CHAPTER

1

INTRODUCTION

The evolution of digital circuit design

n

Compelling issues in digital circuit design

n

How to measure the quality of digital design

n

Valuable references

1.1 A Historical Perspective

The concept of digital data manipulation has made a dramatic impact on our society. One has long grown accustomed to the idea of digital computers. Evolving steadily from mainframe and minicomputers, personal and laptop computers have proliferated into daily life. More significant, however, is a continuous trend towards digital solutions in all other areas of electronics. Instrumentation was one of the first noncomputing domains where the potential benefits of digital data manipulation over analog processing were recognized. Other areas such as control were soon to follow. Only recently have we witnessed the conversion of telecommunications and consumer electronics towards the digital format. Increasingly, telephone data is transmitted and processed digitally over both wired and wireless networks. The compact disk has revolutionized the audio world, and digital video is following in its footsteps.

The idea of implementing computational engines using an encoded data format is by no means an idea of our times. In the early nineteenth century, Babbage envisioned large-scale mechanical computing devices, called *Difference Engines* [Swade93]. Although these engines use the decimal number system rather than the binary representation now common in modern electronics, the underlying concepts are very similar. The Analytical Engine, developed in 1834, was perceived as a general-purpose computing machine, with features strikingly close to modern computers. Besides executing the basic repertoire of operations (addition, subtraction, multiplication, and division) in arbitrary sequences, the machine operated in a two-cycle sequence, called “store” and “mill” (execute), similar to current computers. It even used pipelining to speed up the execution of the addition operation! Unfortunately, the complexity and the cost of the designs made the concept impractical. For instance, the design of Difference Engine I (part of which is shown in Figure 1.1) required 25,000 mechanical parts at a total cost of £17,470 (in 1834!).

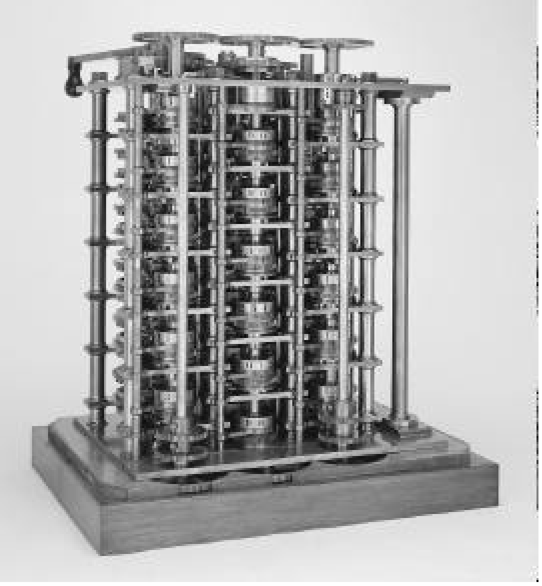


Figure 1.1 Working part of Babbage’s Difference Engine I (1832), the first known automatic calculator (from [Swade93], courtesy of the Science Museum of London).

The electrical solution turned out to be more cost effective. Early digital electronics systems were based on magnetically controlled switches (or relays). They were mainly used in the implementation of very simple logic networks. Examples of such are train safety systems, where they are still being used at present. The age of digital electronic computing only started in full with the introduction of the vacuum tube. While originally used almost exclusively for analog processing, it was realized early on that the vacuum tube was useful for digital computations as well. Soon complete computers were realized. The era of the vacuum tube based computer culminated in the design of machines such as the ENIAC (intended for computing artillery firing tables) and the UNIVAC I (the first successful commercial computer). To get an idea about *integration density*, the ENIAC was 80 feet long, 8.5 feet high and several feet wide and incorporated 18,000 vacuum tubes. It became rapidly clear, however, that this design technology had reached its limits. Reliability problems and excessive power consumption made the implementation of larger engines economically and practically infeasible.

All changed with the invention of the *transistor* at Bell Telephone Laboratories in 1947 [Bardeen48], followed by the introduction of the bipolar transistor by Schockley in 1949 [Schockley49][1]. It took till 1956 before this led to the first bipolar digital logic gate, introduced by Harris [Harris56], and even more time before this translated into a set of integrated-circuit commercial logic gates, called the Fairchild Micrologic family [Norman60]. The first truly successful IC logic family, *TTL (Transistor-Transistor Logic)* was pioneered in 1962 [Beeson62]. Other logic families were devised with higher performance in mind. Examples of these are the current switching circuits that produced the first subnanosecond digital gates and culminated in the *ECL (Emitter-Coupled Logic)* family [Masaki74], which is discussed in more detail in this textbook. TTL had the advantage, however, of offering a higher integration density and was the basis of the first integrated circuit revolution. In fact, the manufacturing of TTL components is what spear-headed the first large semiconductor companies such as Fairchild, National, and Texas Instruments. The family was so successful that it composed the largest fraction of the digital semiconductor market until the 1980s.

Ultimately, bipolar digital logic lost the battle for hegemony in the digital design world for exactly the reasons that haunted the vacuum tube approach: the large power consumption per gate puts an upper limit on the number of gates that can be reliably integrated on a single die, package, housing, or box. Although attempts were made to develop high integration density, low-power bipolar families (such as *I^2L—Integrated Injection Logic* [Hart72]), the torch was gradually passed to the MOS digital integrated circuit approach.

The basic principle behind the MOSFET transistor (originally called IGFET) was proposed in a patent by J. Lilienfeld (Canada) as early as 1925, and, independently, by O. Heil in England in 1935. Insufficient knowledge of the materials and gate stability problems, however, delayed the practical usability of the device for a long time. Once these were solved, MOS digital integrated circuits started to take off in full in the early 1970s. Remarkably, the first MOS logic gates introduced were of the CMOS variety [Wanlass63], and this trend continued till the late 1960s. The complexity of the manufac-

[1] An intriguing overview of the evolution of digital integrated circuits can be found in [Murphy93]. (Most of the data in this overview has been extracted from this reference). It is accompanied by some of the historically ground-breaking publications in the domain of digital IC's.

turing process delayed the full exploitation of these devices for two more decades. Instead, the first practical MOS integrated circuits were implemented in PMOS-only logic and were used in applications such as calculators. The second age of the digital integrated circuit revolution was inaugurated with the introduction of the first microprocessors by Intel in 1972 (the 4004) and 1974 (the 8080) [Shima74]. These processors were implemented in NMOS-only logic, that has the advantage of higher speed over the PMOS logic. Simultaneously, MOS technology enabled the realization of the first high-density semiconductor memories. For instance, the first 4Kbit MOS memory was introduced in 1970 [Hoff70].

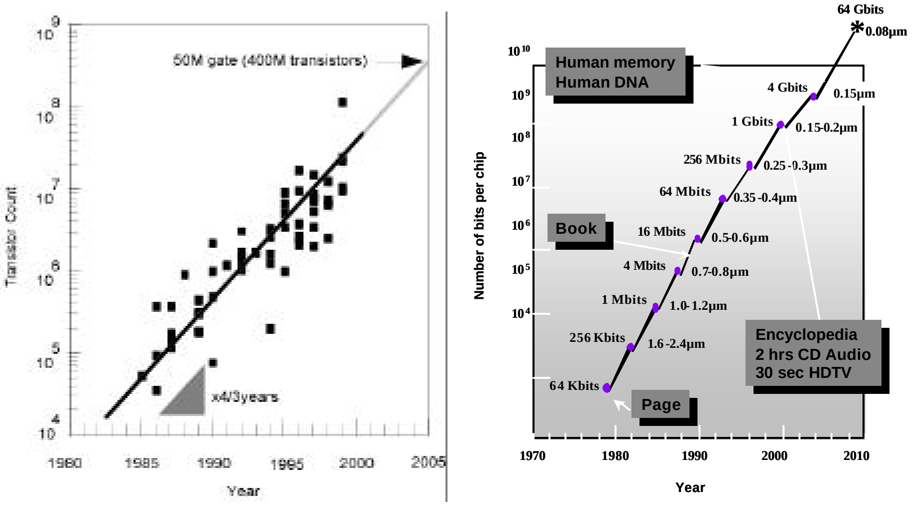
These events were at the start of a truly astounding evolution towards ever higher integration densities and speed performances, a revolution that is still in full swing right now. The road to the current levels of integration has not been without hindrances, however. In the late 1970s, NMOS-only logic started to suffer from the same plague that made high-density bipolar logic unattractive or infeasible: power consumption. This realization, combined with progress in manufacturing technology, finally tilted the balance towards the CMOS technology, and this is where we still are today. Interestingly enough, power consumption concerns are rapidly becoming dominant in CMOS design as well, and this time there does not seem to be a new technology around the corner to alleviate the problem.

Although the large majority of the current integrated circuits are implemented in the MOS technology, other technologies come into play when very high performance is at stake. An example of this is the BiCMOS technology that combines bipolar and MOS devices on the same die. BiCMOS is used in high-speed memories and gate arrays. When even higher performance is necessary, other technologies emerge besides the already mentioned bipolar silicon ECL family—Gallium-Arsenide, Silicon-Germanium and even superconducting technologies. These technologies only play a very small role in the overall digital integrated circuit design scene. With the ever increasing performance of CMOS, this role is bound to be further reduced with time. Hence the focus of this textbook on CMOS only.

1.2 Issues in Digital Integrated Circuit Design

Integration density and performance of integrated circuits have gone through an astounding revolution in the last couple of decades. In the 1960s, Gordon Moore, then with Fairchild Corporation and later cofounder of Intel, predicted that the number of transistors that can be integrated on a single die would grow exponentially with time. This prediction, later called *Moore's law*, has proven to be amazingly visionary. Its validity is best illustrated with the aid of a set of graphs. Figure 1.2 plots the integration density of both logic IC's and memory as a function of time. As can be observed, integration complexity doubles approximately every 1 to 2 years. As a result, memory density has increased by more than a thousandfold since 1970.

An intriguing case study is offered by the microprocessor. From its inception in the early seventies, the microprocessor has grown in performance and complexity at a steady and predictable pace. The number of transistors and the clock frequency for a number of landmark designs are collected in Figure 1.3. The million-transistor/chip barrier was crossed in the late eighties. Clock frequencies double every three years and have reached



(a) Trends in logic IC complexity

(b) Trends in memory complexity

Figure 1.2 Evolution of integration complexity of logic ICs and memories as a function of time .

into the GHz range. This is illustrated in Figure 1.4, which plots the microprocessor trends in terms of complexity and performance at the beginning of the 21st century. An important observation is that, as of now, these trends have not shown any signs of a slow-down.

It should be no surprise to the reader that this revolution has had a profound impact on how digital circuits are designed. Early designs were truly hand-crafted. Every transistor was laid out and optimized individually and carefully fitted into its environment. This is adequately illustrated in Figure 1.5a, which shows the design of the Intel 4004 microprocessor. This approach is, obviously, not appropriate when more than a million devices have to be created and assembled. With the rapid evolution of the design technology, time-to-market is one of the crucial factors in the ultimate success of a component.

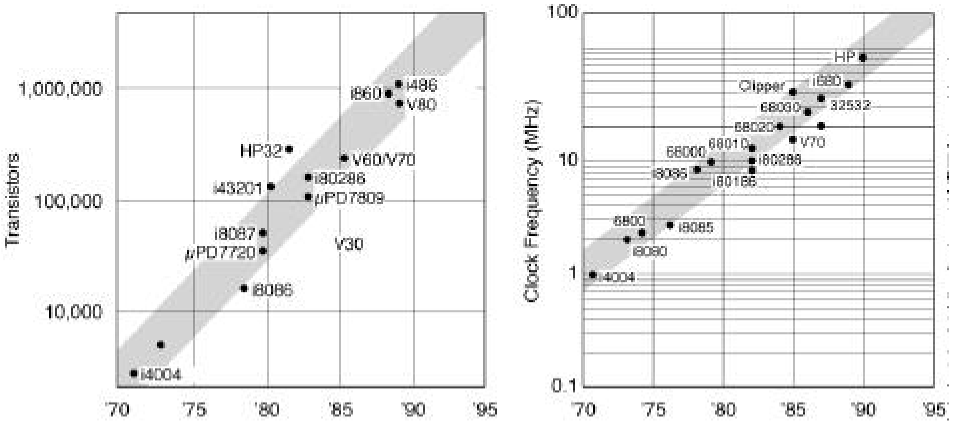
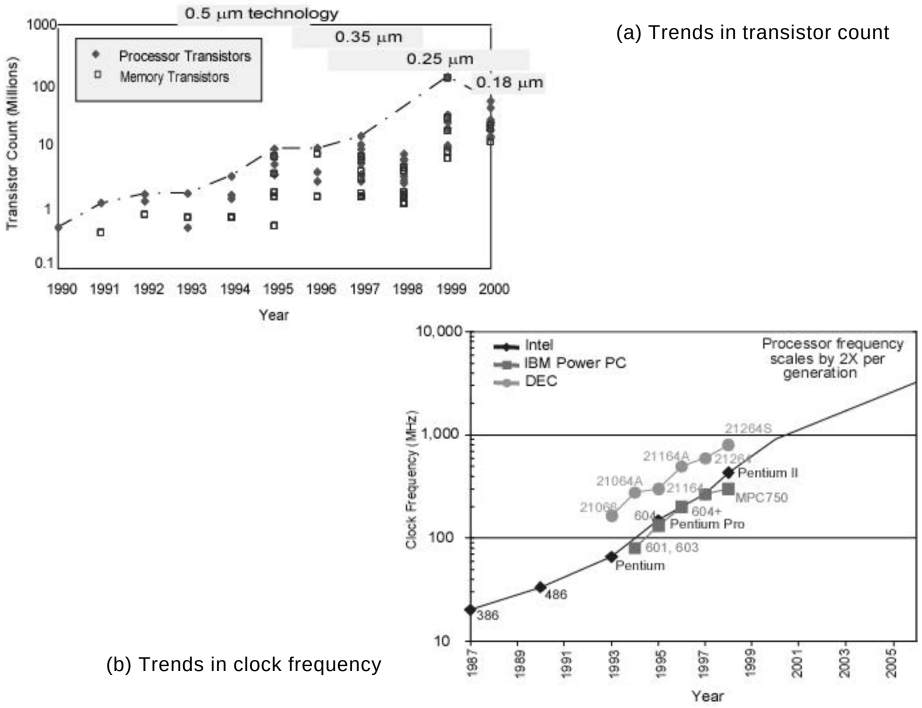


Figure 1.3 Historical evolution of microprocessor transistor count and clock frequency (from [Sasaki91]).



