CHAPTER
3

# Fabrication, Layout, and Simulation

**CHAPTER OUTLINE**



## 3.1 Introduction

In this chapter we describe the close relationship between fabrication, layout, and simulation of integrated circuits. We begin by describing the CMOS fabrication process. The purpose is to give the reader some insight into the steps and issues associated with the manufacture of integrated circuits. Over the years, IC fabrication has become more and more complex. For each technology node, there are multiple processes available to the designer, depending on the target application of the chip. For example, there may be different processes for digital, memory, mixed-signal, low power, and so on. It is not possible to describe all of the pertinent issues in advanced CMOS processes in one chapter. The objective here is to present a generic process to provide an overview of the key steps of the fabrication process.

With this knowledge, the issues surrounding IC layout can be described. A layout is the physical description of the transistors and their connections in the IC design. The layout precisely defines the chip that will ultimately be fabricated. Every transistor and wire in the circuit must be specified in a geometric format in the layout. To ensure that the chip is properly fabricated, the final layout must conform to

a set of design rules that are based on the resolution limits of the manufacturing equipment. The section on IC layout describes transistor layout considerations and some of the basic guidelines for digital CMOS layout, including design rules and their meaning. Full-chip layout issues are not addressed here.

Following the overview of layout, the simulation tool SPICE[1] is described along with the device parameters associated with the LEVEL 1 and BSIM3v3 MOS models. SPICE is the most widely used simulation tool in the industry for detailed circuit analysis. When a circuit is described for SPICE simulation, each MOS transistor requires two pieces of information: geometric and parametric. The geometric information is obtained directly from the layout. This includes the length, width, area, and perimeter information for each transistor. The parametric information is captured in the device models that are extracted from the fabrication process. The simplest device model is the LEVEL 1 model which is described in detail in this chapter. However, the most popular model in use today is BSIM3v3. The key parameters of this model are also described in this chapter. Advanced MOS transistor issues are presented in the context of the BSIM3 model.

The chapter concludes with a brief look at an emerging technology known as silicon-on-insulator (SOI).

For those who are relatively new to silicon integrated circuit technology and device modeling, the contents of this chapter may not be fully appreciated or understood until later chapters are covered. We recommend that this chapter and Chapter 2 be reviewed frequently as they contain the fundamental equations and concepts used throughout the rest of the book.

## 3.2 IC Fabrication Technology

### 3.2.1 Overview of IC Fabrication Process

Silicon transistors and integrated circuits are manufactured on wafers of single-crystal silicon, 200 to 300 millimeters (mm) in diameter and about 0.35 mm to 1.25 mm thick. This thickness is determined by the need to provide enough mechanical strength so that the wafer is not easily broken. The wafers are first polished to a mirror finish by abrasive lapping with finer and finer gritty material, followed by a chemical etch that leaves the surface virtually free of scratches and imperfections. Any defects on the wafer may cause a particular chip to fail and render it useless. This is very important in terms of the overall yield of the wafer. The chips that do not work to specifications are thrown out or must be repaired in some way, and this increases the per unit cost of the chips that actually work.

The fabrication process that forms the devices and circuits involves a sequence of pattern definition steps interspersed with other processes such as oxidation, etching, diffusion or ion implantation (*doping*), and material deposition. A variety of chemical agents and materials are used in this process. There may be 20 to 30 major steps in the fabrication process, each one requiring five or more operations. Approximately 100 to 200 distinct operations are required to produce complete

[1] For those readers not familiar with this simulation tool, a short tutorial on running SPICE is provided in Appendix A.

integrated circuits. After the processing is complete, each wafer is sawed into hundreds or thousands of identical rectangular chips. Today, a complex IC chip may be up to 30 mm on each edge and contain 100 million devices (transistors, resistors, diodes, etc.) or more.

Brief descriptions of some of the nomenclature in processing are as follows:

Oxidation: High-temperature exposure of silicon to oxygen to form $SiO_2$ (silicon dioxide).

Etching: Removal of undesired material from the surface with the use of a liquid or ionized gas etchant.

Diffusion: Doping process to form *n*-type or *p*-type material by high-temperature exposure to donor or acceptor impurities.

Ion Implantation: High-energy bombardment of silicon with donor or acceptor ions from particle accelerators followed by an annealing step to activate implants and repair any damage.

Chemical Vapor Deposition (CVD): Materials such as metal or oxide are deposited out of a gaseous mixture. Metals can also be deposited using sputtering.

The basic flow in the IC fabrication process is shown in Figure 3.1. Designers convert the circuit schematic into a layout consisting of geometric patterns that implement the transistors and interconnections. This geometric information is divided into *layers* such as $n^+$, $p^+$, *p*-well, *n*-well, poly, metal 1, metal 2, etc. Each

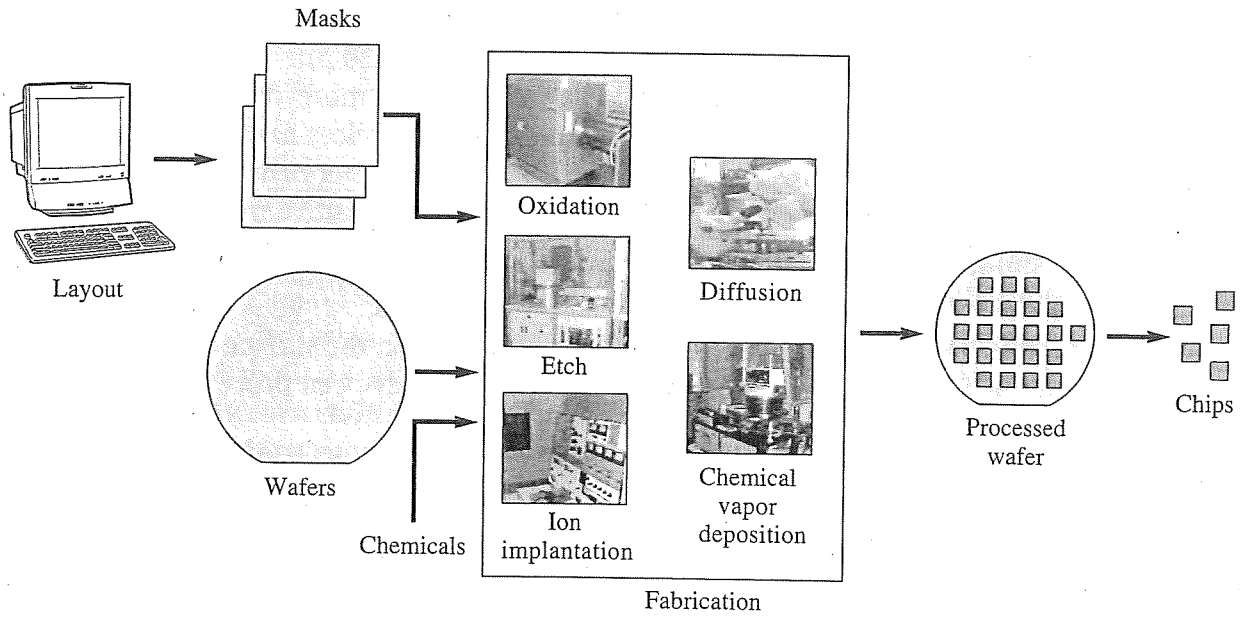



**Figure 3.1**
Overview of chip-making process.

layer corresponds to a major step in the fabrication process. The patterns, which convey precisely what designers want to build, are converted to a set of masks, one for each major step in the fabrication process. The masks are glass plates with patterns that specify the information that will be printed on the integrated circuit in a given step. They have transparent and opaque areas that match the geometric layout data for each separate layer.

The images on the masks are transferred to silicon wafers using an *optical lithography* process to simultaneously produce a large number of chips. The masks are aligned and exposed separately at each die site (or clusters of 2–8 die sites) across the wafer. This is similar to the use of a negative in photography. Many pictures can be produced from one negative. In chip fabrication, a 200 mm wafer may produce 1000 chips that are 5 mm × 5 mm in size. It is this parallel manufacturing of all chips on a wafer that keeps the cost of an integrated circuit relatively low.[2]

### 3.2.2 IC Photolithographic Process

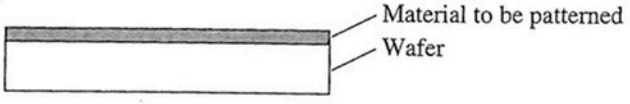
The process of pattern transfer and pattern definition is repeated many times during the fabrication of an IC wafer. Each of these *masking steps* requires that the wafer be coated with a photosensitive emulsion known as *photoresist,* and then optically exposed in desired geometric patterns using a mask: a previously prepared photographic plate. This glass plate with an image of the pattern etched in chrome is generated from the design database. The image is optically projected onto the wafer using a projection aligner which is very much like an enlarger in photography. It projects the image of the mask onto the photoresist on the silicon wafer to *develop* it, that is, make it soluble.

After developing the photoresist, a specific process such as etching or doping is carried out in the exposed areas of the patterned wafer. This entire pattern transfer process is known as *photolithography,* or *optical lithography.* Steady improvements in optical lithography have made it possible to reduce the smallest surface dimensions on an IC chip from about 5 $\mu$m in 1980 to 130 nm and smaller today.[3] The cost per logic gate or memory cell is reduced as more devices and circuits are formed per unit chip area. Furthermore, smaller devices have smaller capacitances and hence can switch faster, leading to better circuit performance.

Photolithography uses light to develop photosensitive material and create the features required in an integrated circuit. The main steps of the processing sequence for patterning are as follows:

1. Place the desired material to be patterned on the surface of the silicon (or on top of the previous material that has already been processed):

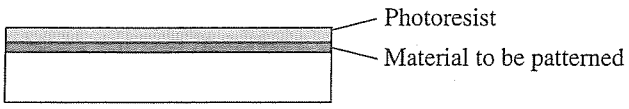
**Step 1**: Apply material to water




[2] However, the masks themselves are becoming increasingly more expensive and could easily run >$1M US in advanced deep submicron technologies.

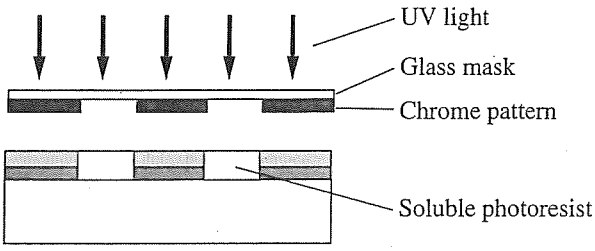
[3] As a reference point, a hair from your head is about 50 $\mu$m in diameter!

2. Spin photoresistive material on top of the first material. Positive photoresist is a material that is difficult to remove unless it is *exposed* to ultraviolet (UV) light, and negative photoresist is difficult to remove unless it is *not exposed* to ultraviolet light. Typically positive photoresist is in use today:



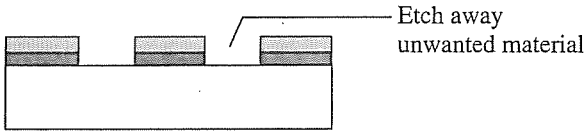

**Step 2**: *Spin* on a photoresistive material

3. Place a glass mask over the wafer and expose the resist material to UV light (positive photoresist assumed). The exposed areas are now soluble in a chemical agent:



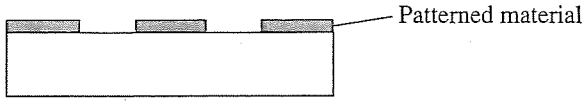

**Step 3**: Pattern photoresist with UV light through glass mask

4. Using the appropriate solvent, remove the resist and expose areas of the underlying material to be patterned. Etch away the underlying material to form the desired pattern:




**Step 4**: Apply specific processing step such as etch, implant, oxidation, after removing soluble photoresist

5. After the previous step is completed, remove the rest of the photoresist:




**Step 5**: Wash off resist

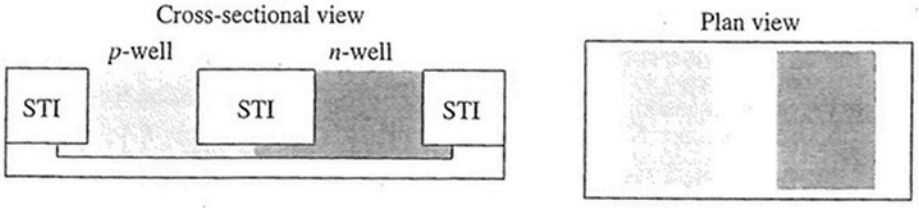
The steps just described are used repeatedly for every major processing step whether it is oxidation, ion implantation, or deposition. However, due to the fine resolutions needed today, photolithography is reaching certain physical limits. Minimum feature sizes less than 0.35 $\mu$m are feasible from the standpoint of device operation but cannot be achieved with standard optical lithography because the wavelength of light is about 0.4 $\mu$m. To overcome this limitation of optical lithography, optical proximity correction (OPC) and phase-shift masks (PSMs) are in common use. These operations either pre-correct the masks to provide the desired final image on the chip, or use masks that phase shift the light to avoid unwanted interference patterns. Such advanced techniques have been used at 130 nm and below.

### 3.2.3 Making Transistors

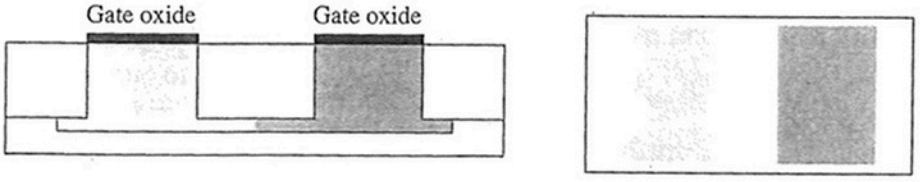
Now that the basic photolithographic step has been illustrated, the manner in which transistors are made can be described. Deep submicron transistors are fabricated

using a number of complex steps. Since our objective here is to highlight the main steps in the process, we will only present an overview of transistor fabrication. The major steps of a CMOS process are as follows.

1. Define well areas and transistor regions. The first set of photolithographic steps is used to specify the areas in which the transistors will be fabricated. NMOS devices are diffused in a *p*-type well. PMOS devices are diffused into an *n*-type well. In recent technologies, twin tubs are used, meaning that an *n*-well and a *p*-well are separately diffused into a common substrate. The transistor areas defined within the well areas are separated from one another using shallow trench isolation (STI). The "trenches" are dug out of the silicon in regions between transistors, and oxide is deposited in these trenches using a chemical vapor deposition process (CVD). The density of transistors today requires the exclusive use of STI.

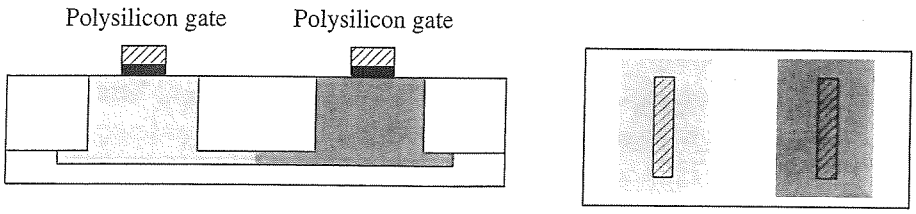



2. Define gate region. The second set of lithographic steps defines the desired patterns for gate electrodes. In a 0.13 $\mu$m process, a clean thermal oxide about 22 Å thick is grown in the transistor areas by exposure to oxygen in a furnace. This is the *gate oxide*. This thin oxide is the insulator in the MOS structure. Threshold adjustment implants are applied to the two gate regions to achieve the desired $V_T$ values for PMOS and NMOS devices, respectively.

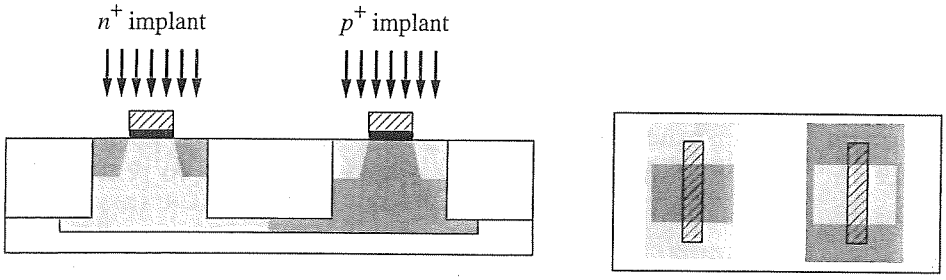



3. Define poly gate. Another CVD process deposits a layer of polycrystalline silicon (*poly*) over the entire wafer. Undesired poly and the underlying thin oxide are removed by chemical or plasma (reactive gas) etching thus producing a self-aligned[4] gate node. The term "self-aligned" refers to the fact that the source and drain regions will automatically align with the poly gate if the gate is placed down first.

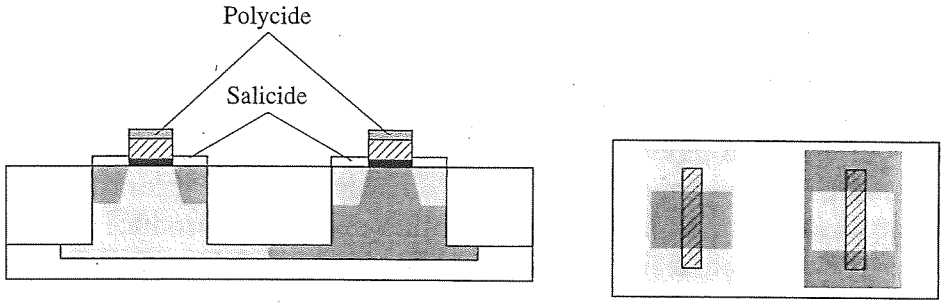
[4] When metal gates were first used, it was difficult to align the gate over the channel region precisely. The use of polysilicon as the gate material eliminated this problem since it would be placed down first and the source and drain regions created afterwards. The chip yields increased dramatically and led to the exclusive use of this approach in MOS technology.




4. Form source/drain regions. A $p^+$ dopant (boron) is introduced into the $n$-well to form the $p$-channel transistor source and drain. Ion implantation is used for each doping step. Then an $n^+$ dopant (phosphorus or arsenic) is introduced into the $p$-well that will become the $n$-channel transistor source and drain. The poly gate requires a doping level that is $n^+$ for NMOS devices and $p^+$ for PMOS devices. This can be performed at the same time as the respective source/drain implants. During a subsequent annealing step, the final junction depth and undesired lateral diffusion under the gate are established.




5. Deposit silicide material. The source, drain, and gate materials have relatively high resistance that may slow down the operation of the transistor. To reduce the resistance, a silicide material is deposited onto the source, drain, and poly regions. The masking step also defines the areas in which contacts to the transistors are to be made. Chemical or plasma etching selectively exposes bare silicon or poly in the contact areas.




