEEP rules except that they are scaled by a factor of 2. **Using the CMOSedu rules, the minimum length of a MOSFET is 1**. If MOSIS specifices a scale factor, λ, of 90 nm using the DEEP rules, where the minimum length is 2, then we would use a scale factor of 180 nm when using the CMOSedu rules with a minimum length of 1. In SPICE we use ".options scale=90n" when using the DEEP rules and ".options scale=180n" when using the CMOSedu rules.

SEM Views of Wells

Before leaving this chapter, let's show a scanning electron microscope (SEM) image of a well. In an SEM electrons are emitted from a cathode made with either tungsten or lanthanum hexaboride (LaB6). Tungsten is generally used for the cathode because it has a high melting point and a low vapor pressure. In some SEMs electrons are emitted via field emission. In either case an electron beam is formed and moved across the surface of an object. To move the electron beam it is passed through pairs of scanning coils and an objective lens. Varying the current through the coils deflects the beam, moving it across the surface of an object, and is used to form the image. The electrons are then attracted towards an anode and collected (as a varying output current).

Figure 2.26 shows an SEM image of a cross-sectional view of a portion of a CMOS memory chip. While most of the fabricated layers seen in this photograph will be covered in the next few chapters, we do point out, in the figure, a p-well (hard to see), a deep n-well, and the p-substrate. Note that in order to view different materials in a cross-section the sample is first stained prior to placement in the SEM and imaging.

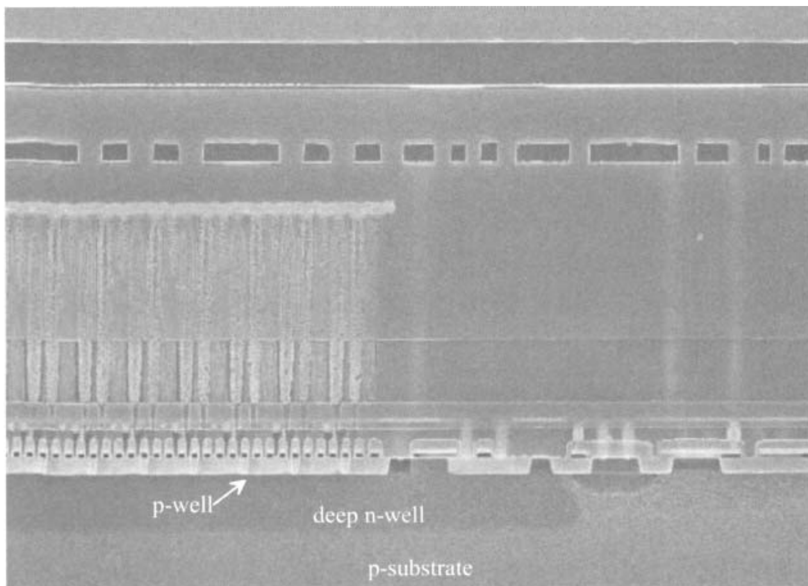


Figure 2.26 SEM image showing the cross-section of a CMOS memory chip.

ADDITIONAL READING

[1] Y. Taur and T. H. Ning, *Fundamentals of Modern VLSI Devices*, Second Edition, Cambridge University Press, 2010. ISBN 978-0521832946

[2] S. A. Campbell, *Fabrication Engineering at the Micro- and Nanoscale*, 3rd ed, Oxford University Press, 2008. ISBN 978-0195320176

[3] M. J. Madou, *Fundamentals of Microfabrication: The Science of Miniaturization*, 2nd ed., CRC Publisher, 2002. ISBN 978-0849308260

[4] R. C. Jaeger, *Introduction to Microelectronic Fabrication*, 2nd ed, volume 5 of the Modular Series on Solid State Devices, Prentice-Hall Publishers, 2002. ISBN 0-20-144494-1

[5] J. D. Plummer, M. D. Deal, and P. B. Griffin, *Silicon VLSI Technology, Fundamentals, Practice, and Modeling*, Prentice-Hall Publishers, 2000. ISBN 978-0130850379

PROBLEMS

2.1 For the layout seen in Fig. 2.27, sketch the cross-sectional views at the places indicated. Is there a parasitic pn junction in the layout? If so, where? Is there a parasitic bipolar transistor? If so, where?

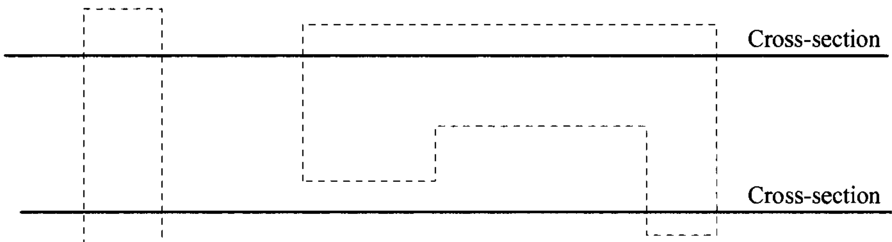


Figure 2.27 Layout used in problem 2.1.

2.2 Sketch (or use a layout tool) the layout of an n-well box that measures 100 by 10. If the scale factor is 50 nm, what is the actual size of the box after fabrication? What is the area before and after scaling? Neglect lateral diffusion or any other fabrication imperfections.