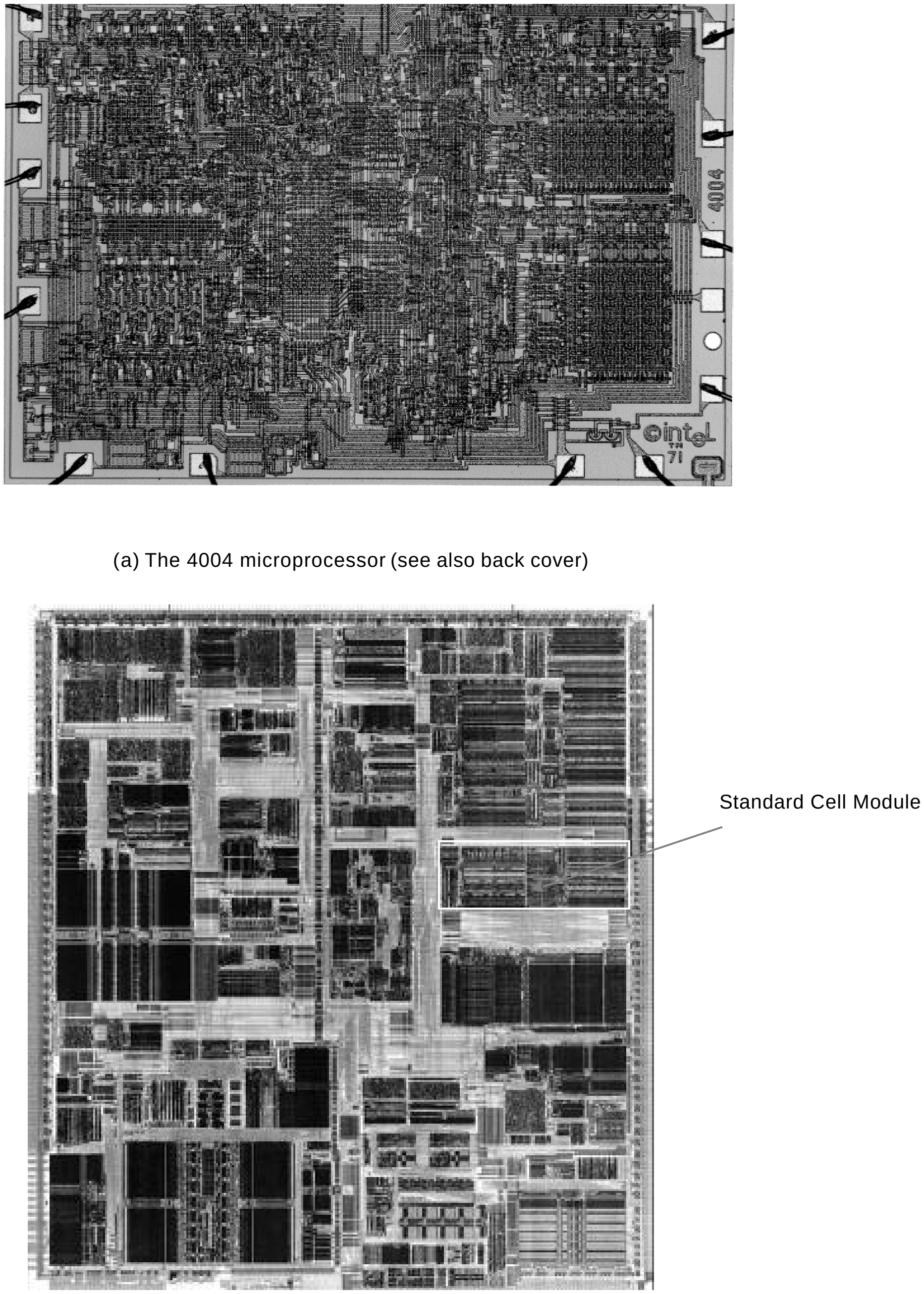
(a) Trends in transistor count

(b) Trends in clock frequency

Figure 1.4 Microprocessor trends at the benning of the 21st century. Observe howthe fraction of the transistors is being devoted to memory is increasing over time ([Young99].

Designers have, therefore, increasingly adhered to rigid design methodologies and strategies that are more amenable to design automation. The impact of this approach is apparent from the layout of one of the later Intel microprocessors, the Pentium, shown in Figure 1.5b. Instead of the individualized approach of the earlier designs, a circuit is constructed in a hierarchical way: a processor is a collection of modules, each of which consists of a number of cells on its own. Cells are reused as much as possible to reduce the design effort and to enhance the chances for a first-time-right implementation. The fact that this hierarchical approach is at all possible is the key ingredient for the success of digital circuit design and also explains why, for instance, very large scale analog design has never caught on.

The obvious next question is why such an approach is feasible in the digital world and not (or to a lesser degree) in analog designs. The crucial concept here, and the most important one in dealing with the complexity issue, is *abstraction*. At each design level, the internal details of a complex module can be abstracted away and replaced by a *black box view* or *model.* This model contains virtually all the information needed to deal with the block at the next level of hierarchy. For instance, once a designer has implemented a multiplier module, its performance can be defined very accurately and can be captured in a model. The performance of this multiplier is in general only marginally influenced by the way it is utilized in a larger system. For all purposes, it can hence be considered a black box with known characteristics. As there exists no compelling need for the system



(a) The 4004 microprocessor (see also back cover)

(b) The Pentium-II™ microprocessor (see also back cover)

Figure 1.5 Comparing the design methodologies of the Intel 4004 (1971) and Pentium-II™ (1997) microprocessors (reprinted with permission from Intel).

designer to look inside this box, design complexity is substantially reduced. The impact of this *divide and conquer* approach is dramatic. Instead of having to deal with a myriad of elements, the designer has to consider only a handful of components, each of which are characterized in performance and cost by a small number of parameters.

This is analogous to a software designer using a library of software routines such as input/output drivers. Someone writing a large program does not bother to look inside those library routines. The only thing he cares about is the intended result of calling one of those modules. Imagine what writing software programs would be like if one had to fetch every bit individually from the disk and ensure its correctness instead of relying on handy “file open” and “get string” operators.

Typically used abstraction levels in digital circuit design are, in order of increasing abstraction, the device, circuit, gate, functional module (e.g., adder) and system levels (e.g., processor), as illustrated in Figure 1.6. A semiconductor device is an entity with a very complex behavior. No circuit designer will ever seriously consider the solid-state physics equations governing the behavior of the device when designing a digital gate. Instead he will use a simplified model that adequately describes the input-output behavior of the transistor. For instance, an AND gate is adequately described by its Boolean expres-

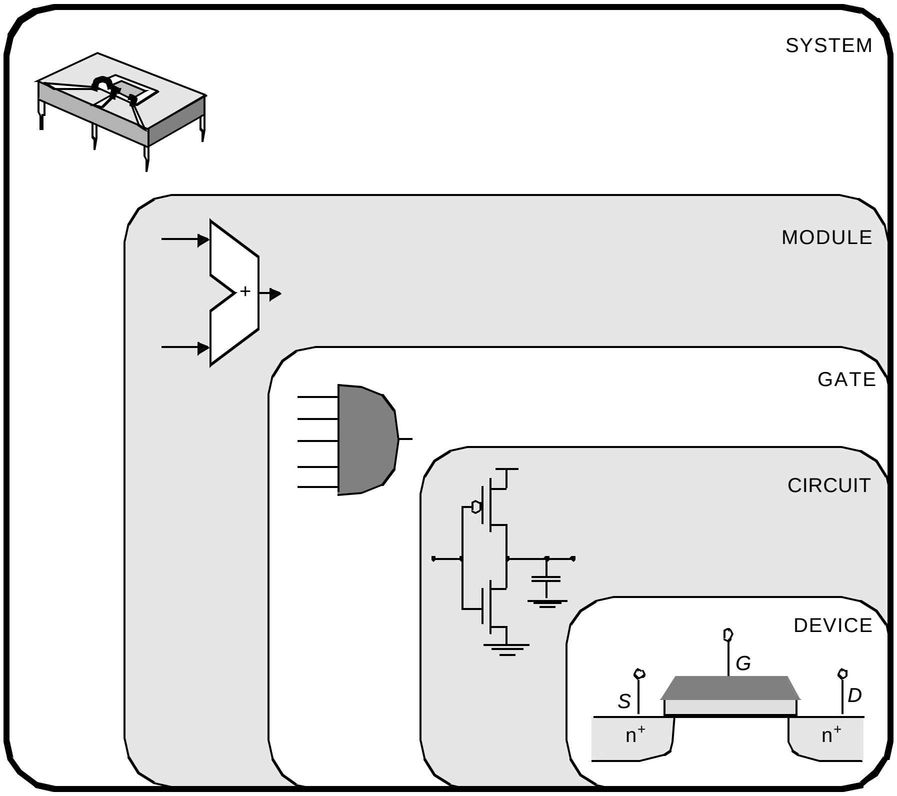


Figure 1.6 Design abstraction levels in digital circuits.

sion ($Z = A.B$), its bounding box, the position of the input and output terminals, and the delay between the inputs and the output.

This design philosophy has been the enabler for the emergence of elaborate *computer-aided design* (CAD) frameworks for digital integrated circuits; without it the current design complexity would not have been achievable. Design tools include simulation at the various complexity levels, design verification, layout generation, and design synthesis. An overview of these tools and design methodologies is given in Chapter 11 of this textbook.

Furthermore, to avoid the redesign and reverification of frequently used cells such as basic gates and arithmetic and memory modules, designers most often resort to *cell libraries*. These libraries contain not only the layouts, but also provide complete documentation and characterization of the behavior of the cells. The use of cell libraries is, for instance, apparent in the layout of the Pentium processor (Figure 1.5b). The integer and floating-point unit, just to name a few, contain large sections designed using the so-called *standard cell approach*. In this approach, logic gates are placed in rows of cells of equal height and interconnected using routing channels. The layout of such a block can be generated automatically given that a library of cells is available.

The preceding analysis demonstrates that design automation and modular design practices have effectively addressed some of the complexity issues incurred in contemporary digital design. This leads to the following pertinent question. If design automation solves all our design problems, why should we be concerned with digital circuit design at all? Will the next-generation digital designer ever have to worry about transistors or parasitics, or is the smallest design entity he will ever consider the gate and the module?

The truth is that the reality is more complex, and various reasons exist as to why an insight into digital circuits and their intricacies will still be an important asset for a long time to come.

- First of all, someone still has to *design and implement the module libraries*. Semiconductor technologies continue to advance from year to year. Until one has developed a fool-proof approach towards “porting” a cell from one technology to another, each change in technology—which happens approximately every two years—requires a redesign of the library.

- Creating an adequate *model* of a cell or module requires an in-depth understanding of its internal operation. For instance, to identify the dominant performance parameters of a given design, one has to recognize the critical timing path first.

- The library-based approach works fine when the design constraints (speed, cost or power) are not stringent. This is the case for a large number of *application-specific designs*, where the main goal is to provide a more integrated system solution, and performance requirements are easily within the capabilities of the technology. Unfortunately for a large number of other products such as microprocessors, success hinges on high performance, and designers therefore tend to push technology to its limits. At that point, the hierarchical approach tends to become somewhat less attractive. To resort to our previous analogy to software methodologies, a programmer tends to “customize” software routines when execution speed is crucial; compilers—or design tools—are not yet to the level of what human sweat or ingenuity can deliver.

- Even more important is the observation that the abstraction-based approach is only correct to a certain degree. The performance of, for instance, an adder can be substantially influenced by the way it is connected to its environment. The interconnection wires themselves contribute to delay as they introduce parasitic capacitances, resistances and even inductances. The impact of the *interconnect parasitics* is bound to increase in the years to come with the scaling of the technology.
- Scaling tends to emphasize some other deficiencies of the abstraction-based model. Some design entities tend to be *global or external* (to resort anew to the software analogy). Examples of global factors are the clock signals, used for synchronization in a digital design, and the supply lines. Increasing the size of a digital design has a profound effect on these global signals. For instance, connecting more cells to a supply line can cause a voltage drop over the wire, which, in its turn, can slow down all the connected cells. Issues such as clock distribution, circuit synchronization, and supply-voltage distribution are becoming more and more critical. Coping with them requires a profound understanding of the intricacies of digital circuit design.
- Another impact of technology evolution is that *new design issues* and constraints tend to emerge over time. A typical example of this is the periodical reemergence of power dissipation as a constraining factor, as was already illustrated in the historical overview. Another example is the changing ratio between device and interconnect parasitics. To cope with these unforeseen factors, one must at least be able to model and analyze their impact, requiring once again a profound insight into circuit topology and behavior.
- Finally, when things can go wrong, they do. A fabricated circuit does not always exhibit the exact waveforms one might expect from advance simulations. Deviations can be caused by variations in the fabrication process parameters, or by the inductance of the package, or by a badly modeled clock signal. *Troubleshooting* a design requires circuit expertise.

For all the above reasons, it is my belief that an in-depth knowledge of digital circuit design techniques and approaches is an essential asset for a digital-system designer. Even though she might not have to deal with the details of the circuit on a daily basis, the understanding will help her to cope with unexpected circumstances and to determine the dominant effects when analyzing a design.

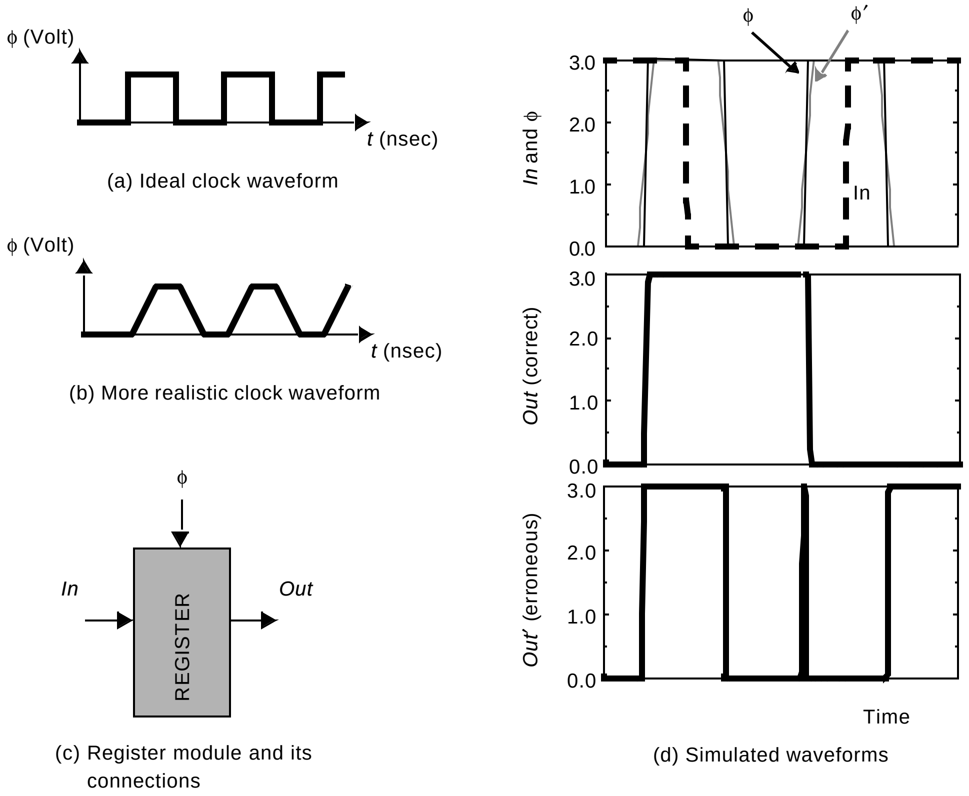
Example 1.1 Clocks Defy Hierarchy

To illustrate some of the issues raised above, let us examine the impact of deficiencies in one of the most important global signals in a design, the *clock*. The function of the clock signal in a digital design is to order the multitude of events happening in the circuit. This task can be compared to the function of a traffic light that determines which cars are allowed to move. It also makes sure that all operations are completed before the next one starts—a traffic light should be green long enough to allow a car or a pedestrian to cross the road. Under ideal circumstances, the clock signal is a periodic step waveform with abrupt transitions between the low and the high values (Figure 1.7a).