Now that the basic processing sequence for transistors has been described, we provide a few more details. Relative to early generations of MOS fabrication, the major changes are shallow trench isolation (STI), extensive channel region engineering, and the use of silicide materials to lower resistance. In Figure 3.2, the final

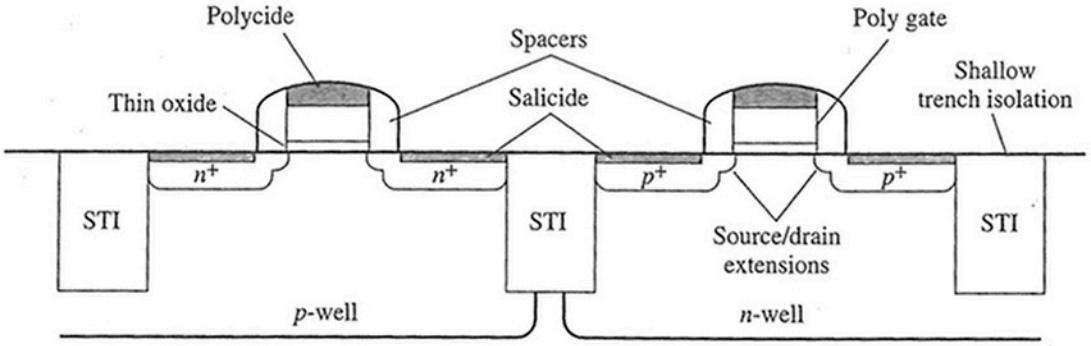



**Figure 3.2**
Deep submicron CMOS transistor structure.

structure of the NMOS and PMOS devices is shown. The *n*-channel and *p*-channel devices are placed in separate wells that sit in heavily doped substrates to avoid *latch-up*, as described later. The STI region is composed of oxide material that is deposited in shallow trenches etched into the silicon. The channel region has a variety of implants to adjust the threshold voltage and to reduce short-channel effects. The source and drain regions feature lightly doped extensions to reduce the possibility of junction breakdown and hot-carrier effects.[5] For the NMOS device, initially phosphorus is implanted and self-aligned to the poly gate edge to form the lightly doped region. Then, oxide is grown and etched to form the "spacers." An arsenic implant follows to create the heavily doped regions. A similar process is used for the PMOS device with boron.

In Figure 3.2, the self-aligned polysilicon gates and their thin oxide are directly above the channel with spacers on either side. Today the dielectric constant, or *k*, of a silicon dioxide gate, is roughly 4. As technology scales, the oxide is so thin that current tunneling through the gate node is likely to occur. To avoid this, thicker materials with higher *k* values are being pursued, the so-called *high-k* gate dielectrics. Target values of *k* for the gate material are expected to be in the vicinity of 10–12 once the process quality issues have been resolved.

To reduce the resistance of the gate and source/drain regions, silicide materials such as $TiSi_2$, $WSi_2$, $PtSi_2$, $CoSi_2$, or TaSi may be used. The silicide material can be seen in Figure 3.2 on top of the polysilicon and the source/drain regions. For example, if titanium is deposited on exposed source/drain and gate regions, it reacts with the silicon surface (during a subsequent heating process) to produce the $TiSi_2$ silicide, while the poly gate reacts with it to produce *polycide*.[6] The application of silicides for both poly and diffusion are carried out in a self-aligned process using

[5] Hot carriers are electrons or holes created through impact ionization with enough energy to surmount the energy barrier between silicon and silicon-dioxide. This topic is discussed in more detail in a later section in this chapter.

[6] When a silicide is used on top of a polysilicon gate, it is referred to as *polycide*.

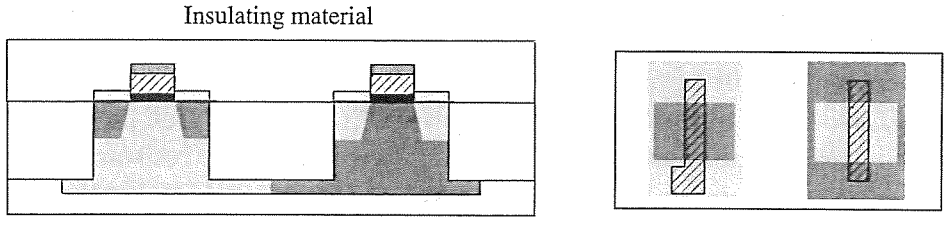
spacers that are the materials shown on either side of the poly gate. Self-aligned silicides are often called *salicides*. One thing to note in the diagram is that the poly gates are tall and thin; they may be 2000 Å high and 1000 Å wide. This gives rise to a fringing capacitance from the side of the poly gate to the surface of the silicon source/ drain region. This contributes to the overlap capacitance which is a combination of fringing capacitance and lateral diffusion underneath the gate.

### 3.2.4 Making Wires

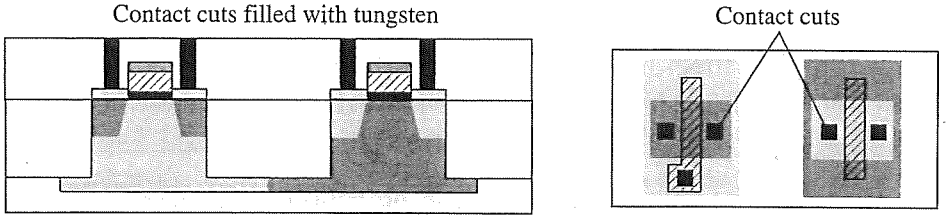
Now that the transistor fabrication has been described, the next step in the process is to make the wires that connect the transistors. These connections, primarily done using a metal such as Al or Cu, are commonly known as *interconnect*. In the early generations of MOS technology, only one or two metal layers were available to wire up the devices. Since there were only a few thousands devices to connect, the wiring process was rather straightforward. With technology scaling, the transistor density has increased tremendously following Moore's Law. As the routing capacity of each layer was exhausted, additional levels of interconnect were required to complete the routing. The number of layers of interconnect has grown over the past 25 years from one layer to over eight layers.

The fabrication of interconnect begins with the first metal layer that is used to make contact with transistor source, drain, and gate terminals, and to connect them to nearby $V_{DD}$, Gnd, and inputs/outputs of other transistors. The starting point for wire fabrication is the structure after transistor fabrication.

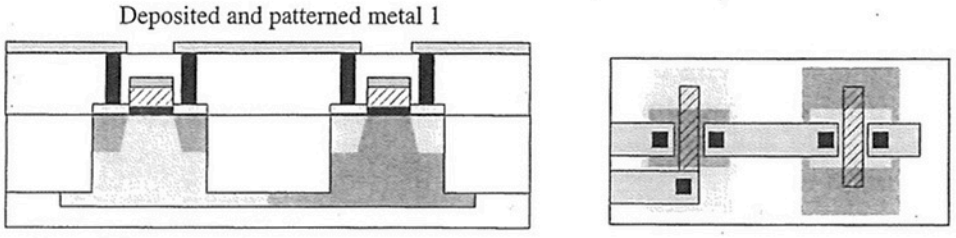
1. Initially, a layer of insulating material is applied and then polished to a flat surface.



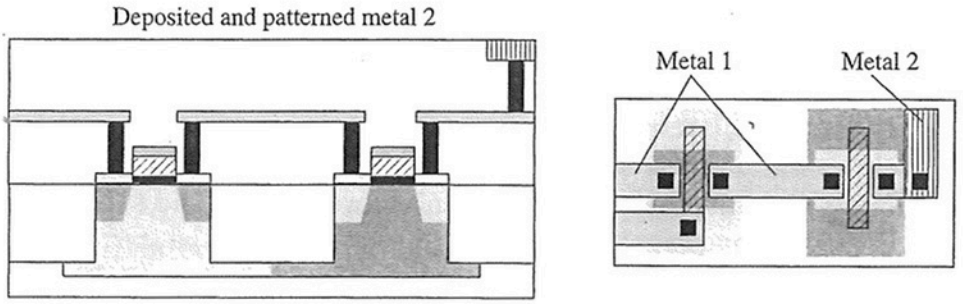

2. Next, contact holes are etched into the insulator and filled with a conducting material such as tungsten.



3. The metal material (either aluminum or copper) is then applied to the surface and patterned to form the desired wires.




4. Next, another passivation layer is applied and polished to a flat surface. Another set of holes are cut into this layer for *vias*. This is the name used to describe material that forms connections between adjacent layers. Then another layer of metal is deposited and patterned. The final structure, including the second layer of metal, is shown below.




Similar steps are used to create the remaining metal layers that form a multilevel metal structure.

Figure 3.3 shows a representative interconnect structure with six layers of metal in a cross-sectional view. Notice that the dimensions of each layer are different since each layer has a specific purpose. For example, the upper layers of metal are used for global signals, clock, and power distribution, and must carry large amounts of current. Their cross sections are made relatively large to keep the resistance levels low. The lower levels are intended for block-level and cell-level routing and are kept small for high density.

The different levels are connected to each other using contacts or vias. Generally speaking, contacts are used to connect wires to transistors, while vias are used to connect one metal layer to another. In the past, aluminum (Al) was used for the metal layers and tungsten (W) was used to implement vias. However, due to increases in resistance and electromigration[7] problems, other materials have been pursued to replace Al. The properties of copper (Cu) have been known to be superior to Al for a long time, but its incompatibility with Si and $SiO_2$ limited its use until a solution to this problem could be developed. Unfortunately, copper diffuses

[7] Electromigration is described in later chapters as the movement of metal material over time due to current flow. Eventually, the metal may break and cause an open circuit (or a short circuit to a neighboring line) that leads to failure of the circuit.

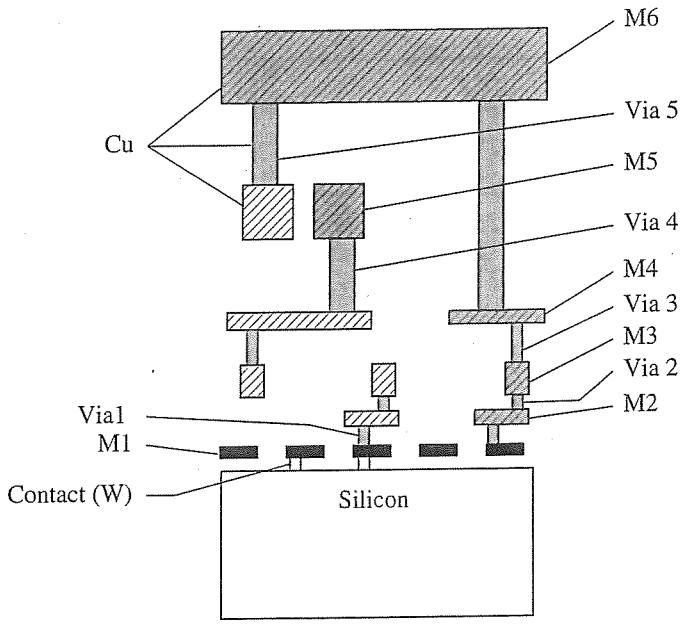


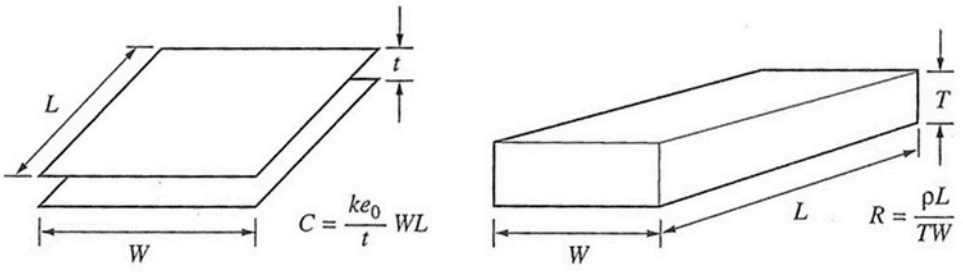

**Figure 3.3**
Multilevel metal interconnect.

very rapidly in silicon, and so great care must be taken to prevent contamination. Typically, a thin copper-cladding material such as TiN is used to surround it to prevent Cu from diffusing into $SiO_2$. This innovation has propelled copper into mainstream usage. Recently, Cu has been used for both wires and vias using a *dual Damascene* process, as depicted in Figure 3.3. Contacts to the poly gate and the source/drain regions are performed using tungsten since it has better adhesion properties than Cu.

Note that vias have resistance associated with them and should be kept as short as possible. This limits the vertical height between metal layers. Often, an array of vias is used to reduce the resistance when connecting between the upper layers of interconnect. Also, to connect from metal 2 to metal 5 requires a sequence of metal 2, via, metal 3, via, metal 4, via, and finally metal 5. Since a direct connect is not possible, the vias are often stacked on top of one another with metal in between. This is referred to as a *stacked via*. Rules exist on the number of vias that can be stacked.

One problem in fabricating deep submicron interconnect is that, as layers are placed on one another, the surface becomes uneven and this may create stresses and strains on the materials in each subsequent layer. Before a new layer of metal is placed on the chip, the surface must be planarized to assist in the photolithographic process of subwavelength geometries, and to avoid problems in the upper layers due to the unevenness of the previous layers. For this purpose, a procedure called chemical mechanical polishing (CMP) is employed whereby a chemical agent and a polishing mechanism are used to remove unwanted material until the surface is highly planar.

Chemical mechanical polishing is used in the front-end process for the planarization of the oxide material in shallow trench isolation, and in the back-end, for




(a) Capacitance of two parallel plates

(b) Resistance of a conductor

**Figure 3.4**
Capacitance and resistance of interconnect.

dielectric planarization and metal etch back. Unfortunately, the degree of planarity is related to the materials below the polishing surface. If there are high- and low-density areas of metal, the results in the two areas will not be uniform as *dishing* may occur. Dishing refers to a sagging of the material in certain areas during the CMP process. To enhance the performance of CMP, some amount of *metal fill* (or *poly fill* for the polysilicon layer) must be specified by the chip designer in the vacant areas to produce more planar surfaces. They are implemented as parallel metal lines during chip layout but are not part of the actual circuit. The natural consequences of introducing metal fill are to increase coupling capacitances of nearby signal lines on the same layer, and on adjacent layers above and below the fills; so it must be used carefully.

### 3.2.5 Wire Capacitance and Resistance

The two most important characteristics of the interconnect today are the capacitance and resistance. We first examine the capacitance of the wires in the interconnect structure. Consider a parallel plate capacitor shown in Figure 3.4a, with length $L$, a width $W$, and a separation of $t$ between the plates.[8]

The basic formulation for a parallel-plate capacitance is given by

$$C = \frac{k\varepsilon_0}{t} WL \tag{3.1}$$

where $\varepsilon_0$ is the permittivity of free space, $k$ is the relative dielectric constant of the insulator, and $t$ is the thickness of the insulating material. Although Equation (3.1) is only suitable for an isolated parallel-plate capacitor, it can give us some insight into the issues associated with deep submicron interconnect.

Consider the configuration of metal 1 and metal 2 wires shown in Figure 3.5. Three metal 1 wires are running adjacent to each other and going into the page. They are situated above a ground plane, which is the substrate. A metal 2 wire runs

[8] Note that these dimensions are *not* the channel length, channel width, and thin-oxide thickness. We have switched to the study of interconnect rather than devices.

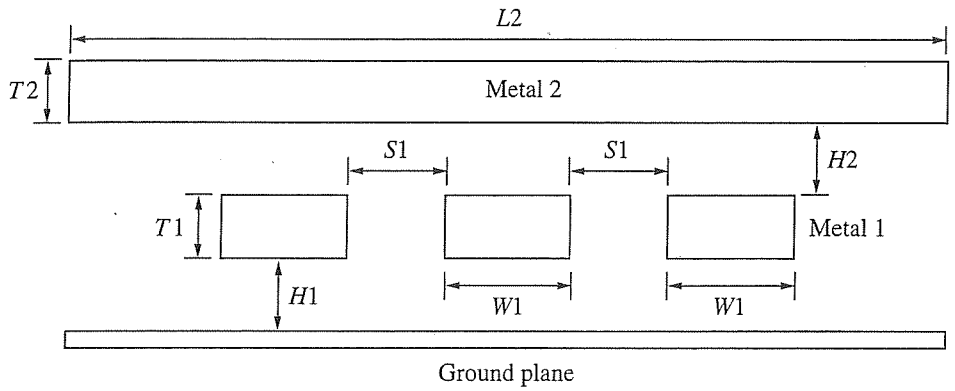



**Figure 3.5**
Wire dimensions for capacitance calculations.

on top of these wires. All the dimensions in the vertical direction are established during the fabrication process. That is, the height above the substrate and between different layers, $H$, and the thickness of the wires, $T$, are fixed values that cannot be modified by the designer. On the other hand, the horizontal dimensions are under the designers' control. Specifically, the width, $W$, spacing between wires, $S$, and the wire length, $L$, are all design variables.

If we focus on the center wire on metal 1 for a moment, there are many sources of capacitance. First, there are two *area* capacitances between it and the substrate below and the metal 2 above that depend on H1 and H2, respectively. Second, there is *lateral* capacitance between adjacent wires on the same level that is dependent on the spacing, S1. Finally, there are *fringing* capacitances between the conductor sidewalls and the upper conductor and lower substrate. Extracting the value of all of these capacitances would require three-dimensional analysis of the structure, and the total capacitance on the middle wire would be the sum of all the capacitances. Since capacitance controls the delay in our signals and determines the power dissipation, we seek to minimize the total capacitance seen by any wire.

There are many ways we can reduce the capacitance from a design perspective. The primary method is to space out the wires by making $S$ as large as possible. From a fabrication perspective, Equation (3.1) tells us that the thickness, $t$, should be made as large as possible and the dielectric constant, $k$, should be made as small as possible. Here, the thickness is the height of the insulator between two metal lines on different layers, called the interlayer dielectric (ILD). It is important in determining the capacitive coupling between wires on adjacent layers. Of course, if $H$ is too large, the overall height of the vias would increase. This tends to increase resistance and reduce reliability. The dielectric constant, $k$, of the insulating material is also important in determining the degree of capacitive coupling. The material between the metal lines is an insulator, typically silicon dioxide. The $k$ for silicon dioxide is approximately 4. Recently, process engineers have developed dielectrics with lower values of $k$ and these are referred to as *low-k* dielectrics. Values of $k \approx 3$ are in production use today. The goal is to reach a $k$ of roughly 2 within the next few years, if possible.