2.3 Lay out a nominally 250 kΩ resistor using the n-well in a serpentine pattern similar to what's seen in Fig. 2.28. Assume that the maximum length of a segment is 100 and the sheet resistance is 2 kΩ/square. Design rule check the finished resistor. If the scale factor in the layout is 50 nm, estimate the fabricated size of the resistor.

2.4 If the fabricated n-well depth, t, is 1 μm, then what are the minimum, typical, and maximum values of the n-well resistivity, ρ? Assume that the measured sheet resistances (minimum, typical, and maximum) are 1.6, 2.0, and 2.2 kΩ/square?

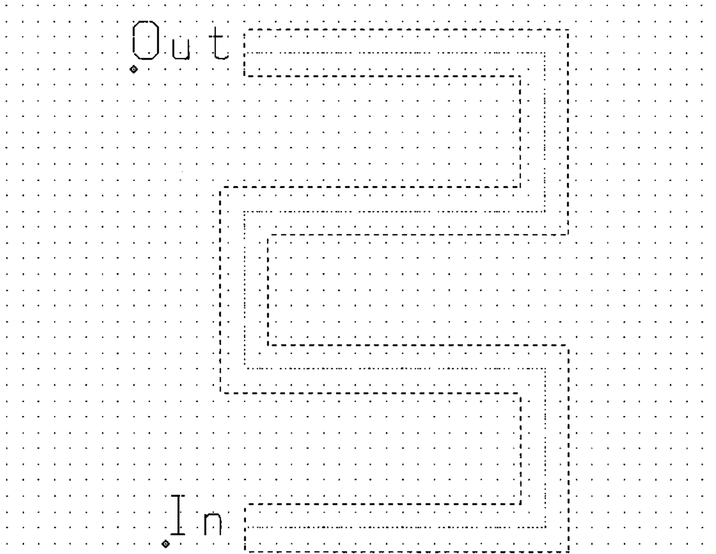


Figure 2.28 Layout of an n-well resistor using a serpentine pattern.

2.5 Normally, the scale current of a pn junction is specified in terms of a scale current density, J_s (A/m^2), and the width and length of a junction (i.e., $I_s = J_s \cdot L \cdot W \cdot scale^2$ neglecting the sidewall component). Estimate the scale current for the diode of Ex. 2.3 if $J_s = 10^{-8}$ A/m^2.

2.6 Repeat problem 2.5, including the sidewall component ($I_s = J_s \cdot L \cdot W \cdot scale^2 + J_s \cdot (2L+2W) \cdot scale \cdot$depth).

2.7 Using the diode of Ex. 2.3 in the circuit of Fig. 2.29, estimate the frequency of the input signal when the AC component of v_{out} is 707 μV (i.e., estimate the 3 dB frequency of the $|v_{out}/v_{in}|$).

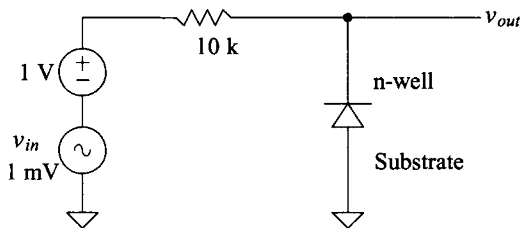


Figure 2.29 Treating the diode as a capacitor. See Problem 2.7.

2.8 Verify the answer given in problem 2.7 with SPICE.

2.9 Using SPICE, show that a diode can conduct significant current from its cathode to its anode when the diode is forward biased.

2.10 Estimate the delay through a 1 MΩ resistor (10 by 2,000) using the values given in Ex. 2.5. Verify the estimate with SPICE.

2.11 If one end of the resistor in problem 2.10 is tied to +1 V and the other end is tied to the substrate that is tied to ground, estimate the depletion capacitance (F/m^2) between the n-well and the substrate at the beginning, the middle, and the end of the resistor. Assume that the resistance does not vary with position along the resistor and that the scale factor is 50 nm, C_{j0} = 25 aF for a 10 by 10 square, m = 0.5, and V_{bi} = 1.

2.12 The diode reverse breakdown current, that is, the current that flows when $|V_D| < BV$ (breakdown voltage), is modeled in SPICE by

$$I_D = IBV \cdot e^{-(V_D + BV)/V_T}$$

Assuming that 10 μA of current flows when the junction starts to break down at 10 V, simulate, using a SPICE DC sweep, the reverse breakdown characteristics of the diode. (The breakdown voltage, BV, is a positive number. When the diode starts to break down $-BV = V_D$. For this diode, breakdown occurs when $V_D = -10$ V.)

2.13 Repeat Ex. 2.3 if the n-well/p-substrate diode is 50 square and the acceptor doping concentration is changed to 10^{15} atoms/cm^3.

2.14 Estimate the storage time, that is, the time it takes to remove the stored charge in a diode, when τ_T = 5 ns, V_F = 5 V, V_R = – 5 V, C_{j0} = 0.5 pF, and R = 1k. Verify the estimate using SPICE.

2.15 Repeat problem 2.14 if the resistor is increased to 10k. Comment on the difference in storage time between using a 1k and a 10k resistor. What dominates the increase the diode's reverse recovery time when using a 10k resistor instead of a 1k resistor?