Consider, for instance, the circuit configuration of Figure 1.7c. The *register* module samples the value of the input signal at the rising edge of the clock signal ϕ. This sampled value is preserved and appears at the output until the clock rises anew and a new input is sampled. Under normal circuit operating conditions, this is exactly what happens, as demonstrated in the simulated response of Figure 1.7d. On the rising edge of clock φ, the input *In* is sampled and appears at the output *Out*.

Assume now that, due to added loading on the clock signal (for instance, connecting more latches), the clock signal is degenerated, and the clock slopes become less steep (clock φ′ in Figure 1.6d). When the degeneration is within bounds, the functionality of the latch is not impacted. When these bounds are exceeded the latch suddenly starts to malfunction as shown in Figure 1.6d (signal *Out′*). The output signal makes unexpected transitions at the falling clock edge, and extra spikes can be observed as well. Propagation of these erroneous values can cause the digital system to go into a unforeseen mode and crash. This example clearly shows how global effects, such as adding extra load to a clock, can change the behavior of an individual module. Observe that the effects shown are not universal, but are a property of the register circuit used.

Besides the requirement of steep edges, other constraints must be imposed on clock signals to ensure correct operation. A second requirement related to *clock alignment*, is illustrated in Figure 1.8. The circuit under analysis consists of two cascaded registers, both operating on the rising edge of the clock φ. Under normal operating conditions, the input *In* gets sampled into the first register on the rising edge of φ and appears at the output exactly one clock period later. This is confirmed by the simulations shown in Figure 1.7b (signal *Out*).



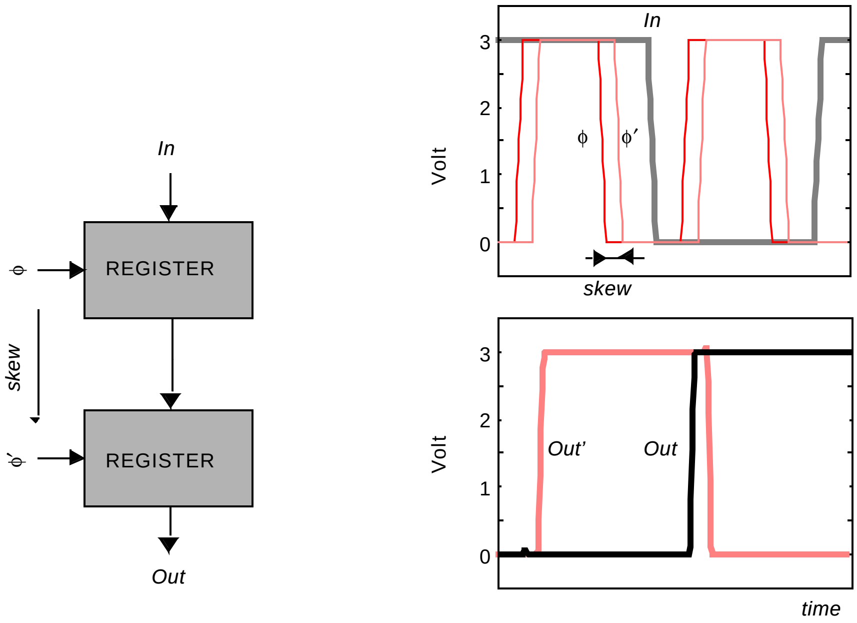
(a) Ideal clock waveform

(b) More realistic clock waveform

(c) Register module and its connections

(d) Simulated waveforms

Figure 1.7 Reduced clock slopes can cause a register circuit to fail.



(a) Two cascaded registers (b) Simulated waveforms

Figure 1.8 Impact of clock misalignment.

Due to delays associated with routing the clock wires, it may happen that the clocks become misaligned with respect to each other. As a result, the registers are interpreting time indicated by the clock signal differently. Consider the case that the clock signal for the second register is delayed—or skewed—by a value δ. The rising edge of the delayed clock ϕ' will postpone the sampling of the input of the second register. If the time it takes to propagate the output of the first register to the input of the second is smaller than the clock delay, the latter will sample the wrong value. This causes the output to change prematurely, as clearly illustrated in the simulation, where the signal *Out′* goes high at the first rising edge of ϕ' instead of the second one.

Clock misalignment, or *clock skew*, as it is normally called, is another example of how global signals may influence the functioning of a hierarchically designed system. Clock skew is actually one of the most critical design problems facing the designers of large, high-performance systems.

The purpose of this textbook is to provide *a bridge between the abstract vision of digital design and the underlying digital circuit and its peculiarities*. While starting from a solid understanding of the operation of electronic devices and an in-depth analysis of the nucleus of digital design—the inverter—we will gradually channel this knowledge into the design of more complex entities, such as complex gates, datapaths, registers, controllers, and memories. The persistent quest for a designer when designing each of the mentioned modules is to identify the dominant design parameters, to locate the section of the design he should focus his optimizations on, and to determine the specific properties that make the module under investigation (e.g., a memory) different from any others.

The text also addresses other compelling (global) issues in modern digital circuit design such as *power dissipation, interconnect, timing, and synchronization*.

1.3 Quality Metrics of A Digital Design

This section defines a set of basic properties of a digital design. These properties help to quantify the quality of a design from different perspectives: cost, functionality, robustness, performance, and energy consumption. Which one of these metrics is most important depends upon the application. For instance, pure speed is a crucial property in a compute server. On the other hand, energy consumption is a dominant metric for hand-held mobile applications such as cell phones. The introduced properties are relevant at all levels of the design hierarchy, be it system, chip, module, and gate. To ensure consistency in the definitions throughout the design hierarchy stack, we propose a bottom-up approach: we start with defining the basic quality metrics of a simple inverter, and gradually expand these to the more complex functions such as gate, module, and chip.

1.3.1 Cost of an Integrated Circuit

The total cost of any product can be separated into two components: the recurring expenses or the *variable cost*, and the non-recurring expenses or the *fixed cost*.

Fixed Cost

The fixed cost is independent of the sales volume, the number of products sold. An important component of the fixed cost of an integrated circuit is the effort in time and manpower it takes to produce the design. This design cost is strongly influenced by the complexity of the design, the aggressiveness of the specifications, and the productivity of the designer. Advanced design methodologies that automate major parts of the design process can help to boost the latter. Bringing down the design cost in the presence of an ever-increasing IC complexity is one of the major challenges that is always facing the semiconductor industry. The Design

Additionally, one has to account for the *indirect costs*, the company overhead that cannot be billed directly to one product. It includes amongst others the company's research and development (R&D), manufacturing equipment, marketing, sales, and building infrastructure.

Variable Cost

This accounts for the cost that is directly attributable to a manufactured product, and is hence proportional to the product volume. Variable costs include the costs of the parts used in the product, assembly costs, and testing costs. The total cost of an integrated circuit is now

$$\text{cost per IC} = \text{variable cost per IC} + \left(\frac{\text{fixed cost}}{\text{volume}}\right) \tag{1.1}$$

The impact of the fixed cost is more pronounced for small-volume products. This also explains why it makes sense to have large design team working for a number of years on a hugely successful product such as a microprocessor.

While the cost of producing a single transistor has dropped exponentially over the past decades, the basic variable-cost equation has not changed:

$$\text{variable cost} = \frac{\text{cost of die + cost of die test + cost of packaging}}{\text{final test yield}} \tag{1.2}$$

As will be elaborated on in Chapter 2, the IC manufacturing process groups a number of identical circuits onto a single *wafer* (Figure). Upon completion of the fabrication, the wafer is chopped into *dies*, which are then individually packaged after being *tested*. We will focus on the cost of the dies in this discussion. The cost of packaging and test is the topic of later chapters.

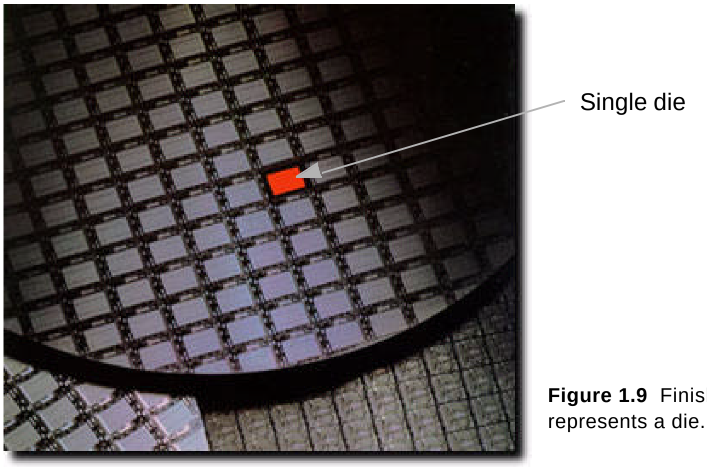


Figure 1.9 Finished wafer. Each square represents a die.

The die cost depends upon the number of good die on a wafer, and the percentage of those that are functional. The latter factor is called the *die yield*.

$$\text{cost of die} = \frac{\text{cost of wafer}}{\text{dies per wafer} \times \text{die yield}} \tag{1.3}$$

The number of dies per wafer is, in essence, the area of the wafer divided by the die area.The actual situation is somewhat more complicated as wafers are round, and chips are square. Dies around the perimeter of the wafer are therefore lost. The size of the wafer has been steadily increasing over the years, yielding more dies per fabrication run. Eq. (1.3) also presents the first indication that the cost of a circuit is dependent upon the chip area—increasing the chip area simply means that less dies fit on a wafer.

The actual relation between cost and area is more complex, and depends upon the die yield. Both the substrate material and the manufacturing process introduce faults that can cause a chip to fail. Assuming that the defects are randomly distributed over the wafer, and that the yield is inversely proportional to the complexity of the fabrication process, we obtain the following expression of the die yield:

$$\text{die yield} = \left(1 + \frac{\text{defects per unit area} \times \text{die area}}{\alpha}\right)^{-\alpha} \tag{1.4}$$

α is a parameter that depends upon the complexity of the manufacturing process, and is roughly proportional to the number of masks. $\alpha = 3$ is a good estimate for today's complex CMOS processes. The defects per unit area is a measure of the material and process induced faults. A value between 0.5 and 1 defects/cm^2 is typical these days, but depends strongly upon the maturity of the process.

Example 1.2 Die Yield

Assume a wafer size of 12 inch, a die size of 2.5 cm^2, 1 defects/cm^2, and $\alpha = 3$. Determine the die yield of this CMOS process run.

The number of dies per wafer can be estimated with the following expression, which takes into account the lost dies around the perimeter of the wafer.

$$\text{dies per wafer} = \frac{\pi \times (\text{wafer diameter}/2)^2}{\text{die area}} - \frac{\pi \times \text{wafer diameter}}{\sqrt{2 \times \text{die area}}}$$

This means 252 (= 296 - 44) potentially operational dies for this particular example. The die yield can be computed with the aid of Eq. (1.4), and equals 16%! This means that on the average only 40 of the dies will be fully functional.

The bottom line is that the number of functional of dies per wafer, and hence the cost per die is a strong function of the die area. While the yield tends to be excellent for the smaller designs, it drops rapidly once a certain threshold is exceeded. Bearing in mind the equations derived above and the typical parameter values, we can conclude that die costs are proportional to the fourth power of the area:

$$\text{cost of die} = f(\text{die area})^4 \tag{1.5}$$

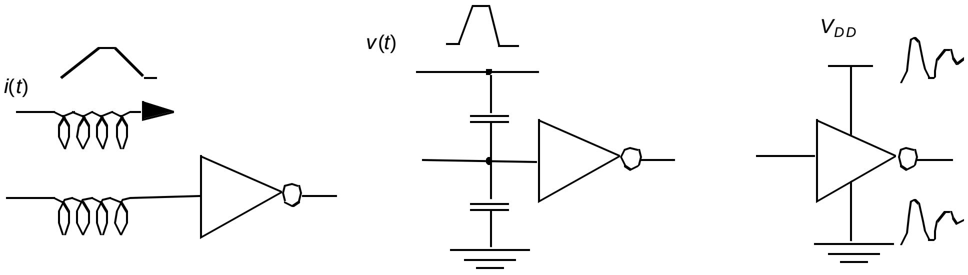
The area is a function that is directly controllable by the designer(s), and is the prime metric for cost. Small area is hence a desirable property for a digital gate. The smaller the gate, the higher the integration density and the smaller the die size. Smaller gates furthermore tend to be faster and consume less energy, as the total gate capacitance—which is one of the dominant performance parameters—often scales with the area.

The *number of transistors* in a gate is indicative for the expected implementation area. Other parameters may have an impact, though. For instance, a complex interconnect pattern between the transistors can cause the wiring area to dominate. The *gate complexity*, as expressed by the number of transistors and the regularity of the interconnect structure, also has an impact on the design cost. Complex structures are harder to implement and tend to take more of the designers valuable time. Simplicity and regularity is a precious property in cost-sensitive designs.

1.3.2 Functionality and Robustness

A prime requirement for a digital circuit is, obviously, that it performs the function it is designed for. The measured behavior of a manufactured circuit normally deviates from the expected response. One reason for this aberration are the variations in the manufacturing process. The dimensions, threshold voltages, and currents of an MOS transistor vary between runs or even on a single wafer or die. The electrical behavior of a circuit can be profoundly affected by those variations. The presence of disturbing noise sources on or off the chip is another source of deviations in circuit response. The word *noise* in the context of digital circuits means *"unwanted variations of voltages and currents at the logic nodes."* Noise signals can enter a circuit in many ways. Some examples of digital noise sources are depicted in Figure 1.10. For instance, two wires placed side by side in an integrated circuit form a coupling capacitor and a mutual inductance. Hence, a voltage or current change on one of the wires can influence the signals on the neighboring wire. Noise on the power and ground rails of a gate also influences the signal levels in the gate.

Most noise in a digital system is internally generated, and the noise value is proportional to the signal swing. Capacitive and inductive cross talk, and the internally-generated power supply noise are examples of such. Other noise sources such as input power supply noise are external to the system, and their value is not related to the signal levels. For these sources, the noise level is directly expressed in Volt or Ampere. Noise sources that are a function of the signal level are better expressed as a fraction or percentage of the signal level. Noise is a major concern in the engineering of digital circuits. How to cope with all these disturbances is one of the main challenges in the design of high-performance digital circuits and is a recurring topic in this book.



(a) Inductive coupling (b) Capacitive coupling (c) Power and ground noise

Figure 1.10 Noise sources in digital circuits.