The discussion of interconnect capacitance will continue in later chapters. We now turn our attention to wire resistance. One of the main reasons for switching from Al to Cu is due to the resistance associated with the two materials. In the past, wire resistance was low and could be ignored. Wires were either too short to be of concern or they were so wide that the resistance was negligible. In modern technologies, the line widths are very small and the resistance has gone up considerably. This increase in resistance has led to a number of issues in design, namely, interconnect delay in signal lines and voltage drop in the power grid. These issues will be elaborated in later chapters. Here, we examine the resistance calculations and their implications for deep submicron wires.

The resistance of a material is given by

$$R = \frac{\rho L}{A} = \frac{\rho L}{TW}$$

where $\rho$ is the resistivity of the material in $\Omega$-cm, $L$ is the length of the wire, $T$ is the thickness of the wire, and $W$ is the wire width. This is illustrated in Figure 3.4b.

The leading term is an important one:

$$R_{sq} = \frac{\rho}{T}$$

where $R_{sq}$ is termed the *sheet resistance* and has the units of *ohms per square* ($\Omega/\square$). The sheet resistance of any metal layer can readily be computed from the resistivity and the thickness, $T$. The two most common materials for metal are aluminum, with $\rho = 2.7\ \mu\Omega$-cm, and copper,[9] with $\rho = 1.7\ \mu\Omega$-cm. As a point of comparison, tungsten has a much higher resistivity of 5.5 $\mu\Omega$-cm.

The resistance of a segment of wire is

$$R = R_{sq}\frac{L}{W}$$

The ratio of $L/W$ is referred to as the *number of squares* of wire. It is the aspect ratio of the wire in terms of length and width. Over the years, wires have been getting longer and narrower which increases the resistance since the number of squares increases. To keep resistance relatively low, the value of $T$ has not been scaled at the same rate as the minimum line width.

In Figure 3.6, the cross-sectional views of the wires are shown as we scale technology. In the 1970s, the line width was fairly large with 5 $\mu$m being a typical value. Today, minimum line widths are below 0.13 $\mu$m and the resistance per square has increased significantly. On the other hand, the thickness of the wire has not scaled as quickly as compared to the width to keep sheet resistivity low. The more important fact is that the number of squares is significantly larger, partially due to the narrower widths but also due to the fact that wires are continuing to get longer.

[9] In practice, the $\rho$ for copper is closer to 2.0 $\mu\Omega$-cm due to the cladding material needed to avoid diffusion into any neighboring oxide.

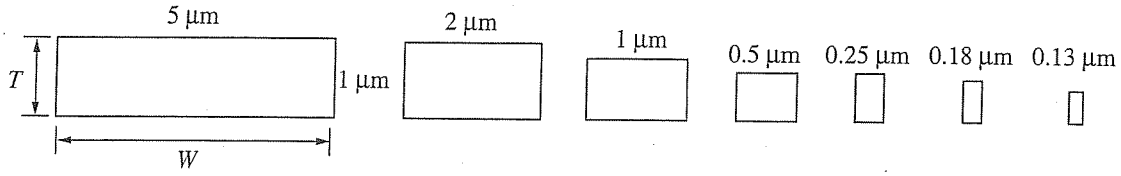



**Figure 3.6**
Interconnect cross-sections as technology scales.

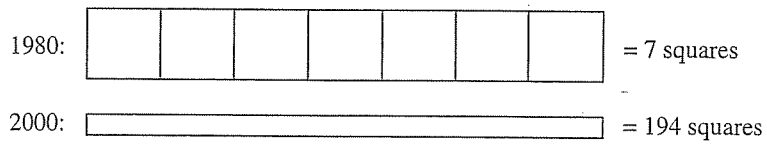
**Comparison of Al and Cu Wire Resistance** **Example 3.1**

**Problem:**

Compute the resistance for an aluminum wire in a 1980s 5 $\mu$m technology and a 0.18 $\mu$m technology of the year 2000. Compare this with a copper wire in the year 2002. First assume that the aluminum wire is 35 $\mu$m with a resistivity of 2.7 $\Omega$-cm and thickness of 1 $\mu$m. Then switch to a copper wire with a resistivity of 1.7 $\Omega$-cm and thickness of 0.4 $\mu$m that is also 35 $\mu$m long.

**Solution:**

Each wire is shown in plan view below. To determine the number of squares of resistance, we divide the length of the wire, $L$, by the width, $W$, for both cases. We can use Figure 3.6 to estimate the thickness. In the first case, the wire is 5 $\mu$m wide, 1 $\mu$m thick, and 35 $\mu$m long. In the second case, the wire is 0.18 $\mu$m wide and 0.5 $\mu$m thick. From these values, we can compute the resistance. Note that even for these short wires the ratio of the resistance values is 50X.




$$R_{5\ \mu\text{m-wire}} = 2.7\ \mu\Omega\text{-cm}/1\ \mu\text{m} \times 7 \approx 0.2\ \Omega \quad (1980)$$

$$R_{0.18\ \mu\text{m-wire}} = 2.7\ \mu\Omega\text{-cm}/0.5\ \mu\text{m} \times 194 \approx 10\ \Omega \quad (2000)$$

For the same wire in a 0.13 $\mu$m process, we have

$$R_{0.13\ \mu\text{m-wire}} = 1.7\ \mu\Omega\text{-cm}/0.4\ \mu\text{m} \times 269 \approx 11\ \Omega \quad (2002)$$

Note that the resistance for the wire in a 0.13 $\mu$m technology has been held to roughly the same value as in the 0.18 $\mu$m technology due to the switch to copper.

## 3.3 Layout Basics

An important step in the design of any IC is the layout of the chip, which defines the various layers associated with the masks used in fabrication. The handoff of the final layout from the designer to the fabrication facility is called *tapeout*. Figure 3.7 shows the tapeout process that involves the flow of mask data from the design house, which is the company designing the chip, to the foundry, which is the company that fabricates the chip. The chip design information is usually represented by the designer in graphical form using a layout tool, and then converted into a binary output format called *GDS-II stream format*.[10] This format describes the polygons and their $(x,y)$ positions in the layout that comprise each of the layers in the design. The geometric information must comply with a set of design rules that the foundry has painstakingly developed for a specific technology, such as 0.13 $\mu$m. The foundry also provides process parameters that are used for simulation purposes.

The goal of the layout process is to implement the design in a compact area while satisfying the *design rules* set by the foundry. Layout design is as much an art as it is a science, but there are some fundamental guidelines that must be adhered to if a chip is to be fabricated successfully. These guidelines, when violated, are flagged by modern CAD tools, called *design rule checkers* (DRCs), that can handle the billions of geometries needed to represent complex layouts. The design rules are

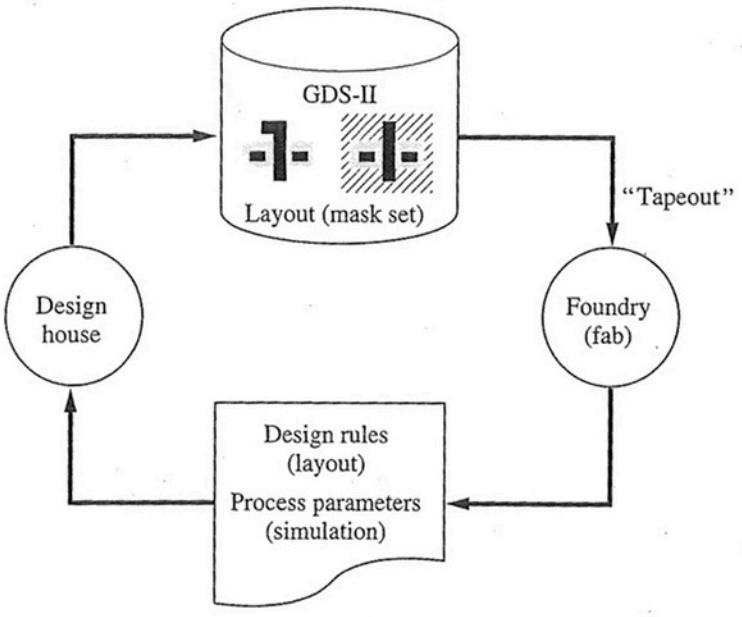



**Figure 3.7**
The flow of data between the design house and the foundry.

[10] GDS is a layout specification format originally developed by the Calma Corporation in the 1980s and later used as an industry standard.

a set of tolerances based on the minimum feature size imposed by a given technology. These tolerances are due to registration error in mask alignment going from one pattern to another, process control due to variation in exposure and etching, and overlap requirements to ensure low ohmic contact where necessary.

The minimum feature size for the layout varies according to the technology node chosen by the design house. Today, new designs are being implemented in 0.13 $\mu$m technology so the design rules are based on this minimum dimension. Since the minimum dimension may change from one technology to another, a technology-independent term, $\lambda$, was introduced in the 1980s. The minimum gate length was taken as $2\lambda$ and all design rules were based on this definition (note: this is not the channel length modulation parameter). As an example, if the minimum gate length is 200 nm, then $\lambda = 100$ nm. Similarly, for $L = 100$ nm, we set $\lambda = 50$ nm. While the notion of a linear scaling factor $\lambda$ is not used in advanced technologies, it is still instructive when describing design rules and it will be used frequently in this book.

In general there are two types of design rules. One set is associated with the resolution possible in a particular layer and the second set is used for alignment constraints between two different layers. These two types are illustrated in Figure 3.8.

In the case of rules concerning resolution, the minimum line width and spacing are defined. In Figure 3.8a, the minimum line width is $3\lambda$ and the minimum spacing is $3\lambda$. This is a typical pair of rules for the upper layers of metal.

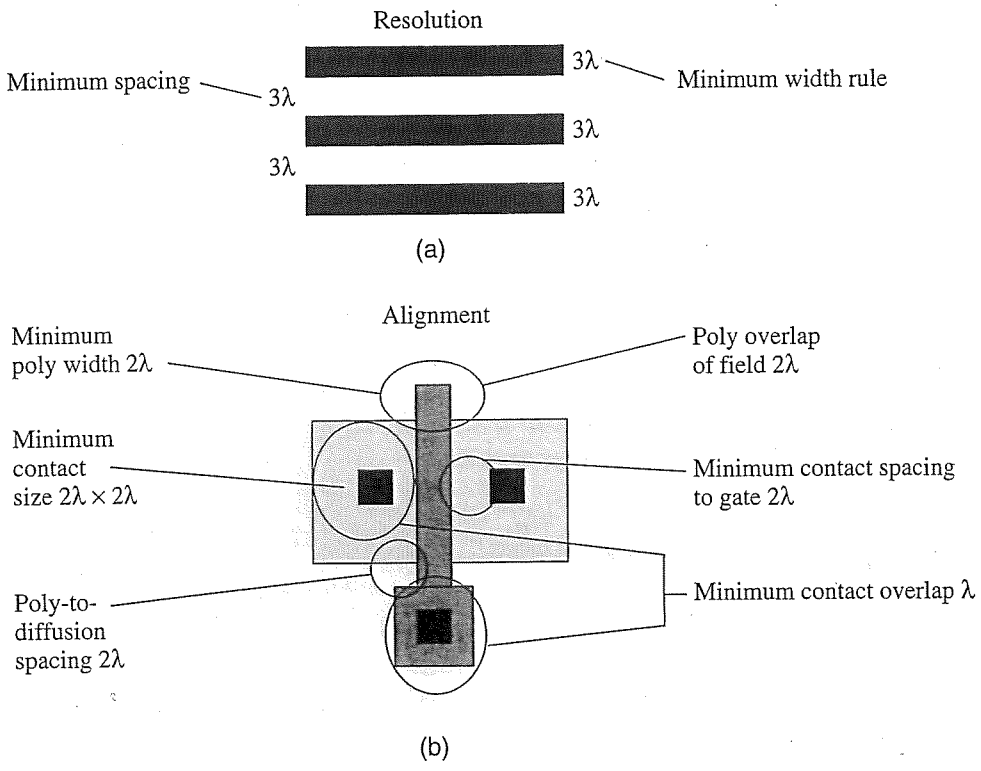



**Figure 3.8**
Two types of design rules: resolution and alignment.

The alignment rules are associated with two different layers, as shown in Figure 3.8b. For example, the poly overlap of field is set to $2\lambda$. This rule is required so that the gate will completely cover the diffusion region. Otherwise, the transistor would become a resistor since the diffusion region would extend from source to drain. Another set of alignment rules are associated with contacts. The contacts are used to connect metal to diffusion or metal to poly. The minimum size contact is $2\lambda \times 2\lambda$. However, to ensure good contact, the two materials must overlap the contact by a minimum amount and, at the same time, must be a safe distance from other layers. In this example, the overlap of the contact by diffusion and metal is set to $\lambda$ while the poly-to-contact spacing is set to $2\lambda$. These rules ensure that good contact is made between the two desired materials (diffusion and metal) but not between two undesired materials (poly and diffusion). Similar rules hold for poly-to-metal contacts.

More specific layout rules, including minimum (and sometimes maximum) sizes of features on all mask layers, accompany every IC manufacturing process. There are over one hundred such rules that must be followed to satisfy the fabrication requirements. These rules are checked by DRC programs and must pass cleanly before a design is accepted by the foundry service. For actual chip layout, the precise information should be obtained from the foundry service. Typically, these rules are specified in absolute dimensions rather than in terms of $\lambda$. The details of new fabrication technologies, including the layout rules, are generally proprietary and are rarely disclosed by manufacturers, except to their customers.

The design rules have implications in terms of the minimum-size transistor that can be fabricated. Consider Figure 3.9a where a minimum size NMOS transistor has been laid out in a *p*-well, with a well contact shown. The minimum channel length is $2\lambda$ as expected. However, the minimum channel width is $4\lambda$, not $2\lambda$ as one

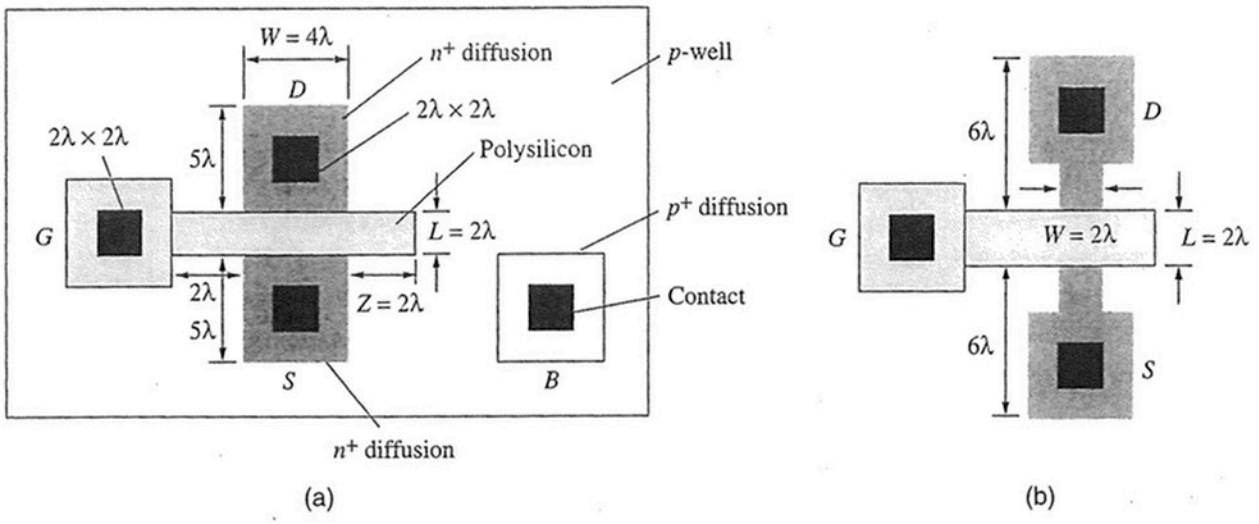



**Figure 3.9**
Alternative layouts of a minimum-size transistor.

might expect. This is due to the fact that the contact is $2\lambda \times 2\lambda$ and the diffusion must overlap it by $1\lambda$ on all sides. This adds $2\lambda$ to the width, which makes the minimum width equal to $4\lambda$. The diffusion regions must extend out a minimum of $5\lambda$ since the contact-to-diffusion edge is $\lambda$, and the contact-to-poly spacing is $2\lambda$. Therefore, the minimum diffusion area is $4\lambda \times 5\lambda$.

In order to create a transistor with $W = 2\lambda$, it will cost some extra area as shown in Figure 3.9b. The contacts must be pushed out and a narrow diffusion region created for the desired transistor width. The *p*-well edges (not shown) would also have to be pushed out due to design rules, and this may affect the spacing rules for nearby *n*-wells. While it is possible to create a minimum-size transistor when needed, it is more convenient to create a device with $W = 4\lambda$. It also has twice as much drive capability. The diffusion area is about the same in both cases.

The other interesting feature shown in Figure 3.9a is the well contact which is composed of a $p^+$ diffusion contact that will eventually connect to a metal line (not shown) that is attached to Gnd. This is the layout view of the bulk terminal of the transistor. The metal line, if added, would run across the well contact and also connect to the source of the NMOS transistor. This would ensure that the *pn* junctions were properly reverse-biased. Of course, the PMOS device has similar minimum size and well contact considerations that were just described for the NMOS device, with the polarities and materials of the opposite type.

## 3.4 Modeling the MOS Transistor for Circuit Simulation

Computer circuit simulators are essential tools in the analysis and design of MOS circuits. The computer program SPICE and its commercial derivatives, such as HSPICE™ and SmartSPICE™, are widely used for such tasks and are used for examples in this text. Other circuit simulators such as Spectre™ require similar modeling considerations.[11] One key point to stress is that a circuit simulator is not a circuit design tool but rather a circuit analysis tool. Novice designers have been known to use SPICE in a circuit design role, but it is not a replacement for thinking about how a circuit should work or how to design a circuit. The role of SPICE is to validate your design and its proper operation. It allows the designer to try various optimization techniques and carry out simulations that include process variations. In this mode, SPICE is one of the most useful tools in the integrated circuits industry.

### 3.4.1 MOS Models in SPICE

The SPICE tool simulates a user-specified circuit description with built-in models for each type of circuit element. The circuit description must include layout dimensions for each device to produce accurate results. The MOS transistor models are based on device physics and empirical equations derived from measurements of

[11] HSPICE is a trademark of Synopsys. SmartSPICE is a trademark of Silvaco. Spectre is a trademark of Cadence.

fabricated devices. The *device models* are developed by a group whose main responsibility is to accurately capture the behavior of semiconductor devices in the form of equations such as those described in Chapter 2 for $V_T$, $I_{DS}$, and capacitance. Once the *model developers* complete the device models, they are coded into a program such as SPICE and given a specific name. For example, LEVEL 1 (quadratic model), LEVEL 2, LEVEL 3, LEVEL 4 or 13 (BSIM1), LEVEL 39 (BSIM2), LEVEL 28, and LEVEL 49 (BSIM3) are well-recognized names within the industry for the different built-in models incorporated into SPICE over the past 25 years. Only one such model is invoked during any given run of SPICE.

Each of these models has a number of *model parameters* associated with them. The number of parameters can vary depending on the complexity of the model. Some of the parameters are related to the *physical* process technology, such as $t_{ox}$, $x_j$, $N_A$, $N_D$, $\mu_0$, and so on. Other *electrical* parameters are associated with the transistor current equations or capacitance equations such as $V_T$, $\gamma$, $\lambda$, and $C_{jo}$. Yet another set are used to capture certain advanced features of the devices such as short-channel effects, narrow-channel effects, and device resistances. Using these parameters, SPICE computes the currents and capacitances of the devices during model evaluation, using the device voltages $V_{GS}$, $V_{DS}$, and $V_{BS}$. If some parameters are not specified, they are computed internally by the program. As an example, if $\gamma$ is not specified, it will be computed using $N_A$.

The actual values assigned to the SPICE model parameters are generated through a process called *parameter extraction*. The group responsible for this activity is associated with the fabrication service, as shown in Figure 3.7, which in this case is the foundry. In parameter extraction, actual devices are measured in a variety of configurations. The measured data is fed to software programs that produce the parameter values. Some of the parameters have a physical meaning, so their values are known in advance, while others are empirical in nature and are only used to fit the equations to the measured data.

Once these parameters are extracted, they are placed in a file called a *model library*. There are many such MOS models in the model library. First, there are different sets of parameters for NMOS and PMOS devices. There are also different parameter sets for ranges of $W$ and $L$ values (called model *binning*) and for process variations. To fully capture all possible realistic scenarios, there may be dozens of different parameter sets in a model library.

After the models have been placed in the library, they are delivered to the design house for use in chip design. These model libraries should be "checked-out" thoroughly for accuracy and consistency using the SPICE program. This is done by running SPICE on a variety of different device sizes and operating conditions. Once they are approved for use, they are made available to the chip designers for actual simulations.

### 3.4.2 Specifying MOS Transistors

In Figure 3.10, an NMOS transistor layout is shown along with its corresponding symbol. The source and drain region extensions are specified by $Y_S$ and $Y_D$. We will assume that $Y_S = Y_D = Y$ for our purposes, but they may differ from device to

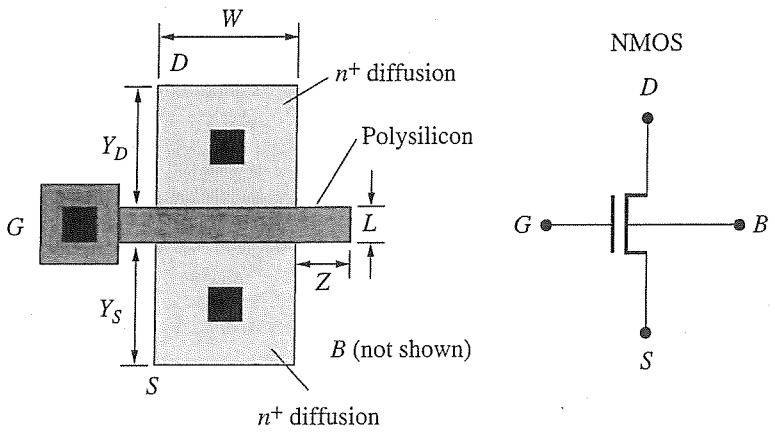



**Figure 3.10**
MOS layout and schematic for SPICE modeling.

device. There is a small poly gate overlap of the field given by $Z \times L$, which is usually small.

In the SPICE program, the user may specify the transistors and their connection with the format given below:

MOSFET

Mxxx D G S B mname *L* = value *W* = value AD = value PD = value AS = value PS = value

| | |
|---|---|
| Mxxx | The instance name of the MOSFET. |
| D | The drain node. |
| G | The gate node. |
| S | The source node. |
| B | The bulk node. |
| mname | The name of the model to be used. This is also where you distinguish between a PMOS and NMOS. |
| *L* | The length of the transistor. |
| *W* | The width of the transistor. |
| AD | The area of drain diffusion bottom region. |
| PD | The drain edges to be used for sidewall capacitance. |
| AS | The area of the source diffusion bottom region. |
| PS | The source edges to be used for sidewall capacitance. |

The specification begins with an "M" and a sequence of letters to form the instance name. This is followed by the drain, gate, source, and bulk connections. Next, the model name is specified. After the model, the dimensions of the device are provided

beginning with $L$ and $W$. The areas of the source and drain are provided as AS and AD, respectively, for the bottom junction capacitances. Typically, the area of the bottom is specified as $W \times Y$. This value is multiplied automatically by the junction capacitance coefficient for the bottom edge. The periphery of the source and drain is also provided to calculate the sidewall capacitance. The two parameters, PS and PD, should be set to $W$ since we are interested in the channel facing edge only.

Note that SPICE first precomputes effective values of $L$ and $W$:

$$\begin{aligned} L_{eff} &= L - \Delta L \\ W_{eff} &= W - \Delta W \end{aligned} \tag{3.2}$$

The values of $\Delta L$ and $\Delta W$ are based on systematic length or width reductions and process-related variations. We will use $W_{eff}$ and $L_{eff}$ below to conform to these precomputed values.

**Example 3.2** **SPICE MOS Transistor Specification**

To illustrate how a MOS transistor is specified to SPICE, we assume that $L = 0.2\ \mu$m, $W = 0.4\ \mu$m, and $Y = 0.5\ \mu$m. Then, AS = AD = W × Y = (0.4 $\mu$m) (0.5 $\mu$m) and PS = PD = 0.4 $\mu$m. If the drain node is `drainn`, and the gate node is `gaten`, while the source and bulk are grounded, we would specify the transistor instance as

```
M1 drainn gaten Gnd Gnd NMOS1 l=0.2u w=0.4u ad=0.2p pd=0.4u as=0.2p ps=0.4u
```

Here, the model invoked from the library is called NMOS1. The length, width, areas, and perimeters are all specified in absolute units of meters. Another alternative is to define a scale factor $\lambda = 0.1\ \mu$m (note this is not the channel length modulation parameter). Then all the transistor dimensions can be scaled relative to this value. This greatly improves the readability of the transistor description. It is specified using the .opt directive. For the same specifications, the device could be re-written as

```
.opt scale=0.1u      * Set scale factor=lambda
M1 drainn gaten Gnd Gnd NMOS1 l=2 w=4 ad=20 pd=4 as=20  ps=4
```

The model used by the simulator is given as NMOS1 in the above input specification. This would be retrieved from the model library and invokes one of the built-in device models. The LEVEL 1 and BSIM3 are the two device models considered in this book. The LEVEL 1 model employs the equations derived in Section 2.4. Today, BSIM3v3 is the workhorse model in industry and is adequate for most deep submicron MOS digital circuit analysis and design, provided parameters are extracted from measured data. This model is based on the description in Section 2.5. We now describe these two models in more detail.

## 3.5 SPICE MOS LEVEL 1 Device Model

The dc characteristics of the MOS LEVEL 1 model are defined by the device electrical parameters[12] VTO, KP, LAMBDA, PHI, and GAMMA. These parameters are directly associated with the model developed in Section 2.4. Ideally, the user should specify these parameters after measuring actual devices and extracting the corresponding values. However, these electrical parameters are computed by SPICE if process parameters, such as NSUB, TOX, and UO, are given; user-specified values always override internal calculation.

To illustrate one flow in SPICE, assume that VTO, NSUB, LAMBDA, TOX, and UO are specified by the library model. The computation of the threshold voltage begins with the calculation of the surface potential, PHI, based on the substrate doping level, NSUB:

$$\text{PHI} = 2 \times \frac{kT}{q} \ln\left(\frac{\text{NSUB}}{n_i}\right) \tag{3.3}$$

The oxide capacitance is computed using TOX:

$$C_{ox} = \frac{\varepsilon_{ox}}{\text{TOX}} \tag{3.4}$$

The body-effect parameter is then computed using NSUB and $C_{ox}$:

$$\text{GAMMA} = \frac{\sqrt{2\varepsilon_{si} q \times \text{NSUB}}}{C_{ox}} \tag{3.5}$$

Finally, the threshold voltage is computed using the following equation:

$$V_T = \text{VTO} + \text{GAMMA}(\sqrt{\text{PHI} - V_{BS}} - \sqrt{\text{PHI}}) \tag{3.6}$$

VTO is positive for enhancement-mode $n$-channel devices and negative for $p$-channel devices.

Next, the current equations are computed using the threshold voltage. If KP is specified, it is used directly in the current expression. If not, the mobility, UO, is used to compute it:

$$\text{KP} = \text{UO} \times C_{ox} \tag{3.7}$$

For the linear region, the current expression is

$$I_{DS} = \frac{W_{eff}}{L_{eff}} \frac{\text{KP}}{2} [2(V_{GS} - V_T)V_{DS} - V_{DS}^2](1 + \text{LAMBDA} \times V_{DS}) \qquad V_{GS} \geq V_T \quad V_{DS} \leq V_{GS} - V_T \tag{3.8}$$

[12] SPICE parameters will be indicated with capitalization of all text when referring to the parameter.

For the saturation region, the current expression is

$$I_{DS} = \frac{W_{eff}}{L_{eff}} \frac{\text{KP}}{2} (V_{GS} - V_T)^2 (1 + \text{LAMBDA} \times V_{DS}) \qquad (3.9)$$

$$V_{GS} \geq V_T \quad V_{DS} \geq V_{GS} - V_T$$

Note that the channel length modulation parameter, LAMBDA, is used in both linear and saturation expressions for continuity of the current. It is important in SPICE that the equations be continuous from one region to another. Otherwise, the numerical methods used in the program have some difficulty converging to a solution. For this reason, the linear and saturation regions both require a term, $(1 + \text{LAMBDA} \times V_{DS})$, even though its effect is most prominent in the saturation region. As mentioned before, if any of the values of GAMMA, KP, or PHI are specified in the model library it would override the calculations given above.

The capacitances of MOS devices, modeled in several parts, are defined so they may be calculated easily from actual circuit layouts. Three constant capacitors, CGSO, CGDO, and CGBO, represent gate-source, gate-drain, and gate-body overlap capacitances. The CGBO term is due to the gate extension into the field region, defined by the region $Z \times L$ in Figure 3.10. The thin-oxide gate capacitance is calculated by the program as a function of applied voltages and distributed among the gate, source, drain, and body regions. The charge storage effects are included only if TOX is specified in the input description. There are two built-in models of the gate capacitances. LEVEL 1 uses the piecewise-linear voltage-dependent capacitance model proposed by Meyer. This model does not necessarily conserve charge and therefore must be used with care. It is not suitable for many circuits that rely on charge conservation of the model for proper operation. A more advanced model is available in higher-level MOS models such as BSIM3, as discussed in the next section.

The junction capacitances, for both source-body and drain-body *pn* junctions, are divided into bottom and periphery components. The junction capacitances are determined by the parameters CJ, CJSW, MJ, MJSW, and PB. Note that CJ is specified in F/m$^2$ and must be multiplied by the area of the source or drain, whereas CJSW is specified in F/m and must be multiplied by the perimeter of the source or drain. CJSW is premultiplied with the junction depth, XJ. The following equations are used to compute the capacitances:

$$C_{JD} = \frac{\text{CJ} \times \text{AD}}{\left(1 - \dfrac{V_J}{\text{PB}}\right)^{\text{MJ}}} + \frac{\text{CJSW} \times \text{PD}}{\left(1 - \dfrac{V_J}{\text{PB}}\right)^{\text{MJSW}}} \quad \text{drain junction}$$

$$C_{JS} = \frac{\text{CJ} \times \text{AS}}{\left(1 - \dfrac{V_J}{\text{PB}}\right)^{\text{MJ}}} + \frac{\text{CJSW} \times \text{PS}}{\left(1 - \dfrac{V_J}{\text{PB}}\right)^{\text{MJSW}}} \quad \text{source junction} \qquad (3.10)$$

When specifying the information for a given transistor, it is important to include AS, AD, PS, and PD so that the correct calculations for junction capacitance can be performed.

There is some overlap among the parameters describing the junction current; for example, the reverse current can be input either as IS (in A) or as JS (in A/m$^2$). Whereas the first is an absolute value, the second is multiplied by AD and AS to give the reverse current of the drain and source junctions, respectively. This flexibility has been provided so that junction characteristics can either be entered as absolute values on model statements or related to areas AD and AS entered on device statements.

The MOS transistor has series resistances associated with the source and drain regions. If needed, these parasitic resistances can be expressed as either RD and RS (in $\Omega$) or RSH (in $\Omega$ per square), the latter being multiplied by the number of squares NRD and NRS specified on the device instance line.

The reader should not be overwhelmed by the number of parameters and their detailed meanings. This is intended to give you the background needed when using SPICE. In fact, the LEVEL 1 device model has more parameters than the ones mentioned above. The complete list of SPICE parameters used in LEVEL 1 is given in Table 3.1 at the end of this chapter together with the corresponding symbols used in this text.

### 3.5.1 Extraction of Parameters for MOS LEVEL 1

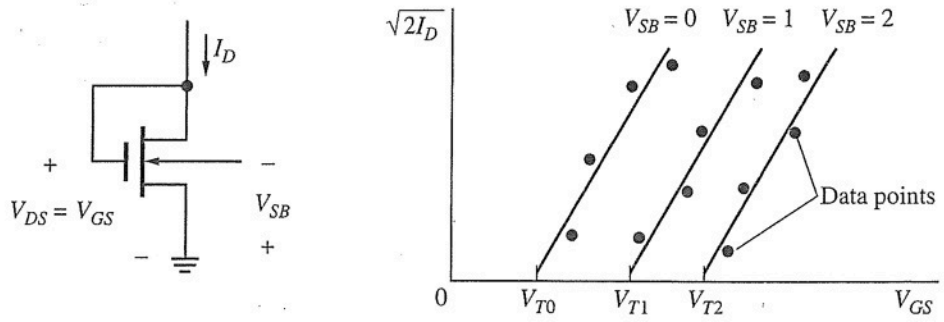
To obtain satisfactory results for the LEVEL 1 model in circuit analysis and design, measured data on samples of MOS transistors must be obtained. The parameters may be extracted by fitting measured data to the device equations over the intended operating range of voltages and currents. These are, of course, the first-order equations associated with long-channel devices, so their use should be restricted to these types of devices. Similar steps for fitting parameters to measured data are used for the short-channel devices.

Methods of taking and reducing data to determine VTO, GAMMA, KP, and LAMBDA are illustrated in Figure 3.11. If the gate is connected to the drain, the device will be in the saturation region of operation, since $V_{DS} > V_{GS} - V_T$. In this condition, the long-channel saturation region equation can be rewritten as

$$\sqrt{2I_D} = \sqrt{\frac{W_{eff}\mathrm{KP}}{L_{eff}}}(V_{GS} - V_T) \tag{3.11}$$

By plotting $\sqrt{2I_D}$ versus $V_{GS}$ as in Figure 3.11a, the $x$-intercept can be extracted as $V_T$ and the slope of the line is $\sqrt{\frac{W_{eff}\mathrm{KP}}{L_{eff}}}$. A number of measurements of the current can be made with $V_{BS} = 0$. This will produce the zero-bias threshold voltage, VTO. The value of KP can be extracted from the slope. By adjusting $V_{BS}$, repeating the measurements will produce additional values of the threshold voltage. Using these $V_T$ values, the body-effect coefficient, GAMMA, can be obtained.

A second set of measurements is used to determine the channel-length modulation parameter LAMBDA, if needed. As seen in Figure 3.11b, the device can be disconnected at the gate and drain and two separate voltages applied. Keeping the device in saturation, two values of $V_{DS}$ can be applied and the corresponding currents measured. From these two measurements, the value of LAMBDA can be extracted.

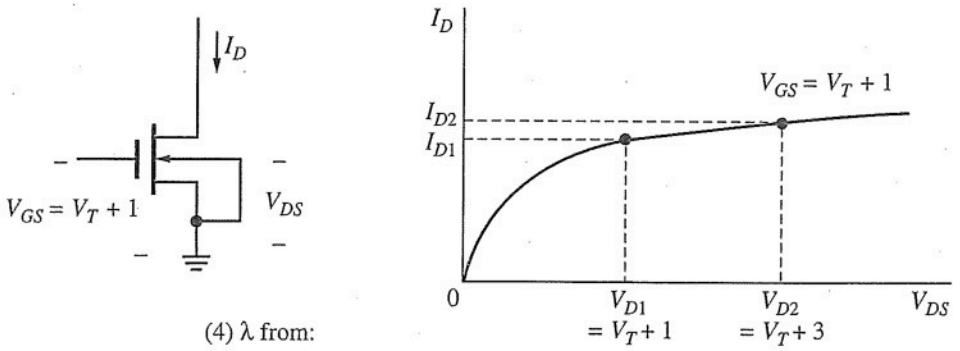



(1) $V_{T0}$ from intercept

(2) $k = k' \dfrac{W}{L}$ from slope: $\sqrt{k} = \dfrac{\sqrt{2I_D}}{V_{GS} - V_T}$

(3) $\gamma = \dfrac{V_T(V_{SB}) - V_{T0}}{\sqrt{V_{SB} - 2\phi_F} - \sqrt{2\phi_F}}$ $\qquad 2\phi_F \cong 0.6$ V

(a)




(4) $\lambda$ from:

$$\frac{I_{D2}}{I_{D1}} = \frac{1 + \lambda V_{D2}}{1 + \lambda V_{D1}}$$

(b)

**Figure 3.11**
MOS transistor parameter extraction.

**Example 3.3 SPICE LEVEL 1 Model Specification**

Assume that after parameter extraction, the following values are determined:

$V_{T0} = 0.5$ V, $k' = \mu_0 C_{ox} = 300$ $\mu$A/V$^2$, $|2\phi_F| = 0.8$ V, $\gamma = 0.4$ V$^{1/2}$ and $\lambda = 0$

Then the corresponding SPICE parameters are

VTO = 0.5 V, KP = 300(10$^{-6}$), PHI = 0.8, GAMMA = 0.4, and LAMBDA = 0.

The library model would be specified as follows:

```
.model NMOS1 nmos level=1 vto=0.5 kp=300u phi=0.8 gamma=0.4 lambda=0
```

In a complete LEVEL 1 specification, many more parameters would be extracted and specified in the model line, but this example is only intended to show the process involved.

After specifying the circuit description and connecting it to the appropriate model library, the user provides a series of control statements to direct the analysis, and plot or print results. The details of this aspect of SPICE are program-specific. Tutorial information of this nature is provided in Appendix A. More complete user guides are available from the CAD vendors, and in the references listed at the end of this chapter.