The steady-state parameters (also called the *static behavior*) of a gate measure how robust the circuit is with respect to both variations in the manufacturing process and noise disturbances. The definition and derivation of these parameters requires a prior understanding of how digital signals are represented in the world of electronic circuits.

Digital circuits (DC) perform operations on *logical* (or *Boolean*) variables. A logical variable x can only assume two discrete values:

$$x \in \{0,1\}$$

As an example, the inversion (i.e., the function that an inverter performs) implements the following compositional relationship between two Boolean variables x and y:

$$y = \bar{x}: \{x = 0 \Rightarrow y = 1; x = 1 \Rightarrow y = 0\} \tag{1.6}$$

A logical variable is, however, a mathematical abstraction. In a physical implementation, such a variable is represented by an electrical quantity. This is most often a node voltage that is not discrete but can adopt a continuous range of values. This electrical voltage is turned into a discrete variable by associating *a nominal voltage level* with each logic state: $1 \Leftrightarrow V_{OH}$, $0 \Leftrightarrow V_{OL}$, where V_{OH} and V_{OL} represent the *high* and the *low* logic levels, respectively. Applying V_{OH} to the input of an inverter yields V_{OL} at the output and vice versa. The difference between the two is called the *logic or signal swing* V_{sw}.

$$\begin{aligned} V_{OH} &= \overline{(V_{OL})} \\ V_{OL} &= \overline{(V_{OH})} \end{aligned} \tag{1.7}$$

The Voltage-Transfer Characteristic

Assume now that a logical variable *in* serves as the input to an inverting gate that produces the variable *out*. The electrical function of a gate is best expressed by its *voltage-transfer characteristic* (VTC) (sometimes called the *DC transfer characteristic*), which plots the output voltage as a function of the input voltage $V_{out} = f(V_{in})$. An example of an inverter VTC is shown in Figure 1.11. The high and low nominal voltages, V_{OH} and V_{OL}, can readily be identified—$V_{OH} = f(V_{OL})$ and $V_{OL} = f(V_{OH})$. Another point of interest of the VTC is the *gate or switching threshold voltage* V_M (not to be confused with the threshold voltage of a transistor), that is defined as $V_M = f(V_M)$. V_M can also be found graphically at the intersection of the VTC curve and the line given by $V_{out} = V_{in}$. The gate threshold voltage presents the midpoint of the switching characteristics, which is obtained when the output of a gate is short-circuited to the input. This point will prove to be of particular interest when studying circuits with feedback (also called *sequential circuits*).

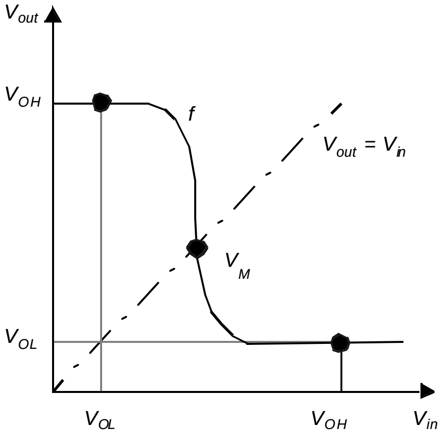


Figure 1.11 Inverter voltage-transfer characteristic.

Even if an ideal nominal value is applied at the input of a gate, the output signal often deviates from the expected nominal value. These deviations can be caused by noise or by the loading on the output of the gate (i.e., by the number of gates connected to the output signal). Figure 1.12a illustrates how a logic level is represented in reality by a range of acceptable voltages, separated by a region of uncertainty, rather than by nominal levels alone. The regions of acceptable high and low voltages are delimited by the $\boldsymbol{V_{IH}}$ and $\boldsymbol{V_{IL}}$ voltage levels, respectively. These represent by definition the points where the gain (= dV_{out} / dV_{in}) of the VTC equals −1 as shown in Figure 1.12b. The region between V_{IH} and V_{IL} is called the *undefined region* (sometimes also referred to as *transition width,* or *TW*). Steady-state signals should avoid this region if proper circuit operation is to be ensured.

Noise Margins

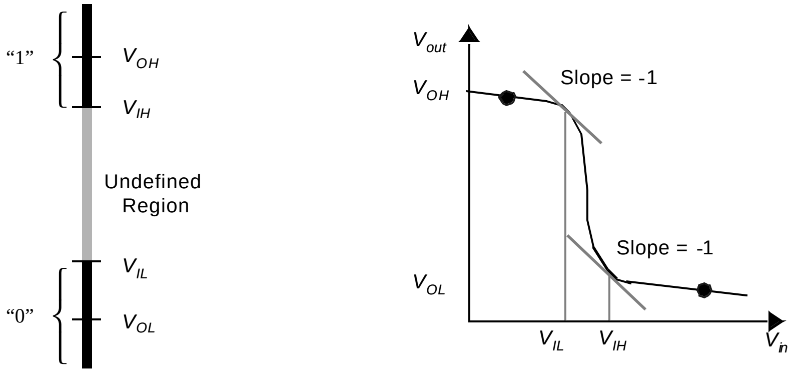
For a gate to be robust and insensitive to noise disturbances, it is essential that the “0” and “1” intervals be as large as possible. A measure of the sensitivity of a gate to noise is given by the noise margins NM_L (*noise margin low*) and NM_H (*noise margin high*), which quantize the size of the legal “0” and “1”, respectively, and set a fixed maximum threshold on the noise value:

$$\begin{aligned} NM_L &= V_{IL} - V_{OL} \\ NM_H &= V_{OH} - V_{IH} \end{aligned} \tag{1.8}$$

The noise margins represent the levels of noise that can be sustained when gates are cascaded as illustrated in Figure 1.13. It is obvious that the margins should be larger than 0 for a digital circuit to be functional and by preference should be as large as possible.

Regenerative Property

A large noise margin is a desirable, but not sufficient requirement. Assume that a signal is disturbed by noise and differs from the nominal voltage levels. As long as the signal is within the noise margins, the following gate continues to function correctly, although its output voltage varies from the nominal one. This deviation is added to the noise injected at the output node and passed to the next gate. The effect of different noise sources may accumulate and eventually force a signal level into the undefined region. This, fortunately, does not happen if the gate possesses the *regenerative property*, which ensures that a dis-



(a) Relationship between voltage and logic levels (b) Definition of V_{IH} and V_{IL}

Figure 1.12 Mapping logic levels to the voltage domain.

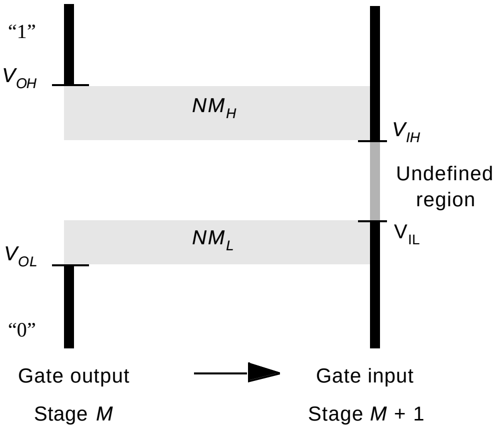
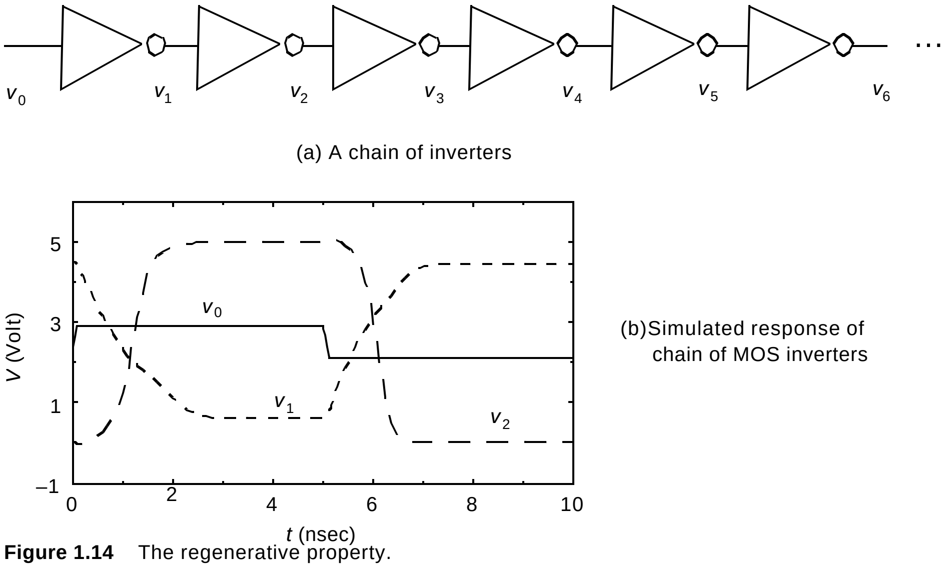


Figure 1.13 Cascaded inverter gates: definition of noise margins.

turbed signal gradually converges back to one of the nominal voltage levels after passing through a number of logical stages. This property can be understood as follows:

An input voltage v_{in} ($v_{in} \in$ "0") is applied to a chain of N inverters (Figure 1.14a). Assuming that the number of inverters in the chain is even, the output voltage v_{out} ($N \to \infty$) will equal V_{OL} if and only if the inverter possesses the regenerative property. Similarly, when an input voltage v_{in} ($v_{in} \in$ "1") is applied to the inverter chain, the output voltage will approach the nominal value V_{OH}.



(a) A chain of inverters

(b)Simulated response of chain of MOS inverters

Figure 1.14 The regenerative property.

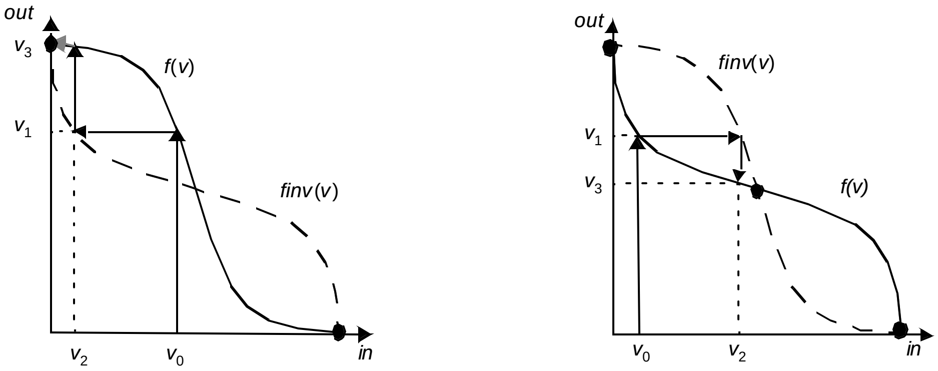
Example 1.3 Regenerative property

The concept of regeneration is illustrated in Figure 1.14b, which plots the simulated transient response of a chain of CMOS inverters. The input signal to the chain is a step-waveform with a degraded amplitude, which could be caused by noise. Instead of swinging from rail to rail,

v_0 only extends between 2.1 and 2.9 V. From the simulation, it can be observed that this deviation rapidly disappears, while progressing through the chain; v_1, for instance, extends from 0.6 V to 4.45 V. Even further, v_2 already swings between the nominal V_{OL} and V_{OH}. The inverter used in this example clearly possesses the regenerative property.

The conditions under which a gate is regenerative can be intuitively derived by analyzing a simple case study. Figure 1.15(a) plots the VTC of an inverter $V_{out} = f(V_{in})$ as well as its inverse function *finv*(), which reverts the function of the *x*- and *y*-axis and is defined as follows:

$$in = f(out) \Rightarrow in = finv(out) \tag{1.9}$$



(a) Regenerative gate

(b) Nonregenerative gate

Figure 1.15 Conditions for regeneration.

Assume that a voltage v_0, deviating from the nominal voltages, is applied to the first inverter in the chain. The output voltage of this inverter equals $v_1 = f(v_0)$ and is applied to the next inverter. Graphically this corresponds to $v_1 = finv(v_2)$. The signal voltage gradually converges to the nominal signal after a number of inverter stages, as indicated by the arrows. In Figure 1.15(b) the signal does not converge to any of the nominal voltage levels but to an intermediate voltage level. Hence, the characteristic is nonregenerative. The difference between the two cases is due to the gain characteristics of the gates. To be regenerative, the VTC should have a transient region (or undefined region) with a gain *greater than* 1 in absolute value, bordered by the two legal zones, where the gain should be *smaller than* 1. Such a gate has two stable operating points. This clarifies the definition of the V_{IH} and the V_{IL} levels that form the boundaries between the legal and the transient zones.

Noise Immunity

While the noise margin is a meaningful means for measuring the robustness of a circuit against noise, it is not sufficient. It expresses the capability of a circuit to “overpower” a noise source. *Noise immunity*, on the other hand, expresses the ability of the system to process and transmit information correctly in the presence of noise [Dally98]. Many digital circuits with low noise margins have very good noise immunity because they *reject a noise source* rather than overpower it.

To study the noise immunity of a gate, we have to construct a noise budget that allocates the power budget to the various power sources. As discussed earlier, the noise sources can be divided into sources that are proportional to the signal swing ($V_{Np}= g\ V_{sw}$), and others that are fixed (V_{Nf}). We assume, for the sake of simplicity, that the noise margin equals half the signal swing (for both H and L). To operate correctly, the noise margin has to be larger than the sum of the noise values.

$$V_{NM} = \frac{V_{sw}}{2} \geq \sum_i V_{Nfi} + \sum_j g_j V_{sw} \tag{1.10}$$

Given a set of noise sources, we can derive the minimum signal swing necessary for the system to be operational,

$$V_{sw} \geq \frac{2\sum_i V_{Nfi}}{1 - 2\sum_j g_j} \tag{1.11}$$

This makes it clear that the signal swing (and the noise margin) has to be large enough to overpower the fixed sources. On the other hand, the impact of the internal sources is strongly dependent upon the noise suppressing capabilities of the gates, i.e. the proportionality or gain factors g_j, which should be as small as possible. In later chapters, we will discuss some differential logic families that suppress most of the internal noise, and hence can get away with very small noise margins and signal swings.

Directivity

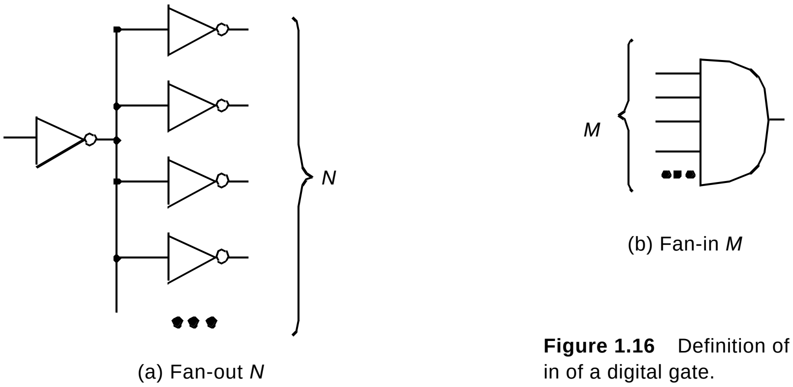
The directivity property requires a gate to be *unidirectional*, that is, changes in an output level should not appear at any unchanging input of the same circuit. If not, an output-signal transition reflects to the gate inputs as a noise signal, affecting the signal integrity.