While these examples illustrate the extraction of specific model parameters for LEVEL 1, the same basic concepts can be used to extract all of the needed parameters for the advanced models in SPICE. These models are kept in library files that are associated with a particular technology and invoked by the user when simulating a given circuit. Normally, the user should not modify the parameters in these files, although the models should be validated using SPICE. Since the models will not be generally viewed by users, those who are new to the field may choose to skip the next section without any loss to continuity.

## *3.6 BSIM3 Model

Over the past 30 years much effort has gone into modeling the MOS transistor. Scaling has made what were previously second- and third-order effects into first-order effects. Analytical equations much more sophisticated than those presented thus far have been developed. More advanced SPICE models are available, such as LEVEL 2 and 3 models, and more recently, BSIM1, BSIM2, BSIM3, and BSIM4.

BSIM (Berkeley Short-channel IGFET Model) was developed at the University of California at Berkeley in the 1980s and 1990s, and work on the latest model continues today. BSIM3v3 is currently the most popular model for deep submicron devices. It is based on a quasi-two-dimensional model of the MOS device, with both physics-based and empirically-based equations. There are over 300 parameters in the complete model so the details of BSIM3v3 are beyond the scope of this book. However, the parameters for the BSIM3v3 are listed in Table 3.2 together with the corresponding symbols at the end of this chapter.

The goal of BSIM3 is to capture the key features of the dc and ac behavior of the deep submicron MOS transistor. Conceptually, the equations of Section 2.5 with velocity saturation form the basis of the model. The model in BSIM3v3 is much more elaborate and considers detailed physical effects as well as measured performance of fabricated devices in its implementation. We first describe the binning feature of BSIM3. Next, the threshold voltage equation, the mobility modeling, and current voltage expressions are presented. Finally, the capacitance modeling is described.

### 3.6.1 Binning Process in BSIM3

One of the limitations of any device model is that it is only as good as the parameter extraction that is carried out during the characterization process. Unfortunately,

even careful parameter extraction for one device does not track across all the $W$ and $L$ sizes encountered in a design. As a result, a number of models are prepared for one technology as illustrated in Figure 3.12. Minimum and maximum widths and lengths are selected and a number of bins are defined. Typically, measured devices lie at the center of the bins (roughly where the numbers are placed) and are used to characterize the entire bin. A large number of parameters are dedicated to the binning process in BSIM, so it is important to understand this concept.

As an example, a total of nine bins are shown in the figure. The number and ranges for the bins will vary from foundry to foundry. Fortunately, in digital design, the minimum length is used for most devices, so the bins labeled 1, 4, and 7 are of most interest. However, the user should realize that different models may be invoked as the width is modified. Certain boundaries of the bins may have discontinuities that produce "interesting" results in the circuit simulator. As a side effect of this *binning process,* two devices near a boundary may have different drive capability if the threshold voltage has a sharp discontinuity across the boundary. It is worthwhile to plot, for example, $V_T$ versus $L$ and $V_T$ versus $W$ in SPICE to identify any possible issues in the models before using them in a design, especially when a new set of models first arrives from the foundry. Otherwise, the fabricated design may not match the expected performance from simulations.

### 3.6.2 Short-Channel Threshold Voltage

The threshold voltage expression was derived earlier in Section 2.3 as

$$V_T = V_{T0} + \gamma(\sqrt{2|\phi_F| + V_{SB}} - \sqrt{2|\phi_F|})$$

There are a number of effects that change the threshold voltage relative to the first-order equation, which we now explore. In BSIM3, the $V_T$ formula starts with the same equation:

$$V_T = \text{VTHO} + \gamma(\sqrt{\text{PHI} - V_{BS}} - \sqrt{\text{PHI}}) \tag{3.12}$$

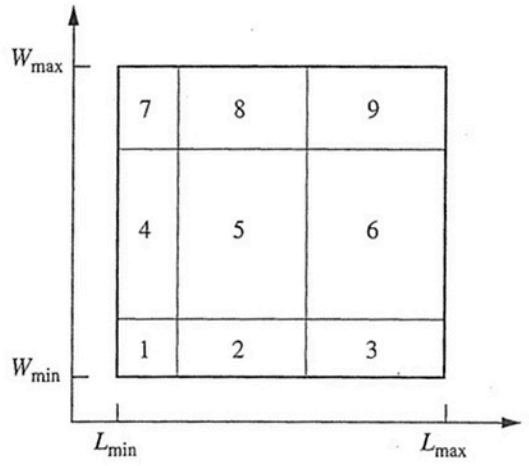



**Figure 3.12**
Model binning process.

Here, the value of VTHO is usually specified or computed using the flat-band voltage, VFB. However, this formulation does not accurately capture the vertical doping profile in the channel region. The $\gamma$ term is actually dependent on the different doping levels in the channel *and* substrate regions. The surface doping is due to the threshold implant, NCH, while the doping in the substrate is NSUB, a much lower value. This variation in the doping profile must be modeled for accuracy.

In BSIM3, the term $2|\phi_F|$ is called PHI and is given by

$$\text{PHI} = 2\frac{kT}{q}\ln\left(\frac{\text{NCH}}{n_i}\right) \tag{3.13}$$

where NCH is the doping of the channel region.

To properly capture the effects of the two doping levels, the threshold voltage equation is modified as follows. First, two $\gamma$ parameters are defined as

$$\begin{aligned}\gamma_1 &= \frac{\sqrt{2\varepsilon_{si}q \times \text{NCH}}}{C_{ox}} \\ \gamma_2 &= \frac{\sqrt{2\varepsilon_{si}q \times \text{NSUB}}}{C_{ox}}\end{aligned} \tag{3.14}$$

The oxide capacitance is computed as usual using

$$C_{ox} = \frac{\varepsilon_{ox}}{\text{TOX}} \tag{3.15}$$

Then, the two new parameters, K1 and K2, are defined using $\gamma_1$ and $\gamma_2$:

$$\begin{aligned}\text{K1} &= f(\gamma_2, \text{PHI}) \\ \text{K2} &= f(\gamma_1, \gamma_2, \text{PHI})\end{aligned} \tag{3.16}$$

Effectively, the K1 and K2 terms model the vertical profile in the channel region as an abrupt transition from NCH to NSUB at some depth $x_t$. Then, $V_T$ is determined by

$$V_T = \text{VTHO} + \text{K1}(\sqrt{\text{PHI} - V_{BS}} - \sqrt{\text{PHI}}) - \text{K2} \times V_{BS} \tag{3.17}$$

There are two additional $V_T$ effects that are worth noting. Both $L_{eff}$ and $W_{eff}$ have an impact on the threshold voltage. The effect of reducing $L_{eff}$ is to reduce the threshold voltage. As the channel length decreases, the depletion regions of the source and drain move closer together and actually aid in the depletion process of the channel region. The effect of this charge sharing is shown in Figure 3.13a. This is referred to as the classical short-channel effect (SCE). However, there is another effect that tends to increase the threshold voltage as $L_{eff}$ is reduced, as shown in Figure 3.13b. Due to oxidation-enhanced diffusion, a process whereby impurities gather at point defects at the two gate edges during oxidation, the surface doping levels in the channel region in short channels may be much higher than in long channels. This so-called reverse short-channel effect (RSCE), or threshold voltage roll-up, causes

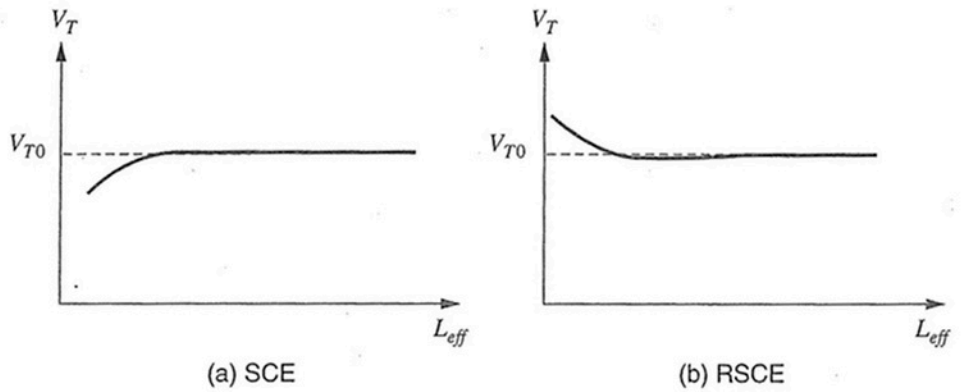



**Figure 3.13**
SCE and RSCE characteristics.

$V_T$ to increase due to the lateral nonuniformity of the doping across the channel. Since SCE and RSCE act in opposite directions, the precise doping profile determines the dominance of one effect over the other as the channel length is reduced. Typically, the $V_T$ "rolls up" as $L_{eff}$ is decreased due to RSCE and then drops off in value as the classical SCE takes over. Both can be incorporated in a short-channel effect (SCE) term $\Delta V_{SCE}$. It is useful to plot $V_T$ versus $L_{eff}$ for a model to observe these effects and also to examine any issues surrounding the binning process.

In some devices, the width can also play a role in determining the threshold voltage. If the channel width is close to the minimum width, the threshold voltage can actually increase in value. This is due to the fact that the poly gate overlap of the field (in Figure 3.10, the area defined by $Z \times L$) induces depletion-layer charge in the field region and this must be compensated for by an additional gate voltage. Its relative importance is reduced as the channel width increases, but it must be accounted for in narrow devices to derive accurate thresholds. This is called the *narrow-channel effect* (NCE). A term can be added to the threshold voltage, $\Delta V_{\text{NCE}}$, to account for this effect. It is also useful to plot $V_T$ versus $W_{eff}$ to observe this effect and also to examine any issues surrounding the binning process.

Yet another factor to be included in the $V_T$ calculation is the effect the drain-source voltage has on lowering the barrier for current flow. Normally, the gate-source voltage is responsible for inverting the surface by $2|\phi_F|$. In short-channel devices, the depletion region around the drain increases as $V_{DS}$ increases, and it extends into the channel region. Conceptually, the drain-source voltage is assisting the gate voltage in the depletion process. It is referred to as drain-induced barrier lowering (DIBL). As the channel length is reduced, the potential barrier between the source and drain is reduced even without the application of a drain-source voltage due to the SCE described above. However, the potential barrier is reduced even further by applying $V_{DS}$ because it enhances the depletion region in the channel.

Figure 3.14 illustrates this effect. At low values of $V_{DS}$, the threshold voltage is the same as the long-channel case. However, as $V_{DS}$ increases, a decrease in $V_{T0}$ is

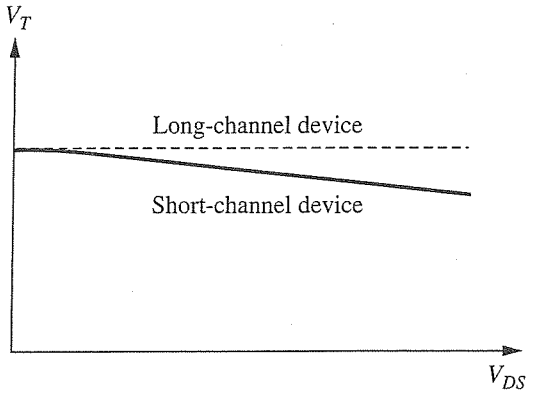



**Figure 3.14**
Drain-induced barrier lowering (DIBL).

observed that is almost linear with $V_{DS}$. An adjustment of the threshold voltage can accommodate the DIBL effect. We can subtract a term $\eta V_{DS}$ from the threshold voltage expression to account for the $V_{DS}$ dependence.

The final form of the threshold voltage expression including all effects mentioned here is given by

$$V_T = \text{VTHO} + \text{K1}(\sqrt{\text{PHI} - V_{BS}} - \sqrt{\text{PHI}}) - \text{K2} \times V_{BS} - \eta V_{DS} - \Delta V_{\text{SCE}} + \Delta V_{\text{NCE}} \tag{3.18}$$

### 3.6.3 Mobility Model

A variety of mobility models are available in BSIM3 to capture the vertical and horizontal field effects. In all these models, the nominal mobility value is UO, and the parameters that modify this nominal value are UA, UB, and UC. One particular model for the vertical field effect is given in the equation below. It includes the effect of the vertical field and the substrate bias voltage, $V_{BS}$:

$$\mu_v = \frac{\text{UO}}{1 + (\text{UA} + \text{UC} + V_{BS})\left(\dfrac{V_{GS} - V_T}{t_{ox}}\right) + \text{UB}\left(\dfrac{V_{GS} - V_T}{t_{ox}}\right)^2} \tag{3.19}$$

The horizontal field component is based on the piecewise-continuous modeling of mobility as a function of the field, as described in Chapter 2. The parameter VSAT is the saturation velocity that is provided to BSIM3 and is used to compute the critical field

$$E_c = \frac{2\text{VSAT}}{\mu_v} \tag{3.20}$$

This relationship is used to guarantee continuity in the velocity versus electric field characteristic.

### 3.6.4 Linear and Saturation Regions

With the threshold voltage and mobility values determined, BSIM3 can compute the current through the device in the linear or saturation region using the operating voltages of the device. These equations are similar in form to those derived in Section 2.5, although much more complex to account for additional effects. The details of these equations are not critical to our discussion here, but the interested reader may consult any of the references listed at the end of this chapter for further information.

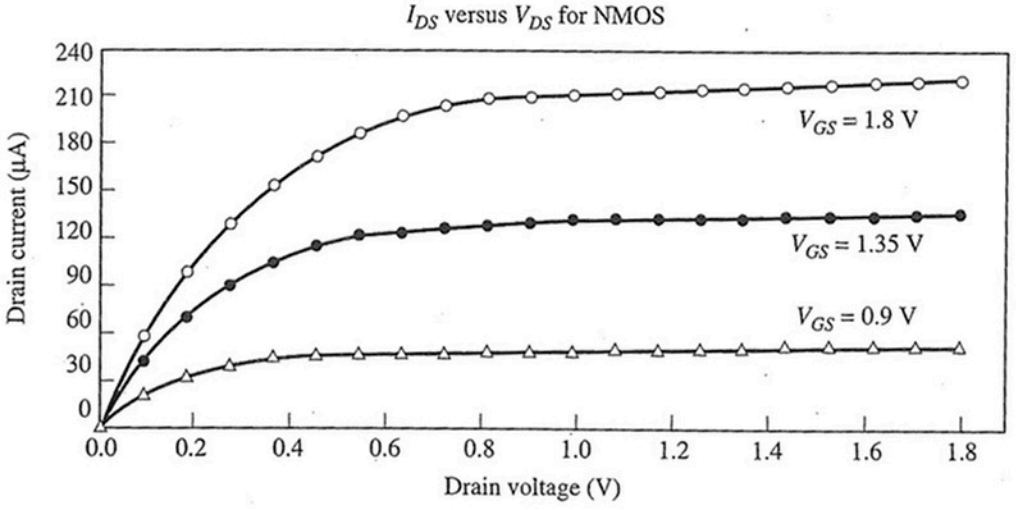
**Example 3.4** **BSIM3 I-V Characteristics for 0.18 $\mu$m CMOS**

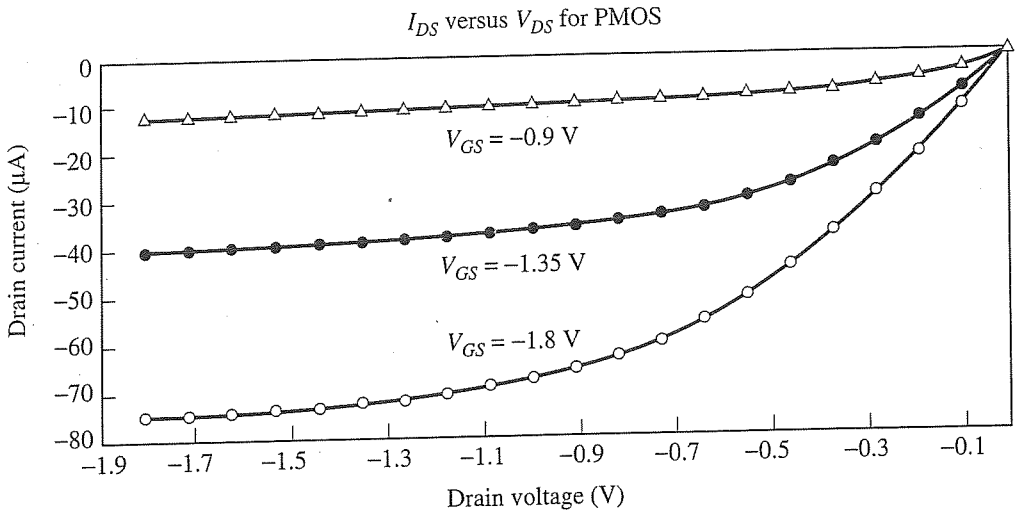
**Problem:**

Using a BSIM3 model for a 0.18 $\mu$m technology, plot $I_{DS}$ versus $V_{DS}$ for both PMOS and NMOS by sweeping $V_{DS}$ from 0 to $V_{DD} = 1.8$ V at intervals of $V_{GS} = 1/2\ V_{DD}$, 3/4 $V_{DD}$ and $V_{DD}$. For convenience, let $W = 0.4$ $\mu$m and $L = 0.2$ $\mu$m, and set $\lambda = 0.1$ $\mu$m. What is the ratio of saturation currents between PMOS and NMOS at $|V_{GS}|=|V_{DS}|=V_{DD}$?

**Solution:**

```
*SPICE Input File
.param      Supply=1.8          * Set value of Vdd
.lib        'bsim3v3.cmos.18um' * Set 0.18um library
.opt        scale=0.1u          * Set lambda
mp    drainp gatep Vdd   Vdd PMOS l=2 w=4 ad=20 pd=4 as=20 ps=4
mn    drainn gaten Gnd   Gnd NMOS l=2 w=4 ad=20 pd=4 as=20 ps=4
Vdd   Vdd    0 'Supply'
Vgsp  Vdd    gatep      dc
Vgsn  gaten  0          dc
Vdsp  Vdd    drainp     dc
Vdsn  drainn 0          dc
.dc   Vdsp   0 'Supply' 'Supply/20' Vgsp 0 'Supply' 'Supply/4'
.dc   Vdsn   0 'Supply' 'Supply/20' Vgsn 0 'Supply' 'Supply/4'
.plot dc     I1(mp)
.plot dc     I1(mn)
.end
```






The ratio of the currents in saturation is obtained by using the topmost curve in each plot and computing the currents at $V_{DS} = 1.8$ V.

$$\text{Ratio} = \frac{I_{DS,N}(V_{GS} = V_{DD}, V_{DS} = V_{DD})}{I_{DS,P}(V_{GS} = V_{DD}, V_{DS} = V_{DD})} = \frac{220\ \mu\text{A}}{78\ \mu\text{A}} \approx 2.8$$

This is close to the ratio of 2.4 obtained in the previous chapter from the velocity saturation model.

**Example 3.5**

**SPICE Plots to Compute $I_{on}$**

**Problem:**

Calculate the current per unit of width in $\mu$A/$\mu$m in the saturation region for both PMOS and NMOS at $V_{GS} = V_{DD}$ for a 2 × minimum width transistor and a 20 × minimum width transistor. Use the 0.18 $\mu$m technology data of Example 3.4.

**Solution:**

```
*SPICE Input File
.param       Supply=1.8                  * Set value of Vdd
.lib         'bsim3v3.cmos.18um'         * Set 0.18um library
.opt  scale=0.1u                         * Set lambda
mp    Gnd    gatep Vdd    Vdd    PMOS l=2 w=40 ad=200 pd=40 as=200 ps=40
mn    Vdd    gaten Gnd    Gnd    NMOS l=2 w=40 ad=200 pd=40 as=200 ps=40
Vdd   Vdd    0     'Supply'
Vgsp  Vdd           gatep 'Supply'
Vgsn  gaten  Gnd           'Supply'
.dc   Vgsp   0     'Supply'     'Supply'
.dc   Vgsn   0     'Supply'     'Supply'
.plot dc     I1(mp)
.plot dc     I1(mn)
.end
```

Plots similar to those in Example 3.4 are produced. From these curves, the current per unit width values are calculated by using the curve associated with $V_{GS} = 1.8V$ and dividing by the width of the device:

| Width | PMOS ($\mu$A/$\mu$m) | NMOS ($\mu$A/$\mu$m) |
|---|---|---|
| 2× | 195 | 550 |
| 20× | 225 | 570 |

The saturation currents can be quickly estimated using these values if we know the device width.

We will use 550 $\mu$A/$\mu$m for the NMOS device and 200 $\mu$A/$\mu$m for the PMOS device, based on these results. Similar results can be obtained for 0.13 $\mu$m technology. These values are referred to as $I_{on}$.

### 3.6.5 Subthreshold Current

The subthreshold conduction region, where $V_{GS} < V_T$, is an important consideration in a multi-million transistor deep submicron design. The key issue is that the number of devices that are leaking current is so large that it consumes appreciable power. Furthermore, any dynamic logic circuits[13] are prone to charge leakage since devices are not fully turned off. A number of models exist in BSIM3 to model this region of operation. One such model is

$$I_{sub} = \beta_{sub} V_{th}^{\,2} e^{(V_{GS} - V_T - V_{effect})/nV_{th}} (1 - e^{-V_{DS}/V_{th}})$$

$$\beta_{sub} = \mathrm{UO} \times C_d W_{eff}/L_{eff} \qquad C_d = \sqrt{\frac{q\varepsilon_{si}\mathrm{NCH}}{\mathrm{PHI}}} \tag{3.21}$$

This model is similar in form to that described in Chapter 2. Note that the low-field mobility is used here since the gate voltage is rather small.

**Example 3.6** **SPICE Plots of BSIM3 Model in Subthreshold Region**

**Problem:**

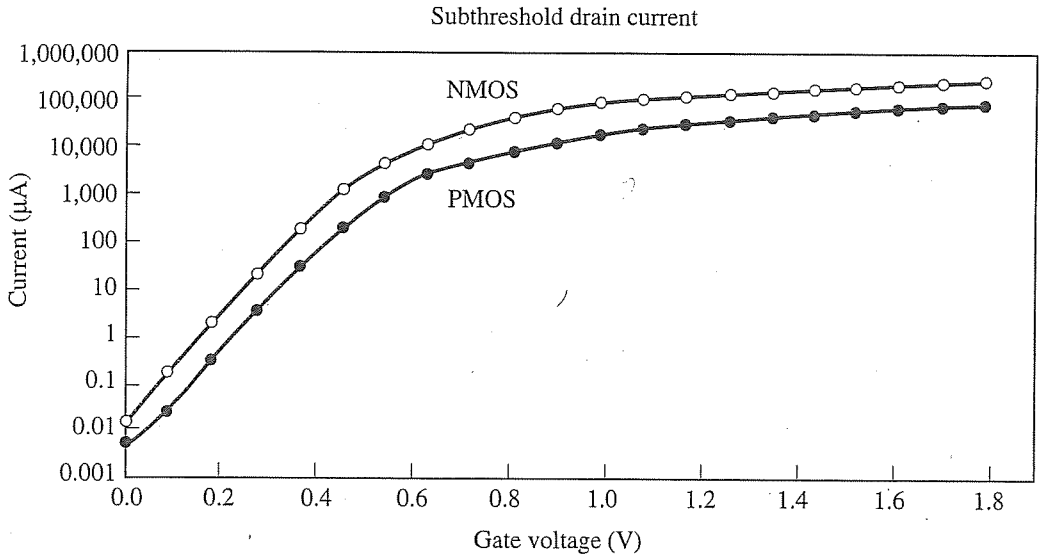
Plot $\log(I_{DS})$ versus $V_{GS}$ for both PMOS and NMOS by sweeping $V_{GS}$ from 0 to $V_{DD}$ with $V_{DS} = V_{DD}$. Compute the current per unit width in the subthreshold region, called $I_{off}$, for each device.

[13] Dynamic circuits are the topic of Chapter 7.

**Solution:**

```
*SPICE Input File
.param       Supply=1.8        * Set value of Vdd
.lib         'bsim3v3.cmos.18um'     * Set 0.18um library
.opt  scale=0.1u          * Set lambda
mp    Gnd    gatep Vdd    Vdd    PMOS l=2 w=4 ad=20 d=4 as=20 ps=4
mn    Vdd    gaten Gnd    Gnd    NMOS l=2 w=4 ad=20 pd=4 as=20 ps=4
Vdd   Vdd    0     'Supply'
Vgsp  Vdd    gatep 'Supply'
Vgsn  gaten  Gnd          'Supply'
.dc   Vgsp   0     'Supply'     'Supply/20'
.dc   Vgsn   0     'Supply'     'Supply/20'
.plot dc     I1(mp)
.plot dc     I1(mn)
.end
```

When $V_{GS}$ = 0.0 V, the $I_{off}$ for the NMOS device is 0.01 nA/0.4 $\mu$m = 25 pA/$\mu$m. For the PMOS device, $I_{off}$ = 0.008 nA/0.4 $\mu$m = 20 pA/$\mu$m. These levels are relatively small compared to $I_{on}$.




### 3.6.6 Capacitance Models

Many of the comments made earlier about the capacitance models for LEVEL 1 also apply to advanced models. However, in the advanced models, charge storage effects are much more accurately represented. Internal to the SPICE program, the capacitances are formulated using charge-conserving equations. From a user's perspective, the information provided to the program is essentially the same. The main difference between LEVEL 1 and BSIM3 is that SPICE will produce much more accurate results, especially for circuits that rely on charge conservation for proper operation.

The nonlinear thin-oxide capacitance due to TOX is calculated by the program as a function of applied voltages and distributed among the gate, source, drain, and body regions. The three capacitances, $C_{gs}$, $C_{gd}$, and $C_{gb}$, are computed based on the region of operation. It is important to point out that *charges* associated with the gate, drain, source, and bulk terminals are computed first, and then capacitances are derived from these terms. In this manner, the total charge in the system can be conserved, thereby producing much more accurate results than LEVEL 1. In addition, three constant capacitors, CGSO, CGDO, and CGBO, represent gate-source, gate-drain, and gate-body overlap capacitances. Today, CGSO and CGDO are due to a combination of fringing capacitances from the edge of the poly to the surface of the silicon and lateral diffusion. The CGBO term is due to the gate extension into the field region and is relatively small.

The nonlinear depletion-layer capacitances, for both source-body and drain-body *pn* junctions, is divided into bottom and periphery parameters, CJ and CJSWG, which vary as the MJ and MJSWG power of junction voltage, respectively. The sidewall capacitance is primarily due to the channel facing edge since the other three edges interface to STI oxide and are therefore small. When specifying the information for a given transistor, it is important to include AS, AD, PS, and PD so that the correct calculations for junction capacitance can be performed. For deep submicron devices, PS and PD should be set equal to $W$ to capture the channel facing edge component. The following equations are used to compute the capacitances:

$$C_{JD} = \frac{\text{CJ} \times \text{AD}}{\left(1 - \dfrac{V_J}{\text{PB}}\right)^{\text{MJ}}} + \frac{\text{CJSWG} \times \text{PD}}{\left(1 - \dfrac{V_J}{\text{PB}}\right)^{\text{MJSWG}}} \quad \text{drain junction}$$

$$C_{JS} = \frac{\text{CJ} \times \text{AS}}{\left(1 - \dfrac{V_J}{\text{PB}}\right)^{\text{MJ}}} + \frac{\text{CJSWG} \times \text{PS}}{\left(1 - \dfrac{V_J}{\text{PB}}\right)^{\text{MJSWG}}} \quad \text{source junction} \tag{3.22}$$

The descriptions of the *pn* junction parameters are otherwise similar to LEVEL 1. Again, the reverse current can be input either as IS (in A) or as JS (in A/m$^2$). Whereas the first is an absolute value, the second is multiplied by AD and AS to give the reverse current of the drain and source junctions, respectively. This flexibility has been provided so that junction characteristics can either be entered as absolute values on model statement or related to areas AD and AS entered on device statements.

### 3.6.7 Source/Drain Resistance

The parasitic drain and source series resistance is expressed as either RD and RS (in Ω) or RSH (in Ω per square), the latter being multiplied by the number of squares, NRD and NRS, input on the device instance line. In addition, an alternative method is available in BSIM3v3 that allows the specification of RSDW, in units of Ω-um, and $R_D + R_S$ is computed using the width. In its basic form,

$$R_{DS} = R_D + R_S = \frac{\text{RDSW}}{W_{eff}} \tag{3.23}$$

## *3.7 Additional Effects in MOS Transistors

In this section, we describe some of the additional features and certain limitations on electrical characteristics of integrated circuit MOS devices.

### 3.7.1 Parameter Variations in Production

MOS transistors have always exhibited broad variations in major device parameters among production lots. As a result, a wide range of devices are measured and parameters are extracted to characterize the statistical variations. Particularly notable are variations in channel length, threshold voltage, and gate-oxide thickness. Additional models are added to the model library based on the extremes of the key parameters. These models are called *process corners* in that they capture parameters that would make the circuit unusually fast or unusually slow. Each corner represents different settings of the parameters that represent typical, fast, and slow situations. Designers perform SPICE simulations at a number of process corners during design. While these matters are progressively coming under better control, it is still quite common to find circuits with a 2 to 1 (or larger) unit-to-unit variation in dc power consumption and speed in a single production facility. A common response to this situation in industry is to use testing to sort the circuits according to speed and/or power consumption. Typically the fastest circuits are sold at a higher price.

### 3.7.2 Temperature Effects

MOS transistors display a temperature dependence that must be considered in circuit design. For example, an increase in temperature slows down the circuits and increases the subthreshold current. Parameters such as mobility and threshold voltage are temperature-dependent. Testing circuits at high temperature is very costly, so means are usually sought to predict high-temperature performance from room temperature tests. The process corners for SPICE simulations mentioned above are used to capture variations in operating temperature on the chip. In this mode, one temperature is specified for the entire design being analyzed. The corners may be defined as typical (room temperature), hot, and cold.

Mobility of carriers in the channel of a MOS transistor is an inverse function of absolute temperature according to the following empirical expression:

$$\mu(T) = \mu_0 \left( \frac{T}{300°\text{K}} \right)^{-1.5} \tag{3.24}$$

where $\mu_0$ is the low-field mobility, and the temperature $T$ is in kelvins (absolute). Thus, for a 100°K temperature increase, mobility may decrease by as much as 40%. Consider the current when the transistor $V_{GS}$ is above the threshold voltage. In the saturation or linear region, the drain-source current is controlled by the carrier mobility. Therefore, the drive current will decrease as we increase the temperature.

The threshold voltage of both NMOS and PMOS enhancement-mode transistors is also affected by temperature variations. The $V_T$ decreases in magnitude by 1.5 to 2 mV/°C with increasing temperature due to changes in $\phi_{GC}$ and $2\phi_F$. Usually the mobility variations are more significant for digital circuit performance.

It is also interesting to observe the effects of temperature variations on the subthreshold current. In the subthreshold region, the current is controlled by the minority carrier concentration which is dependent on $n_i$ as follows:

$$n_i = 1.45 \times 10^{10}\left(\frac{T}{300°\text{K}}\right)^{1.5} e^{(1.12/0.0516 - E_g(T)/2V_{th})} \tag{3.25}$$

where $E_g$ is the bandgap (1.12 eV for silicon). At room temperature, we know that $n_i = 1.45 \times 10^{10}/\text{cm}^3$. Therefore, the subthreshold current will increase as the temperature increases.

From the results above, it is clear that the temperature should be as low as possible for high drive current and low subthreshold current.

**Example 3.7** **SPICE Temperature Variations**

**Problem:**

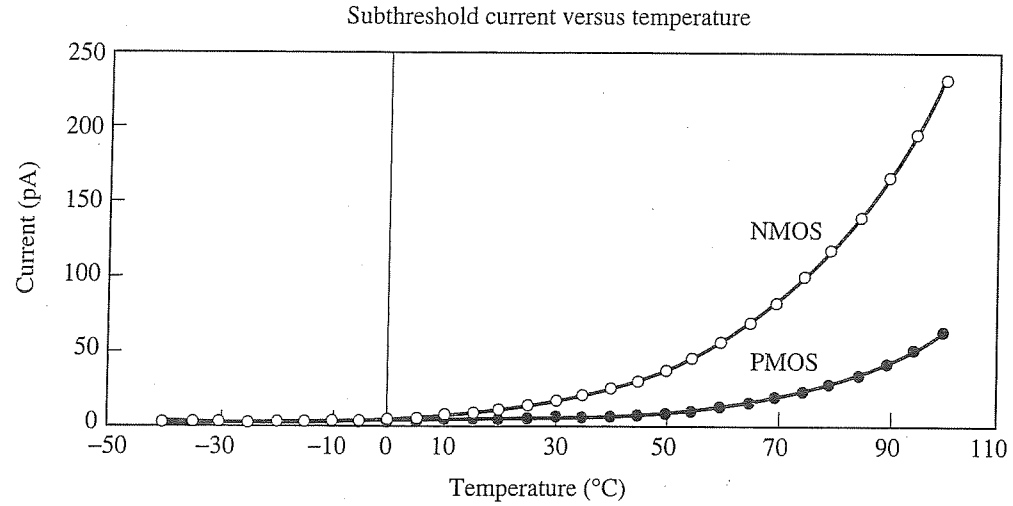
Plot $I_{DS}$ versus temperature for both PMOS and NMOS transistors by sweeping temperature from −40° C to 100° C with $V_{GS} = 0$, $V_{DD}$, and $V_{DS} = V_{DD}$. Explain the results. Use the values given in Example 3.4 for a 0.18 $\mu$m technology.

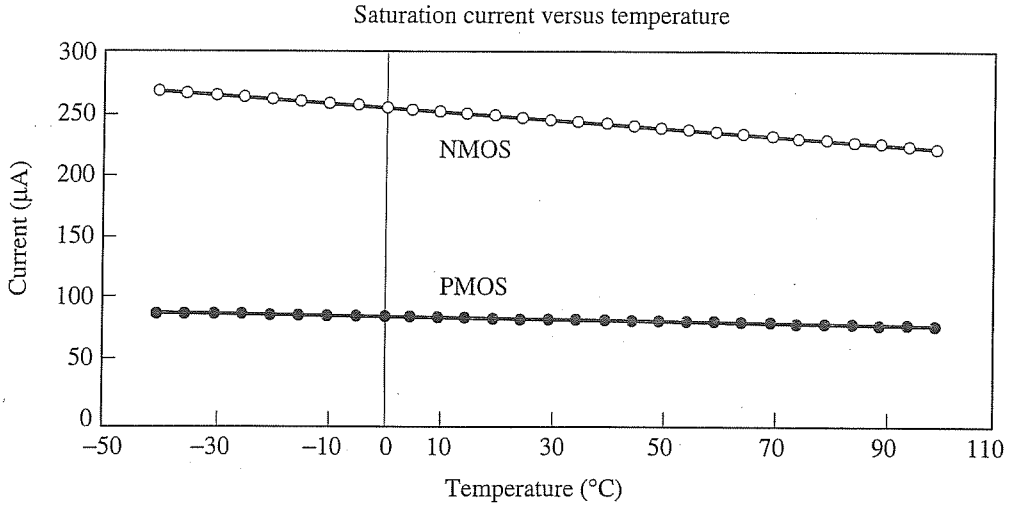
**Solution:**

The following input is used to obtain subthreshold results ($V_{GS} = 0$). A similar file can be used to obtain saturation results ($V_{GS} = V_{DD}$).

```
*SPICE Input File
.param      Supply=1.8          * Set value of Vdd
.lib        'cmos.18um'  * Set 0.18um library
.opt        scale=0.1u          * Set lambda
mp    Gnd   gatep Vdd   Vdd   PMOS l=2 w=4 ad=8 pd=4 s=8 ps=4
mn    Vdd   gaten Gnd   Gnd   NMOS l=2 w=4 ad=8 pd=4 as=8 ps=4
Vdd   Vdd   0     'Supply'
Vgsp  Vdd         gatep 0
Vgsn  gaten Gnd         0
.dc   TEMP  0     100   5
.plot dc    I1(mp)
.plot dc    I1(mn)
.end
```

The results are shown in the following figures. For the subthreshold region, the current increases as temperature increases. This is due to an increase in the minority carrier concentration, according to Equation (3.25), which controls current flow in this region of operation. In the case of the saturation region, the majority carrier current is controlled by the mobility. As temperature is increased, the mobility is reduced as indicated by Equation (3.24), and therefore the current is reduced.




### 3.7.3 Supply Variations

In addition to the process and temperature variations, there are also variations in the supply voltage from one region of the chip to another, depending on the current flow in the power grid. The voltage changes are due to resistance and inductance of the metal lines that comprise the power grid. These voltage changes are referred to as IR drops and $Ldi/dt$ variations. If the voltage on the supply drops, the circuit slows down. If the drop is significant, the circuit may not function properly. Typically, designers account for this variation using process corners. The supply is reduced by 10% uniformly across the whole design during simulation. Even though this is not accurate, the circuits that are simulated are rather small and the simulations will provide useful results for these blocks.