In real gate implementations, full directivity can never be achieved. Some feedback of changes in output levels to the inputs cannot be avoided. Capacitive coupling between inputs and outputs is a typical example of such a feedback. It is important to minimize these changes so that they do not affect the logic levels of the input signals.

Fan-In and Fan-Out

The *fan-out* denotes *the number of load gates N that are connected to the output of the driving gate* (Figure 1.16). Increasing the fan-out of a gate can affect its logic output levels. From the world of analog amplifiers, we know that this effect is minimized by making the input resistance of the load gates as large as possible (minimizing the input currents) and by keeping the output resistance of the driving gate small (reducing the effects of load currents on the output voltage). When the fan-out is large, the added load can deteriorate the dynamic performance of the driving gate. For these reasons, many generic and library components define a *maximum fan-out* to guarantee that the static and dynamic performance of the element meet specification.

The *fan-in* of a gate is defined as the *number of inputs* to the gate (Figure 1.16b). Gates with large fan-in tend to be more complex, which often results in inferior static and dynamic properties.



(a) Fan-out *N*

(b) Fan-in *M*

Figure 1.16 Definition of fan-out and fan-in of a digital gate.

The Ideal Digital Gate

Based on the above observations, we can define the *ideal* digital gate from a static perspective. The ideal inverter model is important because it gives us a metric by which we can judge the quality of actual implementations.

Its VTC is shown in Figure 1.17 and has the following properties: infinite gain in the transition region, and gate threshold located in the middle of the logic swing, with high and low noise margins equal to half the swing. The input and output impedances of the ideal gate are infinity and zero, respectively (i.e., the gate has unlimited fan-out). While this ideal VTC is unfortunately impossible in real designs, some implementations, such as the static CMOS inverter, come close.

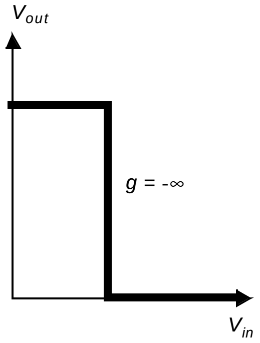


Figure 1.17 Ideal voltage-transfer characteristic.

Example 1.4 Voltage-Transfer Characteristic

Figure 1.18 shows an example of a voltage-transfer characteristic of an actual, but outdated gate structure (as produced by SPICE in the DC analysis mode). The values of the dc-parameters are derived from inspection of the graph.

$$V_{OH} = 3.5\text{ V}; \qquad V_{OL} = 0.45\text{ V}$$

$$V_{IH} = 2.35\text{ V}; \qquad V_{IL} = 0.66\text{ V}$$

$$V_M = 1.64\text{ V}$$

$$NM_H = 1.15\text{ V}; \quad NM_L = 0.21\text{ V}$$

The observed transfer characteristic, obviously, is far from ideal: it is asymmetrical, has a very low value for NM_L, and the voltage swing of 3.05V is substantially below the maximum obtainable value of 5 V (which is the value of the supply voltage for this design).

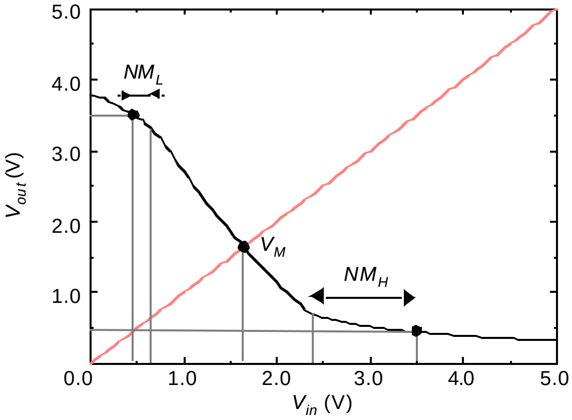


Figure 1.18 Voltage-transfer characteristic of an NMOS inverter of the 1970s.

1.3.3 Performance

From a system designers perspective, the performance of a digital circuit expresses the computational load that the circuit can manage. For instance, a microprocessor is often characterized by the number of instructions it can execute per second. This performance metric depends both on the architecture of the processor—for instance, the number of instructions it can execute in parallel—, and the actual design of logic circuitry. While the former is crucially important, it is not the focus of this text book. We refer the reader to the many excellent books on this topic [for instance, Patterson96]. When focusing on the pure design, performance is most often expressed by the duration of the clock period (*clock cycle time)*, or its rate (*clock frequency*). The minimum value of the clock period for a given technology and design is set by a number of factors such as the time it takes for the signals to propagate through the logic, the time it takes to get the data in and out of the registers, and the uncertainty of the clock arrival times. Each of these topics will be discussed in detail on the course of this text book. At the core of the whole performance analysis, however, lays the performance of an individual gate.

The *propagation delay* t_p of a gate defines how quickly it responds to a change at its input(s). It expresses *the delay experienced by a signal when passing through a gate*. It is measured between the 50% transition points of the input and output waveforms, as shown in Figure 1.19 for an inverting gate.[2] Because a gate displays different response times for rising or falling input waveforms, two definitions of the propagation delay are necessary. The t_{pLH} defines the response time of the gate for a *low to high* (or positive) output transition, while t_{pHL} refers to a *high to low* (or negative) transition. The propagation delay t_p is defined as the average of the two.

$$t_p = \frac{t_{pLH} + t_{pHL}}{2} \tag{1.12}$$

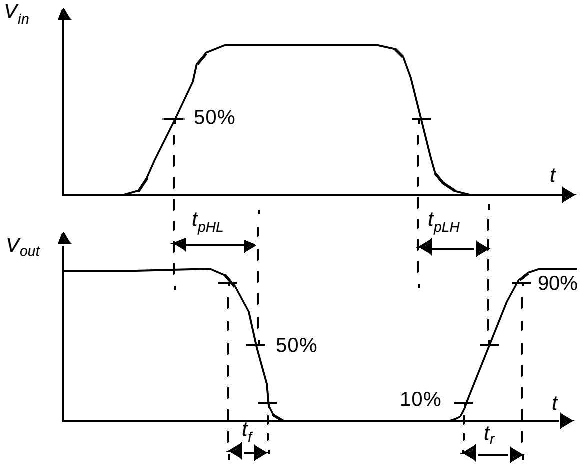


Figure 1.19 Definition of propagation delays and rise and fall times.

CAUTION: : Observe that the propagation delay t_p, in contrast to t_{pLH} and t_{pHL}, is an artificial gate quality metric, and has no physical meaning per se. It is mostly used to compare different semiconductor technologies, or logic design styles.

The propagation delay is not only a function of the circuit technology and topology, but depends upon other factors as well. Most importantly, the delay is a function of the *slopes* of the input and output signals of the gate. To quantify these properties, we introduce the *rise and fall times* t_r and t_f, which are metrics that apply to individual signal waveforms rather than gates (Figure 1.19), and express how fast a signal transits between the different levels. The uncertainty over when a transition actually starts or ends is avoided by defining the rise and fall times between the 10% and 90% points of the wave-

[2] The 50% definition is inspired the assumption that the switching threshold V_M is typically located in the middle of the logic swing.

forms, as shown in the Figure. The rise/fall time of a signal is largely determined by the strength of the driving gate, and the load presented by the node itself, which sums the contributions of the connecting gates (fan-out) and the wiring parasitics.

When comparing the performance of gates implemented in different technologies or circuit styles, it is important not to confuse the picture by including parameters such as load factors, fan-in and fan-out. A uniform way of measuring the t_p of a gate, so that technologies can be judged on an equal footing, is desirable. The de-facto standard circuit for delay measurement is the *ring oscillator*, which consists of an odd number of inverters connected in a circular chain (Figure 1.20). Due to the odd number of inversions, this circuit does not have a stable operating point and oscillates. The period T of the oscillation is determined by the propagation time of a signal transition through the complete chain, or $T = 2 \times t_p \times N$ with N the number of inverters in the chain. The factor 2 results from the observation that a full cycle requires both a low-to-high and a high-to-low transition. Note that this equation is only valid for $2Nt_p >> t_f + t_r$. If this condition is not met, the circuit might not oscillate—one "wave" of signals propagating through the ring will overlap with a successor and eventually dampen the oscillation. Typically, a ring oscillator needs a least five stages to be operational.

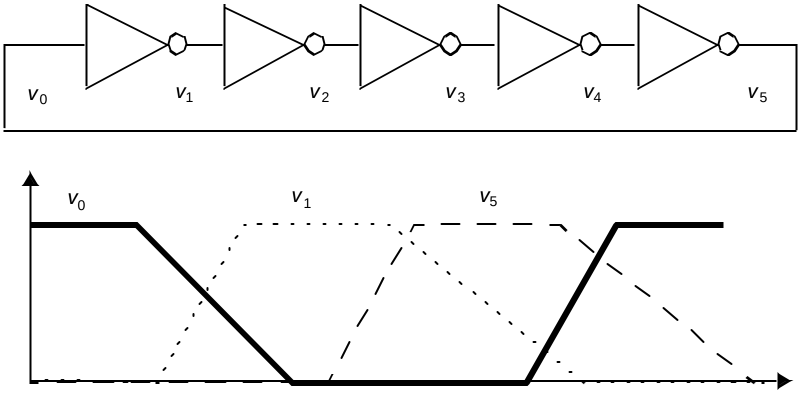


Figure 1.20 Ring oscillator circuit for propagation-delay measurement.

CAUTION: We must be extremely careful with results obtained from ring oscillator measurements. A t_p of 20 psec by no means implies that a circuit built with those gates will operate at 50 GHz. The oscillator results are primarily useful for quantifying the differences between various manufacturing technologies and gate topologies. The oscillator is an idealized circuit where each gate has a fan-in and fan-out of exactly one and parasitic loads are minimal. In more realistic digital circuits, fan-ins and fan-outs are higher, and interconnect delays are non-negligible. The gate functionality is also substantially more complex than a simple invert operation. As a result, the achievable clock frequency on average is 50 to a 100 times slower than the frequency predicted from ring oscillator measurements. This is an average observation; carefully optimized designs might approach the ideal frequency more closely.

Example 1.5 Propagation Delay of First-Order *RC* Network

Digital circuits are often modeled as first-order *RC* networks of the type shown in Figure 1.21. The propagation delay of such a network is thus of considerable interest.

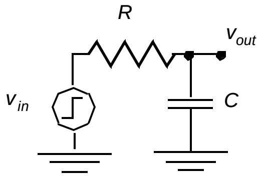


Figure 1.21 First-order *RC* network.

When applying a step input (with v_{in} going from 0 to V), the transient response of this circuit is known to be an exponential function, and is given by the following expression (where $\tau = RC$, the time constant of the network):

$$v_{out}(t) = (1 - \mathrm{e}^{-t/\tau})\,V \tag{1.13}$$

The time to reach the 50% point is easily computed as $t = \ln(2)\tau = 0.69\tau$. Similarly, it takes $t = \ln(9)\tau = 2.2\tau$ to get to the 90% point. It is worth memorizing these numbers, as they are extensively used in the rest of the text.

1.3.4 Power and Energy Consumption

The power consumption of a design determines how much energy is consumed per operation, and much heat the circuit dissipates. These factors influence a great number of critical design decisions, such as the power-supply capacity, the battery lifetime, supply-line sizing, packaging and cooling requirements. Therefore, power dissipation is an important property of a design that affects feasibility, cost, and reliability. In the world of high-performance computing, power consumption limits, dictated by the chip package and the heat removal system, determine the number of circuits that can be integrated onto a single chip, and how fast they are allowed to switch.With the increasing popularity of mobile and distributed computation, energy limitations put a firm restriction on the number of computations that can be performed given a minimum time between battery recharges.

Depending upon the design problem at hand, different dissipation measures have to be considered. For instance, the peak power P_{peak} is important when studying supply-line sizing. When addressing cooling or battery requirements, one is predominantly interested in the average power dissipation P_{av}. Both measures are defined in equation Eq. (1.14):

$$\begin{aligned} P_{peak} &= i_{peak}V_{supply} = max[p(t)] \\ P_{av} &= \frac{1}{T}\int_0^T p(t)dt = \frac{V_{supply}}{T}\int_0^T i_{supply}(t)dt \end{aligned} \tag{1.14}$$

where $p(t)$ is the instantaneous power, i_{supply} is the current being drawn from the supply voltage V_{supply} over the interval $t \in [0,T]$, and i_{peak} is the maximum value of i_{supply} over that interval.

The dissipation can further be decomposed into *static* and *dynamic* components. The latter occurs only during transients, when the gate is switching. It is attributed to the charging of capacitors and temporary current paths between the supply rails, and is, therefore, proportional to the switching frequency: *the higher the number of switching events, the higher the dynamic power consumption*. The static component on the other hand is present even when no switching occurs and is caused by static conductive paths between the supply rails or by leakage currents. It is always present, even when the circuit is in stand-by. Minimization of this consumption source is a worthwhile goal.

The propagation delay and the power consumption of a gate are related—the propagation delay is mostly determined by the speed at which a given amount of energy can be stored on the gate capacitors. The faster the energy transfer (or the higher the power consumption), the faster the gate. For a given technology and gate topology, the product of power consumption and propagation delay is generally a constant. This product is called the *power-delay product* (or PDP) and can be considered as a quality measure for a switching device. The PDP is simply the *energy* consumed by the gate *per switching event*. The ring oscillator is again the circuit of choice for measuring the PDP of a logic family.

An ideal gate is one that is fast, and consumes little energy. The *energy-delay* product (E-D) is a combined metric that brings those two elements together, and is often used as the ultimate quality metric. From the above, it should be clear that the E-D is equivalent to *power-delay*^{2}.

Example 1.6 Energy Dissipation of First-Order *RC* Network

Let us consider again the first-order *RC* network shown in Figure 1.21. When applying a step input (with V_{in} going from 0 to V), an amount of energy is provided by the signal source to the network. The total energy delivered by the source (from the start of the transition to the end) can be readily computed:

$$E_{in} = \int_0^{\infty} i_{in}(t) v_{in}(t) dt = V \int_0^{\infty} C \frac{dv_{out}}{dt} dt = (CV) \int_0^{V} dv_{out} = CV^2 \tag{1.15}$$

It is interesting to observe that the energy needed to charge a capacitor from 0 to V volt with a step input is a function of the size of the voltage step and the capacitance, but is independent of the value of the resistor. We can also compute how much of the delivered energy gets stored on the capacitor at the end of the transition.

$$E_C = \int_0^{\infty} i_C(t) v_{out}(t) dt = \int_0^{\infty} C \frac{dv_{out}}{dt} v_{out} dt = C \int_0^{V} v_{out} dv_{out} = \frac{CV^2}{2} \tag{1.16}$$

This is exactly half of the energy delivered by the source. For those who wonder happened with the other half—a simple analysis shows that an equivalent amount gets dissipated as heat in the resistor during the transaction. We leave it to the reader to demonstrate that during the discharge phase (for a step from V to 0), the energy originally stored on the capacitor gets dissipated in the resistor as well, and turned into heat.

1.4 Summary

In this introductory chapter, we learned about the history and the trends in digital circuit design. We also introduced the important quality metrics, used to evaluate the quality of a design: cost, functionality, robustness, performance, and energy/power dissipation. At the end of the Chapter, you can find an extensive list of reference works that may help you to learn more about some of the topics introduced in the course of the text.

1.5 To Probe Further

The design of digital integrated circuits has been the topic of a multitude of textbooks and monographs. To help the reader find more information on some selected topics, an extensive list of reference works is listed below. The state-of-the-art developments in the area of digital design are generally reported in technical journals or conference proceedings, the most important of which are listed.

JOURNALS AND PROCEEDINGS

IEEE Journal of Solid-State Circuits

IEICE Transactions on Electronics (Japan)

Proceedings of The International Solid-State and Circuits Conference (ISSCC)

Proceedings of the Integrated Circuits Symposium

European Solid-State Circuits Conference (ESSCIRC)

REFERENCE BOOKS

MOS

M. Annaratone, *Digital CMOS Circuit Design*, Kluwer, 1986.

T. Dillinger, *VLSI Engineering*, Prentice Hall, 1988.

E. Elmasry, ed., *Digital MOS Integrated Circuits*, IEEE Press, 1981.

E. Elmasry, ed., *Digital MOS Integrated Circuits II*, IEEE Press, 1992.

L. Glasser and D. Dopperpuhl, *The Design and Analysis of VLSI Circuits*, Addison-Wesley, 1985.

A. Kang and Leblebici, *CMOS Digital Integrated Circuits*, 2nd Ed., McGraw-Hill, 1999.

C. Mead and L. Conway, *Introduction to VLSI Systems*, Addison-Wesley, 1980.

K. Martin, *Digital Integrated Circuit Design*, Oxford University Press, 2000.

D. Pucknell and K. Eshraghian, *Basic VLSI Design*, Prentice Hall, 1988.

M. Shoji, *CMOS Digital Circuit Technology*, Prentice Hall, 1988.

J. Uyemura, *Circuit Design for CMOS VLSI*, Kluwer, 1992.

H. Veendrick, *MOS IC's: From Basics to ASICS*, VCH, 1992.

Weste and Eshraghian, *Principles of CMOS VLSI Design*, Addison-Wesley, 1985, 1993.

High-Performance Design

K. Bernstein et al, *High Speed CMOS Design Styles*, Kluwer Academic, 1998.

A. Chandrakasan, F. Fox, and W. Bowhill, ed., *Design of High-Performance Microprocessor Circuits*, IEEE Press, 2000.

M. Shoji, *High-Speed Digital Circuits*, Addison-Wesley, 1996.

Low-Power Design

A. Chandrakasan and R. Brodersen, ed., *Low-Power Digital CMOS Design*, IEEE Press, 1998.

J. Rabaey and M. Pedram, ed., *Low-Power Design Methodologies*, Kluwer Academic, 1996.

G. Yeap, *Practical Low-Power CMOS Design*, Kluwer Academic, 1998.

Memory Design

B. Prince, *Semiconductor Memories*, Wiley, 1991.

B. Prince, *High Performance Memories*, Wiley, 1996.

D. Hodges, *Semiconductor Memories*, IEEE Press, 1972.

Interconnections and Packaging

H. Bakoglu, *Circuits, Interconnections, and Packaging for VLSI*, Addison-Wesley, 1990.

W. Dally and J. Poulton, *Digital Systems Engineering*, Cambridge University Press, 1998.

E. Friedman, ed., *Clock Distribution Networks in VLSI Circuits and Systems,* IEEE Press, 1995.

J. Lau et al, ed., *Electronic Packaging: Design, Materials, Process, and Reliability*, McGraw-Hill, 1998.

Design Tools and Methodologies

V. Agrawal and S. Seth, *Test Generation for VLSI Chips*, IEEE Press, 1988.

D. Clein, *CMOS IC Layout*, Newnes, 2000.

G. De Micheli, *Synthesis and Optimization of Digital Circuits*, McGraw-Hill, 1994.

S. Rubin, *Computer Aids for VLSI Design*, Addison-Wesley, 1987.

J. Uyemura, *Physical Design of CMOS Integrated Circuits Using L-Edit*, PWS, 1995.

A. Vladimirescu, *The Spice Book*, John Wiley and Sons, 1993.

W. Wolf, *Modern VLSI Design*, Prentice Hall, 1998.

Bipolar and BiCMOS

A. Alvarez, *BiCMOS Technology and Its Applications*, Kluwer, 1989.

M. Elmasry, ed., *BiCMOS Integrated Circuit Design*, IEEE Press, 1994.

S. Embabi, A. Bellaouar, and M. Elmasry, *Digital BiCMOS Integrated Circuit Design*, Kluwer, 1993.

Lynn et al., eds., *Analysis and Design of Integrated Circuits*, McGraw-Hill, 1967.

General

J. Buchanan, *CMOS/TTL Digital Systems Design*, McGraw-Hill, 1990.

H. Haznedar, *Digital Micro-Electronics*, Benjamin/Cummings, 1991.

D. Hodges and H. Jackson, *Analysis and Design of Digital Integrated Circuits*, 2nd ed., McGraw-Hill, 1988.

M. Smith, *Application-Specific Integrated Circuits*, Addison-Wesley, 1997.

R. K. Watts, *Submicron Integrated Circuits*, Wiley, 1989.

REFERENCES

[Bardeen48] J. Bardeen and W. Brattain, “The Transistor, a Semiconductor Triode,” *Phys. Rev.*, vol.74, p. 230, July 15, 1948.

[Beeson62] R. Beeson and H. Ruegg, “New Forms of All Transistor Logic,” *ISSCC Digest of Technical Papers*, pp. 10–11, Feb. 1962.

[Harris56] J. Harris, “Direct-Coupled Transitor Logic Circuitry in Digital Computers,” *ISSCC Digest of Technical Papers*, p. 9, Feb. 1956.

[Hart72] C. Hart and M. Slob, “Integrated Injection Logic—A New Approach to LSI,” *ISSCC Digest of Technical Papers*, pp. 92–93, Feb. 1972.

[Hoff70] E. Hoff, “Silicon-Gate Dynamic MOS Crams 1,024 Bits on a Chip,” *Electronics*, pp.68–73, August 3, 1970.

[Masaki74] A. Masaki, Y. Harada and T. Chiba, “200-Gate ECL Master-Slice LSI,” *ISSCC Digest of Technical Papers*, pp. 62–63, Feb. 1974.

[Masaki92] A. Masaki, “Deep-Submicron CMOS Warms Up to High-Speed Logic,” *Circuits and Devices Magazine*, Nov. 1992.

[Murphy93] B. Murphy, “Perspectives on Logic and Microprocessors,” *Commemorative Supplement to the Digest of Technical Papers, ISSCC Conf.*, pp. 49–51, San Francisco, 1993.

[Norman60] R. Norman, J. Last and I. Haas, “Solid-State Micrologic Elements,” *ISSCC Digest of Technical Papers*, pp. 82–83, Feb. 1960.

[Sasaki91] H. Sasaki, H. Abe, T. Enomoto, and Y. Yano, “Prospect for the Chip Architecture in Sub-Halh-Micron ULSI Era,” *IEICE Transactions*, Vol. E 74, No. 1, pp. 119–129, January 1991.

[Schockley49] W. Schockley, “The Theory of pn Junctions in Semiconductors and pn-Junction Transistors,” *BSTJ*, vol. 28, p. 435, 1949.

[Schutz94] J. Schutz, “A 3.3V, 0.6 mm BiCMOS Superscaler Microprocessor,” *ISSCC Digest of Technical Papers*, pp. 202–203, Feb. 1994.

[Shima74] M. Shima, F. Faggin and S. Mazor, “An N-Channel, 8-bit Single-Chip Microprocessor,” *ISSCC Digest of Technical Papers*, pp. 56–57, Feb. 1974.

[Swade93] D. Swade, “Redeeming Charles Babbage’s Mechanical Computer,” *Scientific American*, pp.86–91, February 1993.

[Wanlass63] F. Wanlass, and C. Sah, “Nanowatt logic Using Field-Effect Metal-Oxide Semiconductor Triodes,” *ISSCC Digest of Technical Papers*, pp. 32–32, Feb. 1963.

1.6 Exercises

1. [E, None, 1.2] Based on the evolutionary trends described in the chapter, predict the integration complexity and the clock speed of a microprocessor in the year 2010. Determine also how much DRAM should be available on a single chip at that point in time, if Moore's law would still hold.
2. [D, None, 1.2] By scanning the literature, find the leading-edge devices at this point in time in the following domains: microprocessor, SRAM, and DRAM. Determine for each of those, the number of integrated devices, the overall area and the maximum clock speed. Evaluate the match with the trends predicted in section 1.2.
3. [D, None, 1.2] Find in the library the latest November issue of the *Journal of Solid State Circuits.* For each of the papers, determine its application class (such as microprocessor, signal processor, DRAM, SRAM, Gate Array), the type of manufacturing technology used (MOS, bipolar, etc.), the minimum feature size, the number of devices on a single die, and the maximum clock speed. Tabulate the results along the various application classes.
4. [E, None, 1.2] Provide at least three examples for each of the abstraction levels described in Figure 1.6.

CHAPTER 2

THE MANUFACTURING PROCESS

Overview of manufacturing process

■

Design rules

■

IC packaging

■

Future Trends in Integrated Circuit Technology

2.1 Introduction

Most digital designers will never be confronted with the details of the manufacturing process that lies at the core of the semiconductor revolution. Yet, some insight in the steps that lead to an operational silicon chip comes in quite handy in understanding the physical constraints that are imposed on a designer of an integrated circuit, as well as the impact of the fabrication process on issues such as cost.

In this chapter, we briefly describe the steps and techniques used in a modern integrated circuit manufacturing process. It is not our aim to present a detailed description of the fabrication technology, which easily deserves a complete course [Plummer00]. Rather we aim at presenting the general outline of the flow and the interaction between the various steps. We learn that a set of *optical masks* forms the central interface between the intrinsics of the manufacturing process and the design that the user wants to see transferred to the silicon fabric. The masks define the patterns that, when transcribed onto the different layers of the semiconductor material, form the elements of the electronic devices and the interconnecting wires. As such, these patterns have to adhere to some constraints in terms of minimum width and separation if the resulting circuit is to be fully functional. This collection of constraints is called the *design rule set*, and acts as the contract between the circuit designer and the process engineer. If the designer adheres to these rules, he gets a guarantee that his circuit will be manufacturable. An overview of the common design rules, encountered in modern CMOS processes, will be given. Finally, an overview is given of the *IC packaging* options. The package forms the interface between the circuit implemented on the silicon die and the outside world, and as such has a major impact on the performance, reliability, longevity, and cost of the integrated circuit.

2.2 Manufacturing CMOS Integrated Circuits

A simplified cross section of a typical CMOS inverter is shown in Figure 2.1. The CMOS process requires that both *n*-channel (NMOS) and *p*-channel (PMOS) transistors be built in the same silicon material. To accommodate both types of devices, special regions called *wells* must be created in which the semiconductor material is opposite to the type of the channel. A PMOS transistor has to be created in either an *n*-type substrate or an *n*-well, while an NMOS device resides in either a *p*-type substrate or a *p*-well. The cross section

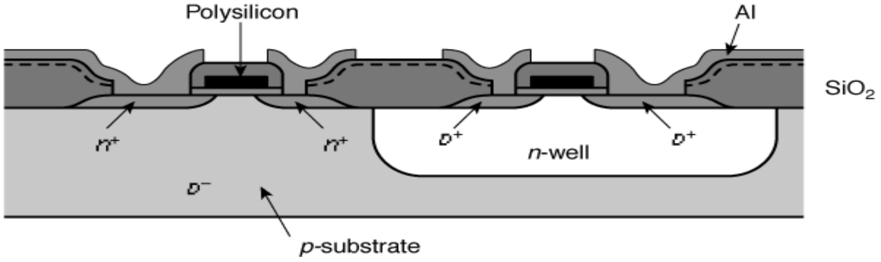


Figure 2.1 Cross section of an *n*-well CMOS process.

shown in Figure 2.1 features an *n*-well CMOS process, where the NMOS transistors are implemented in the *p*-doped substrate, and the PMOS devices are located in the *n*-well. Modern processes are increasingly using a *dual-well* approach that uses both *n*- and *p*-wells, grown on top on a epitaxial layer, as shown in Figure 2.2. We will restrict the remainder of this discussion to the latter process (without loss of generality).

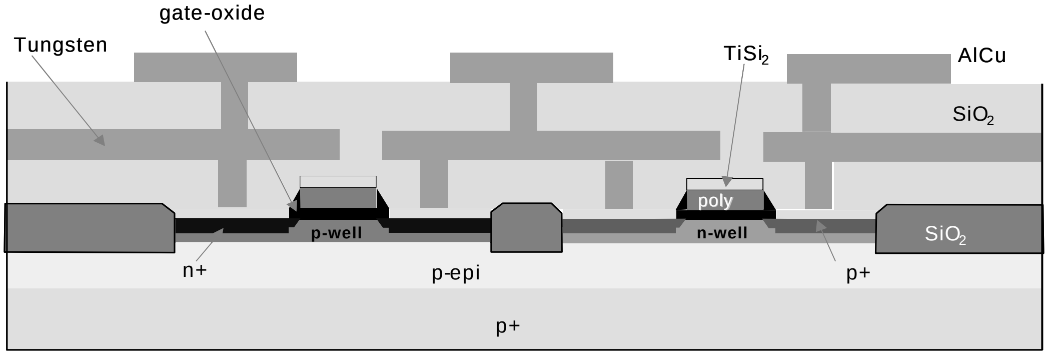


Figure 2.2 Cross section of modern dual-well CMOS process.

The CMOS process requires a large number of steps, each of which consists of a sequence of basic operations. A number of these steps and/or operations are executed very repetitively in the course of the manufacturing process. Rather than diving directly into a description of the overall process flow, we first discuss the starting material followed by a detailed perspective on some of the most-often recurring operations.

2.2.1 The Silicon Wafer

The base material for the manufacturing process comes in the form of a single-crystalline, lightly doped *wafer*. These wafers have typical diameters between 4 and 12 inches (10 and 30 cm, respectively) and a thickness of at most 1 mm, and are obtained by cutting a single-crystal ingot into thin slices (Figure 2.3). A starting wafer of the p^--type might be doped around the levels of 2×10^{21} impurities/m^3. Often, the surface of the wafer is doped more heavily, and a single crystal *epitaxial layer* of the opposite type is grown over the surface before the wafers are handed to the processing company. One important metric is the defect density of the base material. High defect densities lead to a larger fraction of non-functional circuits, and consequently an increase in cost of the final product.