### 3.7.4 Voltage Limitations

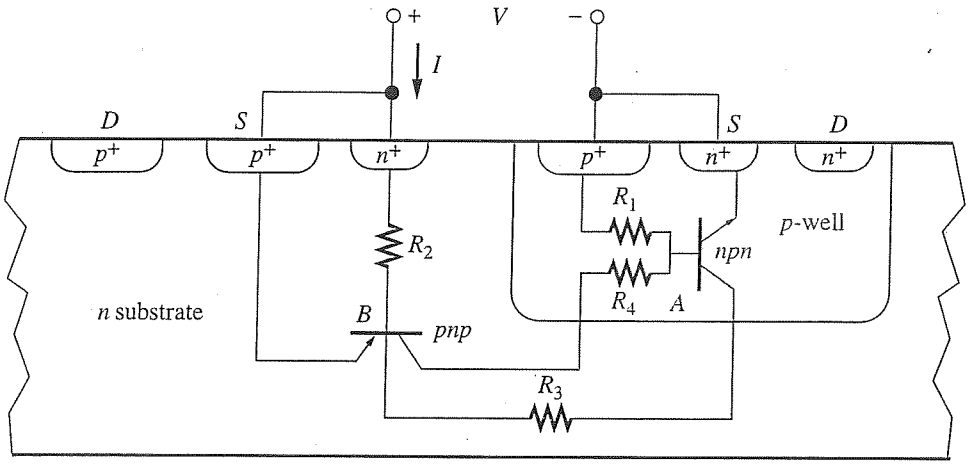
One effect that is not generally considered as a major design issue today is the *hot-carrier effect* (HCE). When $V_{GS}$ is large and $V_{DS}$ is large, carriers achieve high energy levels that can cause impact ionization near the drain end to generate hole-electron pairs. In NMOS devices, the holes are swept into the bulk to create substrate current while the "hot" electrons are injected into the oxide, thereby increasing the threshold voltage and potentially damaging the device. The effect is also seen in PMOS devices as "hot" holes. Device technology was augmented to introduce lightly doped drain and source diffusions called LDD regions. The typical source/drain doping levels are $5 \times 10^{19}$ to $10^{20}/\text{cm}^3$ whereas LDD regions are in the range of $5 \times 10^{17}/\text{cm}^3$. This reduces the field in these regions and reduces the generation of hot carriers. An additional benefit is that the *pn* junction breakdown voltage increases. However, the series resistance of the source and drain increases as a side effect. In devices today, the supply voltage is being reduced as technology scales, and the carriers no longer gain sufficient energy through collisions to surmount the potential barrier. As a result, the likelihood of hot carriers or junction breakdown is diminishing. The doping levels in these LDD regions are now in the range of $4 \times 10^{18}$ to $8 \times 10^{18}/\text{cm}^3$, effectively decreasing the series resistance.

Note that if we apply a very large $V_{DS}$ which continues to increase, eventually the source and drain depletion regions will coalesce and *punch-through* will occur. This is essentially the drain depletion region extending under the channel region to the point where it reaches the source depletion region. In this situation, the current will increase dramatically and possibly damage the device. To prevent this, the channel region is doped with additional *p*-type material to limit the extent of the depletion regions. The presence of punch-through in a device reflects a problem in the fabrication technology. It should not occur for normal operating voltages of the MOS transistor.

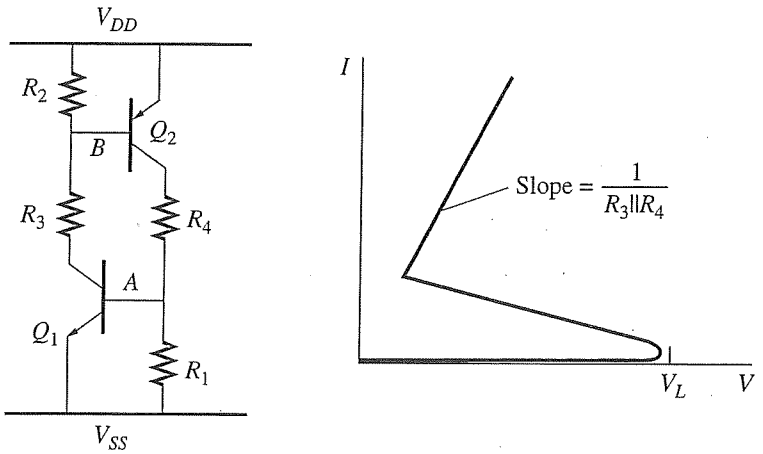
### 3.7.5 CMOS Latch-up

All MOS transistor integrated circuits have undesired and potentially troublesome parasitic bipolar transistors that will conduct if one or more *pn* junctions become forward-biased. The potential consequences of bipolar transistor actions are much more serious in CMOS circuits. Figure 3.15 illustrates that a *pnp* transistor is possible with the *n*-type body as its base, while an *npn* transistor is possible with an $n^+$ source or drain electrode as its emitter, the *p*-well as its base, and the *n*-body as a collector.

In older technologies, the undesired parasitic circuit, shown schematically in Figure 3.15a, was present. The bipolar transistors originate as just described. The resistors $R_1$ and $R_2$ (also parasitic elements in the sense that they make no contribution to the desired MOS circuit function) originate in the bulk semiconductor material of the body and the *p*-type well. The other two resistors, $R_3$ and $R_4$, play a minor role in this scenario. Consider the extracted circuit in Figure 3.15b. If any appreciable current flows through $R_1$ raising the voltage at node *A*, the base-emitter junction of the bipolar transistor $Q_1$ will become forward-biased and turn on. This will reduce the voltage at node *B*, which will cause current to flow in $R_2$. This, in




(a) Origin of parasitic elements




(b) *IV* characteristic of *pnp* – *npn* structure

**Figure 3.15**
Latch-up in CMOS circuits.

turn, will eventually turn on $Q_2$, which acts to pull node $A$ even higher. More current will flow in $Q_1$ with the appropriate current gain, and this positive feedback action will increase the currents around the loop until breakdown occurs. Clearly, low values of resistance are desirable in order to make it more difficult to forward-bias the junctions.

Another way to view the behavior of the circuit is to look at the two-terminal current-voltage characteristics of this parasitic circuit, also depicted in Figure 3.15b. This type of device is referred to as a *thyristor* or *silicon-controlled rectifier* (SCR). As the voltage across the device increases, the current initially increases slowly. Above some critical voltage, $V_L$, both bipolar transistors begin to conduct and the current rises sharply from leakage levels to a value limited by resistors $R_3$ and $R_4$, often many milliamperes. This phenomenon is known as CMOS *latch-up*. In a sense, the high current state is latched and will not change until the power is turned off or one of the devices is permanently damaged. It can occur even at normal operating voltages

if voltages applied to input or output pins cause forward-biasing of *pn* junctions within the chip.

The solution to latch-up problems is to prevent junctions from ever becoming forward-biased and to limit externally applied voltages at levels safely below $V_L$. Practical solutions to the latch-up problem involve special care in device and circuit design to reduce bipolar transistor current gain and reduce the values of $R_1$ and $R_2$. The resistance can be reduced by introducing more well contacts to create parallel resistors that have a much smaller effective resistance. The extra contacts should be placed close to the source terminal of the NMOS and PMOS devices since $R_1$ and $R_2$ are shunt resistances across the bipolar base-emitter junctions, which correspond to the source-bulk junctions of the MOS transistor.

Circuits connected to input and output pins are most critical with respect to latch-up, especially during the chip-testing process. When power is switched on and off, voltages applied to the pins of a chip frequently go outside the normal range causing latch-up. Design rules have been defined to keep the wells separated. Another approach is to use *guard rings*. If we surround the NMOS devices with a $p^+$ region connected to $V_{SS}$ and surround the PMOS devices with an $n^+$ region connected to $V_{DD}$, we can effectively isolate the devices to reduce the bipolar transistor gain, and also reduce the effective resistance. Recently, the use of STI for isolation and twin-tub CMOS in heavily doped substrates have reduced the need for additional guard rings, and made latch-up a relatively minor issue in deep submicron design.

## *3.8 Silicon-on-Insulator (SOI) Technology

A variety of special CMOS processes are under development to allow Moore's Law to continue well into the next decade. One of the more promising technologies is silicon-on-insulator (SOI), which gained popularity in the mid-1990s at a number of leading semiconductor companies.

Early attempts at SOI involved the use of silicon-on-sapphire (SOS) technology in the 1970s. The technology was effective, although costly as its name implies. Today, the insulating substrate is created by a high-energy ion-implantation of oxygen atoms well below the surface of a lightly doped silicon wafer. Then the wafer is annealed at a high temperature to produce a buried oxide layer. This process is referred to as Separation by Implanted Oxygen (SIMOX). Next, the *n*-wells and *p*-wells are formed in the substrate and separated by the STI process. Finally, the NMOS and PMOS devices are formed in the wells.

The basic structure of the CMOS-SOI technology is shown in Figure 3.16. This profile is the same as the standard bulk CMOS process except that a buried oxide has been introduced in place of the substrate. The $SiO_2$ layer is between 300–400 nm thick, while the silicon surface region is 100–200 nm thick. As a result of the insulating layer, all sides of the diffusion region have a greatly reduced capacitance. In fact, there is no longer a bottom edge junction capacitance; only a channel-edge capacitance remains relative to standard CMOS. The advantage of this structure is that the SOI circuits are much faster, dissipate less power, and are not susceptible to

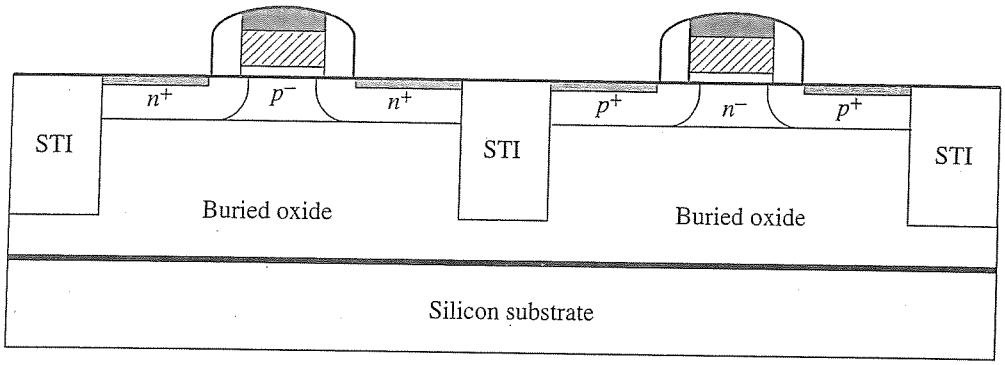



**Figure 3.16**
Structure of CMOS-SOI technology.

latch-up. This makes it a promising technology for high-speed and low-power design. Furthermore, there is reduced concern about substrate currents or the interaction of mixed-signal designs through the substrate. When technology scales below 65 nm, the doping levels required in bulk CMOS may be too high to be practical. At this point, SOI may be an attractive alternative.

Of course, like all technologies, there are a host of disadvantages or limitations. In SOI, the bulk is floating, which leads to a number of issues with the threshold voltage. The body-bias shifts around depending on the switching characteristics of the MOS device. As a result, the $V_T$ shifts dynamically and this changes the transient characteristics of the circuit. The substrate region can be tied to the appropriate supply to remove this "history-dependent" feature, but this requires additional area. Another issue is related to a *kink* that exists in the *I-V* characteristics of the device around $V_{DS} = V_{DD}/2$. The $V_T$ reduces in value and the current increases relative to the expected level and this has a noticeable effect on the transient operation. Self-heating is yet another concern since the devices are encased in oxides in all directions and it is difficult to transfer the heat generated by the devices. There is also a larger effect of DIBL on these types of devices. If these and other issues can be resolved, SOI is a strong contender for next generation CMOS technology.

Two main forms of SOI exist. The first type is the partially depleted (PD) SOI in which the depletion regions of the source/drain extend into the body but do not completely deplete all the charge in that region. In PD-SOI, the same bulk CMOS fabrication process can be used, with the addition of the SIMOX formation. It can be viewed as an evolutionary step. In fully depleted (FD) SOI, a much thinner silicon surface region and lower well doping levels are used, and the entire body of the transistor is depleted. The thickness of the silicon layer controls the threshold voltage. The development of the thinner silicon layer presents a processing challenge. However, this technology may be important to limit subthreshold current levels in sub-90 nm devices. As a result, many foundries are projecting that they will introduce PD-SOI at the 90 nm technology node, and fully depleted (FD) SOI at the 65 nm node.

## *3.9 SPICE Model Summary

**Table 3.1**
MOS LEVEL 1 parameters for SPICE

| Symbol | Name | Parameter | Units | Default | Example |
|---|---|---|---|---|---|
| | LEVEL | Model Index | | 1 | |
| $V_{T0}$ | VTO | Zero-bias threshold voltage | V | 0.0 | 1.0 |
| $k'$ | KP | Transconductance parameter | $A/V^2$ | 2.0E–5 | 3.1E–5 |
| $\gamma$ | GAMMA | Bulk threshold parameter | $V^{1/2}$ | 0.0 | 0.37 |
| $2\lvert\phi_F\rvert$ | PHI | Surface potential | V | 0.6 | 0.65 |
| $\lambda$ | LAMBDA | Channel-length modulation | $V^{-1}$ | 0.0 | 0.02 |
| $r_d$ | RD | Drain ohmic resistance | $\Omega$ | 0.0 | 1.0 |
| $r_s$ | RS | Source ohmic resistance | $\Omega$ | 0.0 | 1.0 |
| | IS | Bulk junction saturation current | A | 1.0E–14 | 1.0E–15 |
| $\phi_B$ | PB | Bulk junction potential | V | 0.8 | 0.87 |
| | CGSO | Gate-source overlap capacitance per meter channel width | F/m | 0.0 | 4.0E–11 |
| | CGDO | Gate-drain overlap capacitance per meter channel width | F/m | 0.0 | 4.0E–11 |
| | CGBO | Gate-bulk overlap capacitance per meter channel length | F/m | 0.0 | 2.0E–10 |
| | RSH | Drain and source diffusion sheet resistance | $\Omega$/ square | 0.0 | 10.0 |
| $C_{j0}$ | CJ | Zero-bias bulk junction bottom capacitance per square meter of junction area | $F/m^2$ | 0.0 | 2.0E–4 |
| *mj* | MJ | Bulk junction bottom grading coefficient | | 0.5 | 0.5 |
| | CJSW | Zero-bias bulk junction sidewall capacitance per meter of junction parameter | F/m | 0.0 | 1.0E–9 |
| *mjsw* | MJSW | Bulk junction sidewall grading coefficient | | 0.33 | |
| | JS | Bulk junction saturation current per square meter of junction area | $A/m^2$ | | 1.0E–8 |
| $t_{ox}$ | TOX | Oxide thickness | m | 1.0E–7 | 1.0E–7 |
| $N_A$ or $N_D$ | NSUB | Substrate doping | $cm^{-3}$ | 0.0 | 4.0E15 |

(*continued*)

**Table 3.1**
(Continued)

| Symbol | Name | Parameter | Units | Default | Example |
|---|---|---|---|---|---|
| $Q_{ss}/q$ | NSS | Surface state density | $cm^{-2}$ | 0.0 | 1.0E10 |
| | NFS | Fast surface state density | $cm^{-2}$ | 0.0 | 1.0E10 |
| | TPG | Type of gate material:<br>+1 opposite to substrate<br>−1 same as substrate<br>0 Al gate | | 1.0 | |
| $x_j$ | XJ | Metallurgical junction depth | m | 0.0 | 1.0E−6 |
| $L_D$ | LD | Lateral diffusion | m | 0.0 | 0.8E−6 |
| $\mu_0$ | UO | Surface mobility | $cm^2/V{\cdot}s$ | 600 | 700 |

**Table 3.2**
BSIM3v3 model parameters for SPICE

| Name | General, *W*, *L*, TOX Parameters | Units |
|---|---|---|
| LEVEL | MOSFET model selector | — |
| BINUNIT | Binning unit selector | — |
| DWB | Coefficient of the substrate bias' dependence of the width offset | $m/V^{1/2}$ |
| DWG | Coefficient of the gate-voltage dependence of the width offset | m/V |
| LINT | Channel-length offset for dc I-V characteristics without bias | m |
| LL | Coefficient of length dependence for channel-length offset in I-V calculations | m |
| LLN | Power exponent of the length dependence in the calculation of the I-V and C-V channel-length offsets | — |
| LMAX | Maximum channel length | m |
| LMIN | Minimum channel length | m |
| LW | Coefficient of width dependence in the calculation of the dc channel-length offset | m |
| LWL | Coefficient of length and width dependence in the calculation of the dc channel-length offset | m |
| LWN | Power exponent of the width dependence in the calculation of the I-V and C-V channel-length offsets | — |
| MOBMOD | Mobility model selector | — |

(*continued*)

**Table 3.2**
(Continued)

| Name | General, *W*, *L*, TOX Parameters | Units |
|---|---|---|
| NCH | Channel doping concentration | $cm^{-3}$ |
| TNOM | Device temperature | °C |
| TOX | Gate-oxide thickness | m |
| TOXM | Gate-oxide thickness at which the parameter set was extracted | m |
| VERSION | Establishes the version of the BSIM3 model to be used in simulation | — |
| WINT | Channel-width offset for dc I-V characteristics | m |
| WL | Coefficient of width dependence for channel-width offset in I-V calculation | m |
| WLN | Power exponent of the length dependence in the calculation of the I-V and C-V channel-width offsets | — |
| WMAX | Maximum channel width | m |
| WMIN | Minimum channel width | m |
| WW | Coefficient of width dependence in the calculation of the C-V channel width offset | m |
| WWL | Coefficient of length and width dependence in the calculation of the dc channel-width offset | m |
| VTH0 | Threshold voltage of long-channel device at zero $V_{BS}$ | V |
| W0 | Channel-width offset to calculate narrow width's effect on the threshold voltage | m |
| WK1 | Width sensitivity | $V^{1/2}\mu m$ |
| **Name** | **Mobility Parameters** | **Units** |
| A0 | Bulk charge effect coefficient for channel length | — |
| A1 | First nonsaturation factor | $V^{-1}$ |
| A2 | Second nonsaturation factor | $V^{-1}$ |
| AGS | Gate-bias coefficient of the body-charge coefficient, $A_{bulk}$ | $V^{-1}$ |
| B0 | Channel-width coefficient for the calculation of the body-charge coefficient, $A_{bulk}$ | m |
| B1 | Channel-width offset for the calculation of the body-charge coefficient, $A_{bulk}$ | M |
| KETA | Body-bias coefficient of the bulk charge coefficient | $V^{-1}$ |
| LU0 | Length sensitivity | $\mu m/V$ |

*(continued)*

**Table 3.2**
(Continued)

| Name | Mobility Parameters | Units |
|---|---|---|
| LUA | Binning parameter for UA | |
| LUB | U0 sensitivity to effective channel length | $cm^2\mu m/(Vsec)$ |
| LUC | Binning parameter for UC | |
| LVSAT | Binning parameter for VSAT | |
| PRWB | Body-effect coefficient of RDSW | $V^{-1/2}$ |
| PRWG | Gate-bias effect coefficient of RDSW | $V^{-1}$ |
| PU0 | Binning parameter for U0 | |
| PUB | Binning parameter for UB | |
| RDSW | Parasitic drain/source resistance per unit width | $\Omega$-$\mu$m |
| U0 | Zero-field mobility at TNOM | $cm^2$/V-sec |
| UA | First-order mobility degradation coefficient | M/V |
| UB | Second-order mobility degradation coefficient | $m^2/V^2$ |
| UC | Body-effect of mobility degradation coefficient | $V^{-1}$ |
| VSAT | Carrier saturation velocity at TNOM | m/s |
| WA0 | Binning parameter for A0 | |
| WR | Exponent of the effective device width for the calculation of RDSW | — |
| WU0 | Width sensitivity | $\mu$m/V |
| WUA | Binning parameter for UA | |

| Name | Subthreshold, Substrate, DIBL Parameters | Units |
|---|---|---|
| CDSC | Drain/Source to channel coupling capacitance | $F/m^2$ |
| CDSCB | Body-bias sensitivity of CDSC | $F/m^2$-V |
| CDSCD | Drain-bias sensitivity of CDSC | $F/m^2$-V |
| CIT | Interface state capacitance | $F/m^2$ |
| DSUB | $L_{eff}$–dependence exponent of the DIBL effects on the threshold voltage | — |
| ETA0 | Sub-threshold region drain-induced barrier-lowering (DIBL) coefficient for $V_{th}$ | — |
| ETAB | Bulk-bias coefficient of the DIBL effects | $V^{-1}$ |
| LNFACTOR | Binning parameter for NFACTOR | |
| NFACTOR | Subthreshold turn-on swing factor | — |
| PETA0 | Binning parameter for ETA0 | |
| VOFF | Offset voltage in sub-threshold region | V |
| ALPHA0 | First parameter of impact ionization current | m/V |

(*continued*)

**Table 3.2**
(Continued)

| Name | Subthreshold, Substrate, DIBL Parameters | Units |
|---|---|---|
| ALPHA1 | Substrate current parameter | $V^{-1}$ |
| BETA0 | The second parameter of the substrate current due to impact ionization | $V^{-1}$ |
| DELTA | Effective $V_{DS}$ smoothing parameter | V |
| DROUT | $L_{eff}$ dependence exponent in the DIBL correction on the Early voltage | — |
| PCLM | Channel-length modulation parameter for $I_d$ | — |
| PDIBLC1 | First coefficient of DIBL's correction on the Early voltage | — |
| PDIBLC2 | Second coefficient of DIBL's correction on the Early voltage | — |
| PDIBLCB | Body-effect coefficient to the DIBL's correction on the Early voltage | — |
| PSCBE1 | Substrate current induced body effect exponent 1 | V/m |
| PSCBE2 | Substrate current induced body effect exponent 2 | m/V |
| PVAG | Gate-bias dependence of Early voltage | — |
| **Name** | **Temperature Parameters** | **Units** |
| AT | Temperature coefficient for saturation velocity | m/sec |
| KT1 | Temperature coefficient for $V_{th}$ | V |
| KT1L | Channel-length coefficient of the threshold voltage's temperature dependence | Vm |
| KT2 | Body bias coefficient of $V_{th}$ temperature effect | Vm |
| PKT1 | Binning parameter for KT1 | |
| PRT | Temperature coefficient of RDSW | $\Omega\ \mu m^{WR}$ |
| PUA1 | Binning parameter for UA1 | |
| UA1 | Temperature coefficient for UA | m/V |
| UB1 | Temperature coefficient for UB | $m^2/V^2$ |
| UC1 | Temperature coefficient for UC | $m/V^2$ |
| UTE | Temperature coefficient for the zero-field universal mobility U0 | — |
| WUA1 | Binning parameter for UA1 | |
| **Name** | **Capacitance Parameters** | **Units** |
| CAPMOD | Capacitance model selector | — |
| CF | Fringing-field capacitance per side | F/m |

(*continued*)

**Table 3.2**
(Continued)

| Name | Capacitance Parameters | Units |
|---|---|---|
| CGDO | Voltage-independent gate-drain overlap capacitance per unit gate width | F/m |
| CGSO | Voltage-independent gate-source overlap capacitance per unit gate width | F/m |
| CJ | Source/Drain bottom junction capacitance | $F/m^2$ |
| CJSWG | Zero-bias gate-edge sidewall bulk junction capacitance | F/m |
| CJSW | Source/drain sidewall junction capacitance per unit length at the isolation sidewall, when the device temperature is equal to TNOM | F/m |
| CTA | Junction capacitance CJ temperature coefficient | 1/°K |
| CTP | Junction sidewall capacitance CJSW temperature coefficient | 1/°K |
| DLC | Effective channel-length offset for capacitance calculations | m |
| ELM | Elmore constant of the channel, used in BSIM3's non-quasi-static (NQS) model | — |
| JS | Source-bulk and drain-bulk junction saturation current per unit area when the device temperature ($T_{device}$) is equal to TNOM | $A/m^2$ |
| JSW | Sidewall bulk junction saturation current | A/m |
| MJ | Grading coefficient of the bottom-wall junction capacitance | — |
| MJSWG | Grading coefficient of the gate-edge sidewall junction capacitance | — |
| MJSW | Grading coefficient of the isolation-side sidewall junction capacitance | — |
| NQSMOD | A flag for the non-quasi-static model | — |
| PB | Built-in potential of the bottom-wall junction capacitance | V |
| PTP | Junction potential PHP temperature coefficient | V/°K |
| PTA | Junction potential PB temperature coefficient | V/°K |
| TLEVC | Temperature equation level selector for junction capacitances and potentials | — |
| XPART | Charge partition flag | — |
| XTI | Junction saturation current densities' temperature exponent | — |

## REFERENCES


1. R. C. Jaeger, *Introduction to Microelectronic Fabrication,* 2nd ed. Prentice-Hall, Upper Saddle River, NJ, 2002.
2. H. Veendrick, *Deep-Submicron CMOS ICs,* 2nd ed. Kluwer Academic Publishers, Boston, MA, 2000.
3. D. Foty, *MOSFET Modeling with SPICE: Principles and Practice,* Prentice-Hall PTR, 1997.
4. N. Arora, *MOSFET Models for VLSI Circuit Simulation,* Springer-Verlag, 1993.
5. L. W. Nagel, "SPICE2, A Computer Program to Simulate Semiconductor Circuits," *ERL Memorandum ERL-M520,* University of California, Berkeley, May 1975.
6. G. Massobrio and P. Antognetti, *Semiconductor Device Modeling with SPICE,* McGraw-Hill, 1993.
7. W. Liu, *MOSFET Models with SPICE Simulation including BSIM3v3 and BSIM4,* Wiley-Interscience, New York, 2001.
8. A. Vladimirescu, *The SPICE Book,* John Wiley and Sons, 1994.
9. HSPICE User's Manual, Synopsys Inc., Sunnyvale, CA, 1999.
10. K. Bernstein and N. Rohrer, *SOI Circuit Design Concepts,* Kluwer Academic Publishers, Boston, MA, 2000.
11. J. Plummer, M. Deal, P. Griffen, *Silicon VLSI Technology,* Prentice-Hall, Upper Saddle River, NJ, 2000.
12. J. E. Meyer, "MOS Models and Circuit Simulation," *RCA Review,* Vol. 32, March 1971, pp. 42–63.


## PROBLEMS

For the SPICE problems below, use either 0.18 $\mu$m or 0.13 $\mu$m process technology with BSIM3 model library files, depending on what is available at your site. For convenience, we use $L = 200$ nm for a 0.18 $\mu$m technology and $L = 100$ nm for a 0.13 $\mu$m process.

**P3.1.** Figure P3.1 shows a setup used to perform parameter extraction of a LEVEL 1 model on a long-channel NMOS transistor at room temperature.

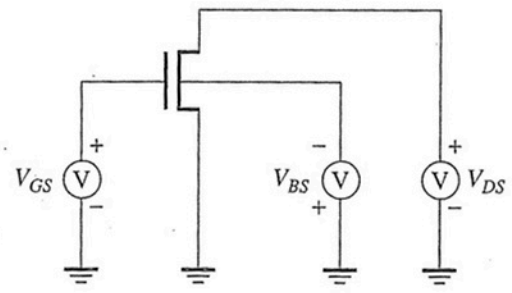



**Figure P3.1**
Experimental setup to measure I-V characteristics of an NMOS transistor.

From this setup, the current is measured for a number of different operating conditions as listed in Table P3.1.

**Table P3.1**
I-V characteristics for a long-channel NMOS transistor

| $V_{GS}$ [V] | $V_{DS}$ [V] | $V_{SB}$ [V] | $I_D$ [$\mu$A] |
|---|---|---|---|
| 1.2 | 1.2 | 0 | 1000 |
| 0.8 | 1.2 | 0 | 280 |
| 0.8 | 0.8 | 0 | 276 |
| 1.2 | 1.2 | 0.4 | 741 |

Using the data in Table P3.1, along with $W/L = 4.75$, $t_{ox} = 22$ Å and $2|\phi_F| = 0.88$ V find:

(a) the threshold voltage, $V_{T0}$

(b) the channel modulation parameter, $\lambda$

(c) the electron mobility, $\mu_n$

(d) the body effect coefficient gamma, $\gamma$

**P3.2.** Consider the long-channel MOS device shown in Figure P3.2. At $t = 0$, the drain and source are at an initial voltage of 3.3 V. At $t = 0^+$, the current source turns on with a value of 10 $\mu$A. On the graph, sketch the voltage at the source and drain nodes as a function of time until a steady-state behavior is reached. Estimate the voltages for the end points of every region of operation on the graph. Also, calculate the slope in every region. The key to this problem is to include the proper capacitances in each region of operation.

For this problem, assume that $k' = \mu_n C_{ox} = 180$ $\mu$A/V$^2$, $V_{T0} = 0.6$ V, $V_T = 1.1$ V (Assume this is the correct threshold voltage when $V_{BS}$ is not 0 V), and $W = 1.5$ $\mu$m, $L = 1$ $\mu$m (so the quadratic model can be used here). Use $t_{ox} = 100$ Å to compute the gate capacitances. Assume $C_{SB} = C_{DB} = 15$ fF are fixed junction capacitances.

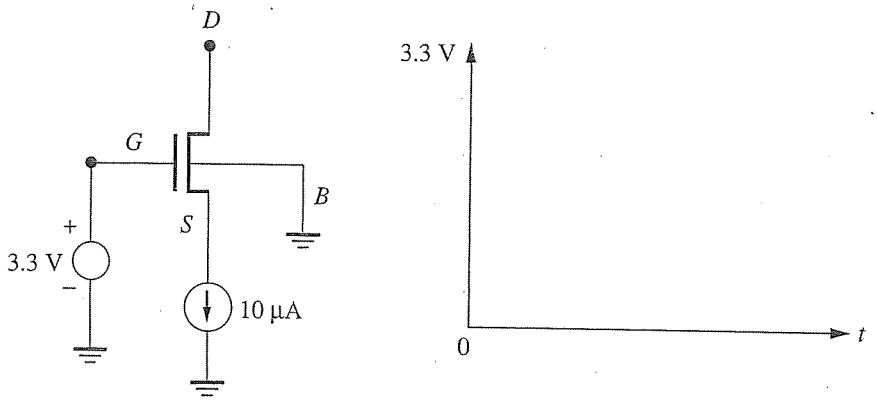



**Figure P3.2**

Check your results in SPICE. You will need to initialize the source and drain nodes to 3.3 V, and use a Level 1 MOS model to obtain the results.

**P3.3.** Figure P3.3 shows a circuit used to measure $V_{T0}$. First describe why gate and drain are connected together for the purposes of this measurement. Then using SPICE simulation, estimate $V_{T0}$ for a transistor with $L =$ 100 nm, $W = 400$ nm. You do not need to specify the capacitance information for this measurement. Why? Where should the bulk terminal be connected? Why?

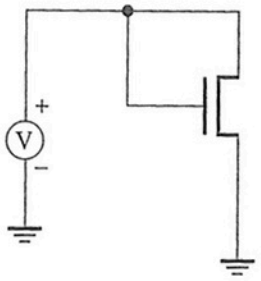



**Figure P3.3**

**P3.4.** Figure P3.4 shows a circuit that can be used to investigate the degree of channel-length modulation in short-channel transistors. Using SPICE simulation, determine whether a transistor with $L = 100$ nm, $W =$ 400 nm exhibits channel-length modulation. What is the range of voltage that should be applied to drain for the purposes of this measurement? Estimate the value of $\lambda$ from the results of simulation.

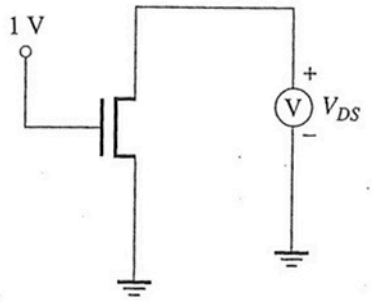



**Figure P3.4**

**P3.5.** Figure P3.5 shows a circuit used to measure the effective value of the body effect factor ($\gamma$), by measuring $V_T$ at different source voltages.

Assuming that PHI = 0.88 V, use SPICE and a 0.13 $\mu$m technology library model to find VTO and the effective value of GAMMA for a transistor with $L = 100$ nm $W = 400$ nm to be used in the simplified threshold voltage equation:

$$V_T = \text{VTO} + \text{GAMMA}(\sqrt{\text{PHI} - V_{BS}} - \sqrt{\text{PHI}})$$

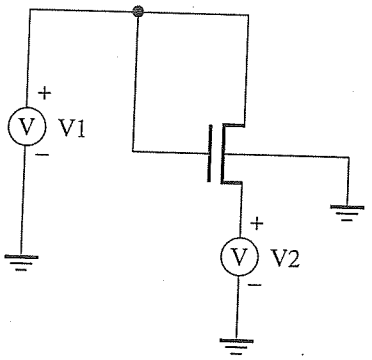



**Figure P3.5**

P3.6. Using SPICE simulation find the value of $\gamma$ for a transistor with $L = 100$ nm and $W = 400$ nm. Use the circuit shown in Figure P3.6 to measure the body effect. What is the range of the voltage that can be applied to bulk ($V_B$)? Compute the current ($I_{DS}$) with $V_B = -1$ V and $V_B = 0$ V when $V_G = 1.2$ V. What is the effect on current as $V_T$ increases?

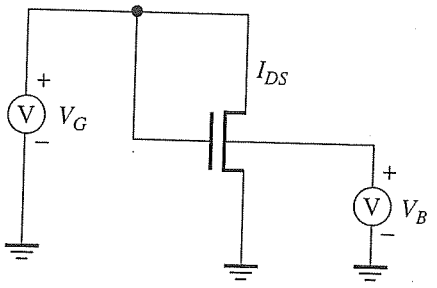



**Figure P3.6**

P3.7. Using SPICE and a 0.18 $\mu$m model, plot the subthreshold current versus $V_{BS}$, and saturation current versus $V_{BS}$ for an NMOS device with $W = 400$ nm and $L = 200$ nm. Specify the range for $V_{BS}$ as 0 to $-2.0$ V. Explain the results.

P3.8. Determine which of the following $p$-channel transistors have the highest and lowest magnitude of threshold voltage and describe why. Assume a 0.13 $\mu$m technology.

- A transistor with: $L = 0.1\ \mu$m, $W = 0.8\ \mu$m, $V_{SB} = 0.0$ V, $V_{DS} = -1.2$ V
- A transistor with: $L = 0.1\ \mu$m, $W = 0.4\ \mu$m, $V_{SB} = -0.5$ V, $V_{DS} = -1$ V
- A transistor with: $L = 0.1\ \mu$m, $W = 0.1\ \mu$m, $V_{SB} = -0.5$ V, $V_{DS} = -1$ V
- A transistor with: $L = 0.2\ \mu$m, $W = 0.4\ \mu$m, $V_{SB} = 0.0$ V, $V_{DS} = -1.2$ V
- A transistor with: $L = 0.2\ \mu$m, $W = 0.8\ \mu$m, $V_{SB} = 0.0$ V, $V_{DS} = -1.2$ V
- A transistor with: $L = 0.2\ \mu$m, $W = 0.1\ \mu$m, $V_{SB} = -0.5$ V, $V_{DS} = -1$ V

**P3.9.** Using SPICE simulation, plot log $I_{DS}$ versus $V_{GS}$ while varying $V_{DS}$ for an NMOS device with $L = 100$ nm, $W = 400$ nm and PMOS with $L = 100$ nm, $W = 1.0$ $\mu$m. Which device exhibits more DIBL? Why do PMOS transistors typically have a higher $V_T$ than NMOS transistors?

**P3.10.** This problem concerns the use of SPICE to examine effect of temperature on the $I_{on}$ and $I_{off}$ characteristics of NMOS and PMOS transistors in 0.13 $\mu$m technology, which use a 1.2 V power supply. The device sizes are $W_n = 0.8$ $\mu$m, $L_n = 0.1$ $\mu$m, $W_p = 1.6$ $\mu$m, $L_p = 0.1$ $\mu$m. Plot the following:

(a) Subthreshold region current for $V_{GS}$ from 0 to 0.4 V for temperatures $T = -60$, 25, and 125°C.

(b) Above threshold region current for $V_{GS}$ from 0.4 to 1.2 V for temperatures $T = -60$, 25, and 125°C.

(c) Explain the relationship between current and temperature for the two regions, above threshold and subthreshold, and explain the trends using equations involving temperature.

**P3.11.** What issues prompted the switch from Al to Cu? Why are low-$k$ dielectrics being developed for deep submicron processes? What is the target value for $k$ in future technologies? Why are high-$k$ dielectrics being developed?

**P3.12.** What is the difference between a self-aligned poly gate and self-aligned silicide?

**P3.13.** Compute the length of an aluminum wire and a copper wire that has a resistance of 100 Ω. You can assume that the wire thickness is 0.8 $\mu$m and the wire width is 1 $\mu$m in both cases.

**P3.14.** Assume that two metal conductors with thickness 1 $\mu$m and length 100 $\mu$m are spaced apart by a distance $S$ by silicon dioxide. Assuming that the $k$ for $SiO_2$ is roughly 4, plot the capacitance as $S$ is increased from 0.5 $\mu$m to 5 $\mu$m. On the same graph, plot the capacitance for a material with a $k = 2$. Explain both results.