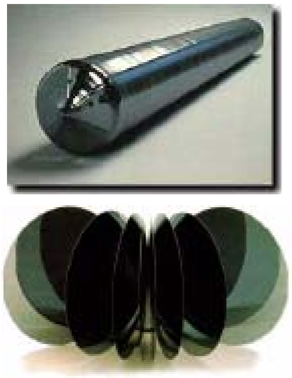


Figure 2.3 Single-crystal ingot and sliced wafers (from [Fullman99]).

2.2.2 Photolithography

In each processing step, a certain area on the chip is masked out using the appropriate optical mask so that a desired processing step can be selectively applied to the remaining regions. The processing step can be any of a wide range of tasks including oxidation, etching, metal and polysilicon deposition, and ion implantation. The technique to accomplish this selective masking, called *photolithography*, is applied throughout the manufacturing process. Figure 2.4 gives a graphical overview of the different operations involved in a typical photolitographic process. The following steps can be identified:

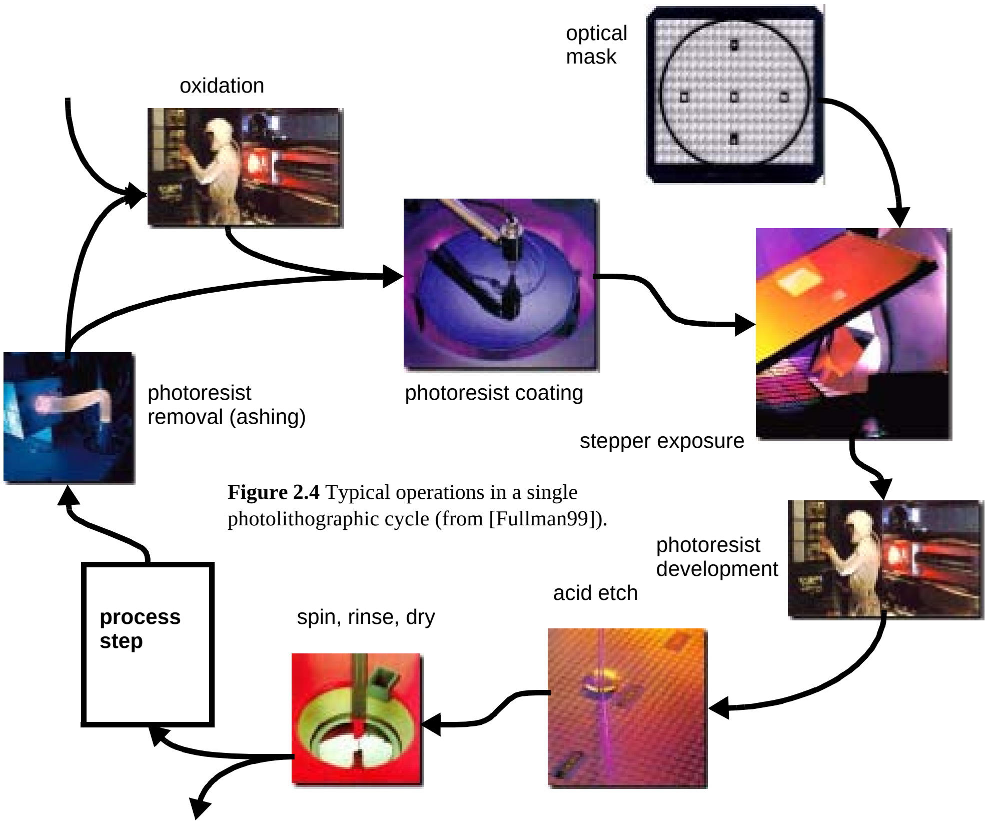


Figure 2.4 Typical operations in a single photolithographic cycle (from [Fullman99]).

1. *Oxidation layering* — this optional step deposits a thin layer of SiO_2 over the complete wafer by exposing it to a mixture of high-purity oxygen and hydrogen at approximately 1000°C. The oxide is used as an insulation layer and also forms transistor gates.
2. *Photoresist coating* — a light-sensitive polymer (similar to latex) is evenly applied while spinning the wafer to a thickness of approximately 1 μm. This material is originally soluble in an organic solvent, but has the property that the polymers cross-

link when exposed to light, making the affected regions insoluble. A photoresist of this type is called *negative*. A positive photoresist has the opposite properties; originally insoluble, but soluble after exposure. By using both positive and negative resists, a single mask can sometimes be used for two steps, making complementary regions available for processing. Since the cost of a mask is increasing quite rapidly with the scaling of technology, a reduction of the number of masks is surely of high priority.

3. *Stepper exposure* — a glass mask (or reticle), containing the patterns that we want to transfer to the silicon, is brought in close proximity to the wafer. The mask is opaque in the regions that we want to process, and transparent in the others (assuming a negative photoresist). The glass mask can be thought of as the negative of one layer of the microcircuit. The combination of mask and wafer is now exposed to ultra-violet light. Where the mask is transparent, the photoresist becomes insoluble.

4. *Photoresist development and bake* — the wafers are developed in either an acid or base solution to remove the non-exposed areas of photoresist. Once the exposed photoresist is removed, the wafer is “soft-baked” at a low temperature to harden the remaining photoresist.

5. *Acid Etching* — material is selectively removed from areas of the wafer that are not covered by photoresist. This is accomplished through the use of many different types of acid, base and caustic solutions as a function of the material that is to be removed. Much of the work with chemicals takes place at large wet benches where special solutions are prepared for specific tasks. Because of the dangerous nature of some of these solvents, safety and environmental impact is a primary concern.

6. *Spin, rinse, and dry* — a special tool (called SRD) cleans the wafer with deionized water and dries it with nitrogen. The microscopic scale of modern semiconductor devices means that even the smallest particle of dust or dirt can destroy the circuitry. To prevent this from happening, the processing steps are performed in ultra-clean rooms where the number of dust particles per cubic foot of air ranges between 1 and 10. Automatic wafer handling and robotics are used whenever possible. This explains why the cost of a state-of-the-art fabrication facility easily ranges in the multiple billions of dollars. Even then, the wafers must be constantly cleaned to avoid contamination, and to remove the left-over of the previous process steps.

7. *Various process steps* — the exposed area can now be subjected to a wide range of process steps, such as ion implantation, plasma etching, or metal deposition. These are the subjects of the subsequent section.

8. *Photoresist removal (or ashing)* — a high-temperature plasma is used to selectively remove the remaining photoresist without damaging device layers.

We illustrate the use of the photolitographic process for one specific example, the patterning of a layer of SiO_2, in Figure 2.5. The sequence of process steps shown in the Figure patterns exactly one layer of the semiconductor material, and may seem very complex. Yet, the reader has to bear in mind that same sequence patterns the layer of **the complete surface of the wafer**. It is hence a very parallel process, transferring hundreds of

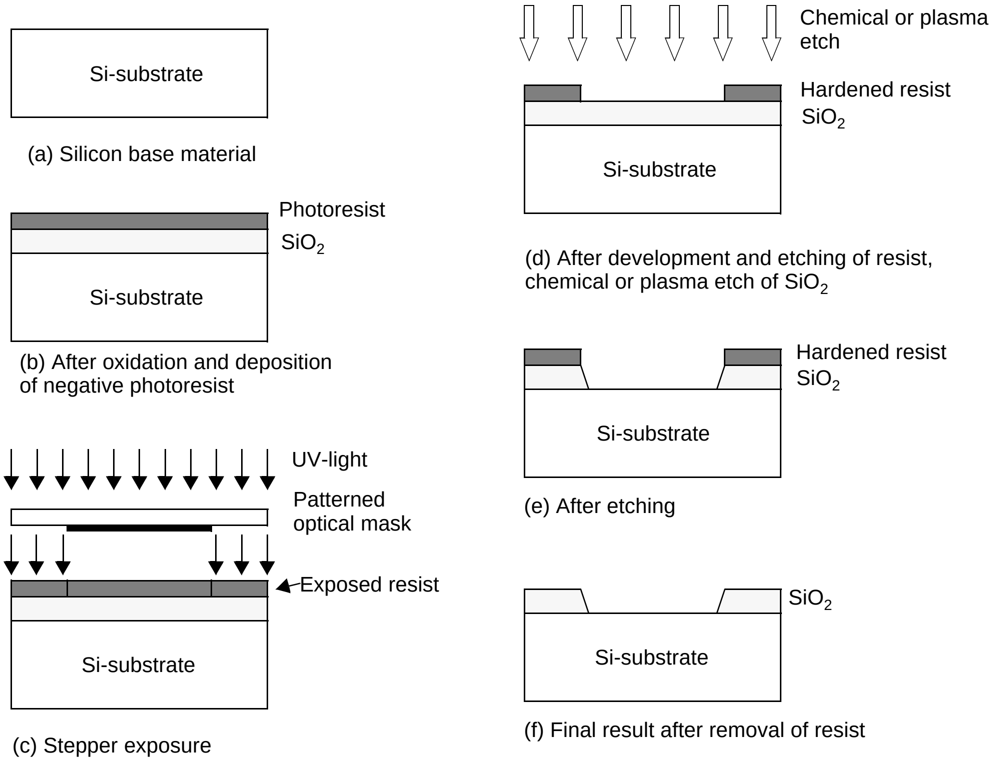


Figure 2.5 Process steps for patterning of SiO_2.

millions of patterns to the semiconductor surface simultaneously. The concurrent and scalable nature of the optolithographical process is what makes the cheap manufacturing of complex semiconductor circuits possible, and lies at the core of the economic success of the semiconductor industry.

The continued scaling of the minimum feature sizes in integrated circuits puts an enormous burden on the developer of semiconductor manufacturing equipment. This is especially true for the optolithographical process. The dimensions of the features to be transcribed surpass the wavelengths of the optical light sources, so that achieving the necessary resolution and accuracy becomes harder and harder. So far, engineering engineering has extended the lifetime of this process at least until the 100 nm (or 0.1 μm) process generation. Techniques such as optical-mask correction (OPC) pre-warp the drawn patterns to account for the diffraction phenomena, encountered when printing close to the limits of optical lithography. This adds substantially to the cost of mask making. In the foreseeable future, other solutions that offer a a finer resolution such as extreme-ultraviolet (EUV), X-ray or electron-beam may be needed. These techniques, while fully functional, are currently less attractive from an economic viewpoint.

2.2.3 Some Recurring Process Steps

Diffusion and Ion Implantation

Many steps of the integrated circuit manufacturing process require a chance in the dopant concentration of some parts of the material. The creation of the source and drain regions, well and substrate contacts, the doping of the polysilicon, and the adjustments of the device threshold are examples of such. There exist two approaches for introducing these dopants—diffusion and ion implantation. In both techniques, the area to be doped is exposed, while the rest of the wafer is coated with a layer of buffer material, typically SiO_2.

In *diffusion implantation*, the wafers are placed in a quartz tube embedded in a heated furnace. A gas containing the dopant is introduced in the tube. The high temperatures of the furnace, typically 900 to 1100 °C, cause the dopants to diffuse into the exposed surface both vertically and horizontally. The final dopant concentration is the greatest at the surface and decreases in a gaussian profile deeper in the material.

In *ion implantation*, dopants are introduced as ions into the material. The ion implantation system directs and sweeps a beam of purified ions over the semiconductor surface. The acceleration of the ions determines how deep they will penetrate the material, while the beam current and the exposure time determine the dosage. The ion implantation method allows for an independent control of depth and dosage. This is the reason that ion implantation has largely displaced diffusion in modern semiconductor manufacturing.

Ion implantation has some unfortunate side effects however, the most important one being lattice damage. Nuclear collisions during the high energy implantation cause the displacement of substrate atoms, leading to material defects. This problem is largely resolved by applying a subsequent *annealing* step, in which the wafer is heated to around 1000°C for 15 to 30 minutes, and then allowed to cool slowly. The heating step thermally vibrates the atoms, which allows the bonds to reform.

Deposition

Any CMOS process requires the repetitive deposition of layers of a material over the complete wafer, to either act as buffers for a processing step, or as insulating or conducting layers. We have already discussed the oxidation process, which allows a layer of SiO_2 to be grown. Other materials require different techniques. For instance, silicon nitride (Si_3N_4) is used as a sacrificial buffer material during the formation of the field oxide and the introduction of the stopper implants. This silicon nitride is deposited everywhere using a process called *chemical vapor deposition* or CVD, which uses a gas-phase reaction with energy supplied by heat at around 850°C.

Polysilicon, on the other hand, is deposited using a chemical deposition process, which flows silane gas over the heated wafer coated with SiO_2 at a temperature of approximately 650°C. The resulting reaction produces a non-crystalline or amorphous material called polysilicon. To increase to conductivity of the material, the deposition has to be followed by an implantation step.

The Aluminum interconnect layers are typically deployed using a process known as *sputtering*. The aluminum is evaporated in a vacuum, with the heat for the evaporation delivered by electron-beam or ion-beam bombarding. Other metallic interconnect materials such as Copper require different deposition techniques.

Etching

Once a material has been deposited, etching is used to selectively form patterns such as wires and contact holes. The *wet etching* process was described earlier, and makes use of acid or basic solutions. For instance, hydrofluoric acid buffered with ammonium fluoride is typically used to etch SiO_2.

In recent years, *dry* or *plasma etching* has made a lot of inroad. A wafer is placed into the etch tool's processing chamber and given a negative electrical charge. The chamber is heated to 100°C and brought to a vacuum level of 7.5 Pa, then filled with a positively charged plasma (usually a mix of nitrogen, chlorine and boron trichloride). The opposing electrical charges cause the rapidly moving plasma molecules to align themselves in a vertical direction, forming a microscopic chemical and physical “sandblasting” action which removes the exposed material. Plasma etching has the advantage of offering a well-defined directionality to the etching action, creating patterns with sharp vertical contours.

Planarization

To reliably deposit a layer of material onto the semiconductor surface, it is essential that the surface is approximately flat. If no special steps were taken, this would definitely not be the case in modern CMOS processes, where multiple patterned metal interconnect layers are superimposed onto each other. Therefore, a *chemical-mechanical planarization* (CMP) step is included before the deposition of an extra metal layer on top of the insulating SiO_2 layer. This process uses a slurry compound—a liquid carrier with a suspended abrasive component such as aluminum oxide or silica—to microscopically plane a device layer and to reduce the step heights.

2.2.4 Simplified CMOS Process Flow

The gross outline of a potential CMOS process flow is given in Figure 2.6. The process starts with the definition of the *active regions*, this is the regions where transistors will be constructed. All other areas of the die will be covered with a thick layer of silicon dioxide (SiO_2), called the *field oxide*. This oxide acts as the insulator between neighboring devices, and is either grown (as in the process of Figure 2.1), or deposited in etched trenches (Figure 2.2) — hence the name *trench insulation*. Further insulation is provided by the addition of a reverse-biased *np*-diode, formed by adding an extra p^+ region, called the *channel-stop implant* (or *field implant*) underneath the field oxide. Next, lightly doped *p*- and *n*-wells are formed through ion implantation. To construct an NMOS transistor in a *p*-well, heavily doped *n*-type *source* and *drain* regions are implanted (or diffused) into the lightly doped *p*-type substrate. A thin layer of SiO_2, called the *gate oxide*, separates the region between the source and drain, and is itself covered by conductive polycrystalline silicon (or polysilicon, for short). The conductive material forms the *gate* of the transistor. PMOS transistors are constructed in an *n*-well in a similar fashion (just reverse *n*’s and

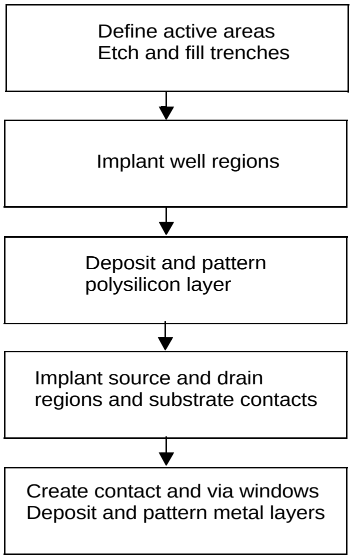
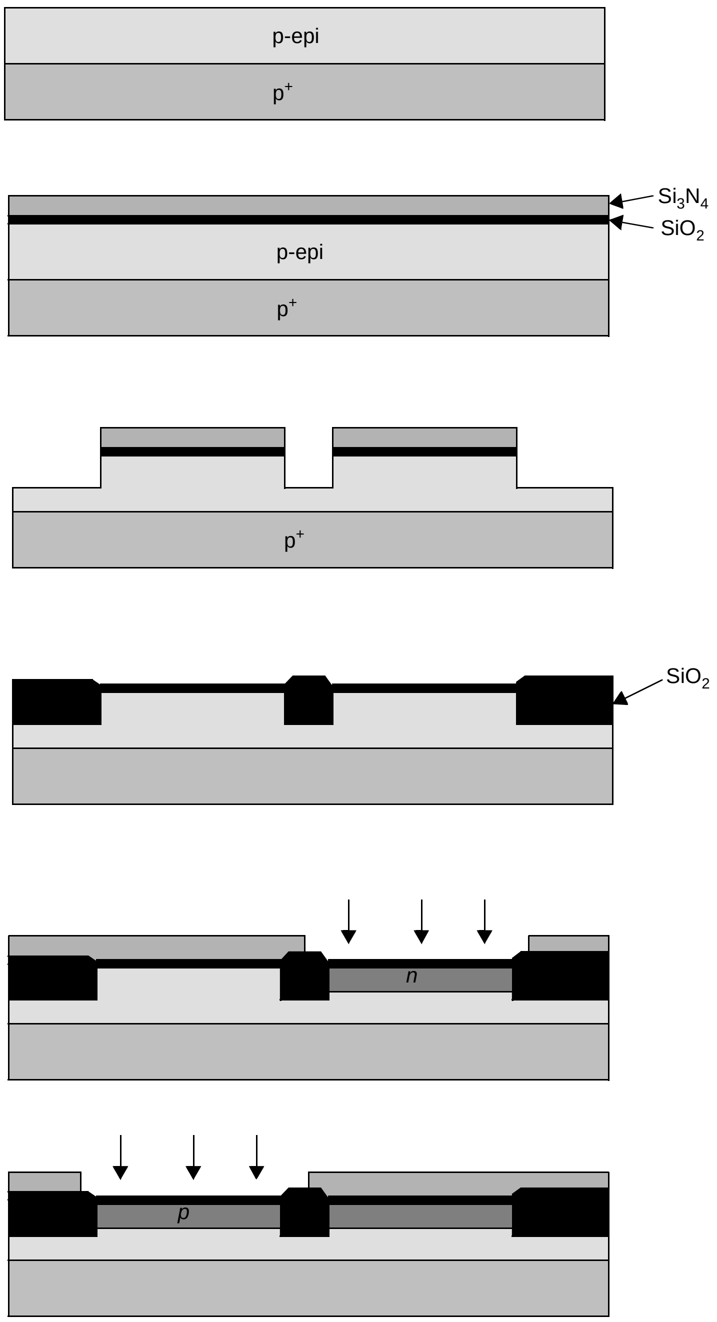


Figure 2.6 Simplified process sequence for the manufacturing of a n-dual-well CMOS circuit.

p's). Multiple insulated layers of metallic (most often Aluminum) wires are deposited on top of these devices to provide for the necessary interconnections between the transistors.

A more detailed breakdown of the flow into individual process steps and their impact on the semiconductor material is shown graphically in Figure 2.7. While most of the operations should be self-explanatory in light of the previous descriptions, some comments on individual operations are worthwhile. The process starts with a *p*-substrate surfaced with a lightly doped *p*-epitaxial layer (a). A thin layer of SiO_2 is deposited, which will serve as the gate oxide for the transistors, followed by a deposition of a thicker sacrificial silicon nitride layer (b). A plasma etching step using the complementary of the active area mask creates the trenches, used for insulating the devices (c). After providing the channel stop implant, the trenches are filled with SiO_2 followed by a number of steps to provide a flat surface (including inverse active pattern oxide etching, and chemical-mechanical planarization). At that point, the sacrificial nitride is removed (d). The *n*-well mask is used to expose only the *n*-well areas (the rest of the wafer is covered by a thick buffer material), after which an implant-annealing sequence is applied to adjust the well-doping. This is followed by a second implant step to adjust the threshold voltages of the PMOS transistors. This implant only impacts the doping in the area just below the gate oxide (e). Similar operations (using other dopants) are performed to create the p-wells, and to adjust the thresholds of the NMOS transistors (f). A thin layer of polysilicon is chemically deposited, and patterned with the aid of the polysilicon mask. Polysilicon is used both as gate electrode material for the transistors as well as an interconnect medium (g). Consecutive ion implantations are used to dope the source and drain regions of the PMOS (p^+) and NMOS (n^+) transistors, respectively (h), after which the thin gate oxide not covered by the polysilicon is etched away[1]. The same implants are also used to dope

[1] Most modern processes also include extra implants for the creation of the lightly-doped drain regions (LDD), and the creation of gate spacers at this point. We have omitted these for the sake of simplicity.



(a) Base material: p+ substrate with p-epi layer

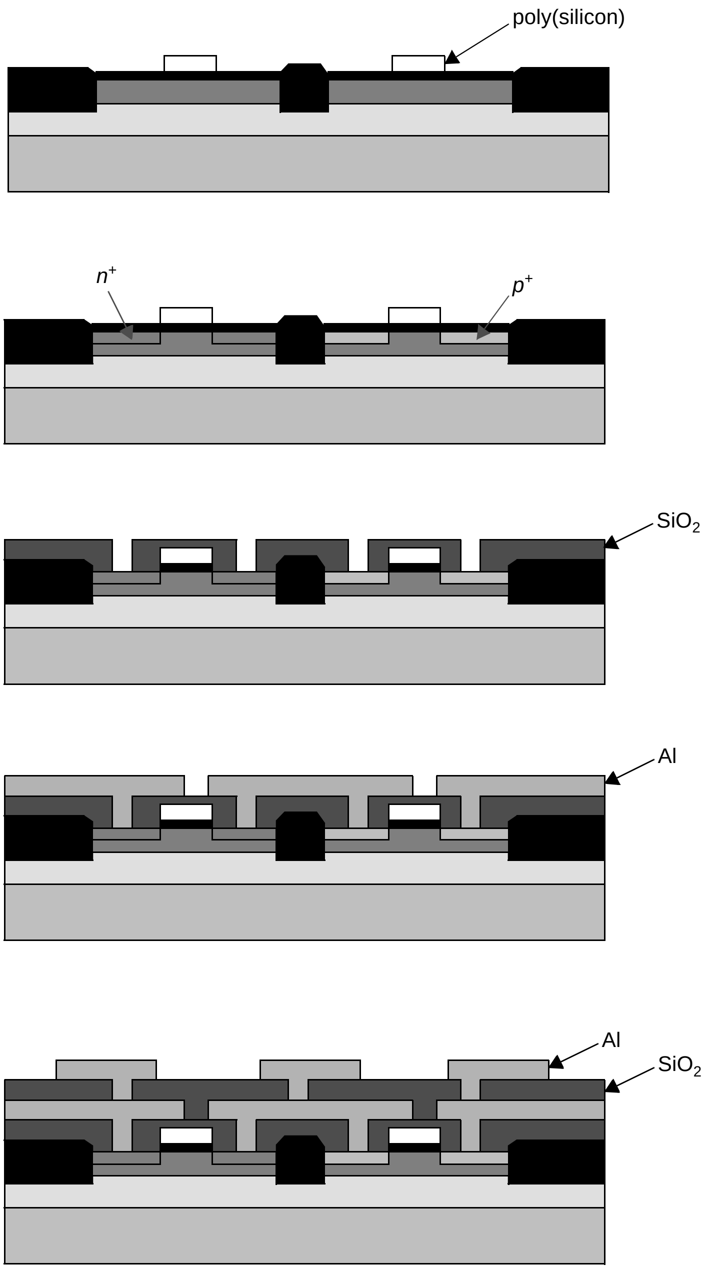
(b) After deposition of gate-oxide sacrificial nitride (acts as a buffer layer)

(c) After plasma etch of insulating trenches using the inverse of the active area mask

(d) After trench filling, CMP planarization, and removal of sacrificial nitride

(e) After *n*-well and V_{Tp} adjust implants

(f) After *p*-well and V_{Tn} adjust implants



(g) After polysilicon deposition and etch

(h) After n^+ source/drain and p^+ source/drain implants. These steps also dope the polysilicon.

(i) After deposition of SiO_2 insulator and contact hole etch.

(j) After deposition and patterning of first Al layer.

(k) After deposition of SiO_2 insulator, etching of via's, deposition and patterning of second layer of Al.

Figure 2.7 Process flow for the fabrication of an NMOS and a PMOS transistor in a dual-well CMOS process. Be aware that the drawings are stylized for understanding, and that the aspects ratios are not proportioned to reality.

the polysilicon on the surface, reducing its resistivity. Undoped polysilicon has a very high resistivity. Note that the polysilicon gate, which is patterned before the doping, actually defines the precise location of the channel region, and hence the location of the source and drain regions. This procedure allows for a very precise positioning of the two regions relative to the gate, and hence is called the *self-aligned process*. The process continues with the deposition of the metallic interconnect layers. These consists of a repetition of the following steps (i-k): deposition of the insulating material (most often SiO_2), etching of the contact or via holes, deposition of the metal (most often Aluminum and Copper, although Tungsten is often used for the lower layers), and patterning of the metal. Intermediate planarization steps ensure that the surface remains reasonable flat, even in the presence of multiple interconnect layers. After the last level of metal is deposited, a final passivation or *overglass* is deposited for protection. This layer would be CVD SiO_2, although often an additional layer of nitride is deposited as it is more impervious to moisture. The final processing step is to etch openings to the pads used for bonding.

A cross-section of the final artifact is shown in Figure 2.8. Observe how the transistors occupy only a small fraction of the total height of the structure. The interconnect layers take up the majority of the vertical dimension.

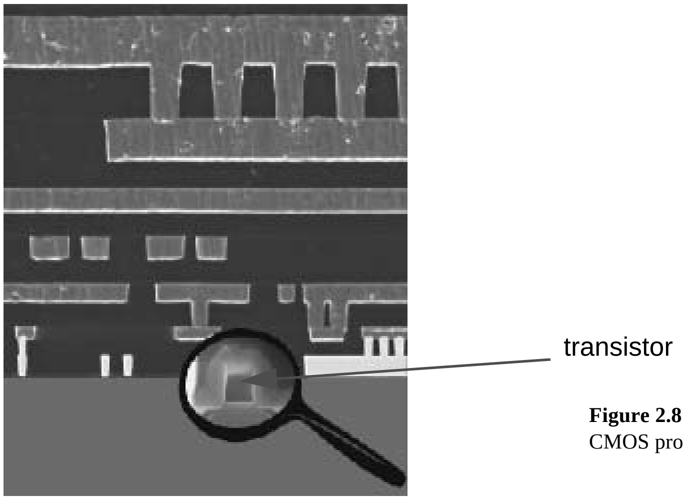


Figure 2.8 Cross-section of state-of-the-art CMOS process.

2.3 Design Rules — The Contract between Designer and Process Engineer

As processes become more complex, requiring the designer to understand the intricacies of the fabrication process and interpret the relations between the different masks is a sure road to trouble. The goal of defining a set of design rules is to allow for a ready translation of a circuit concept into an actual geometry in silicon. The design rules act as the interface or even the contract between the circuit designer and the process engineer.

Circuit designers in general want tighter, smaller designs, which lead to higher performance and higher circuit density. The process engineer, on the other hand, wants a reproducible and high-yield process. Design rules are, consequently, a compromise that attempts to satisfy both sides.

The design rules provide a set of guidelines for constructing the various masks needed in the patterning process. They consist of minimum-width and minimum-spacing constraints and requirements between objects on the same or on different layers.

The fundamental unity in the definition of a set of design rules is the *minimum line width*. It stands for the minimum mask dimension that can be safely transferred to the semiconductor material. In general, the minimum line width is set by the resolution of the patterning process, which is most commonly based on optical lithography. More advanced approaches use electron-beam. EUV or X-ray sources that offer a finer resolution, but are less attractive from an economical viewpoint today.

Even for the same minimum dimension, design rules tend to differ from company to company, and from process to process. This makes porting an existing design between different processes a time-consuming task. One approach to address this issue is to use advanced CAD techniques, which allow for migration between compatible processes. Another approach is to use *scalable design rules*. The latter approach, made popular by Mead and Conway [Mead80], defines all rules as a function of a single parameter, most often called λ. The rules are chosen so that a design is easily ported over a cross section of industrial processes. Scaling of the minimum dimension is accomplished by simply changing the value of λ. This results in a *linear scaling* of all dimensions. For a given process, λ is set to a specific value, and all design dimensions are consequently translated into absolute numbers. Typically, the minimum line width of a process is set to 2λ. For instance, for a 0.25 μm process (i.e., a process with a minimum line width of 0.25 μm), λ equals 0.125 μm.

This approach, while attractive, suffers from some disadvantages:

1. Linear scaling is only possible over a limited range of dimensions (for instance, between 0.25 μm and 0.18 μm). When scaling over larger ranges, the relations between the different layers tend to vary in a nonlinear way that cannot be adequately covered by the linear scaling rules.

2. Scalable design rules are conservative. As they represent a cross section over different technologies, they have to represent the worst-case rules for the whole set. This results in over-dimensioned and less-dense designs.

For these reasons, scalable design rules are normally avoided by industry.[2] As circuit density is a prime goal in industrial designs, most semiconductor companies tend to use *micron rules*, which express the design rules in absolute dimensions and can therefore exploit the features of a given process to a maximum degree. Scaling and porting designs between technologies under these rules is more demanding and has to be performed either manually or using advanced CAD tools.

For this textbook, we have selected a "vanilla" 0.25 μm CMOS process as our preferred implementation medium. The rest of this section is devoted to a short introduction and overview of the design rules of this process, which fall in the micron-rules class. A complete design-rule set consists of the following entities: a set of layers, relations

[2] While not entirely accurate, lambda rules are still useful to estimate the impact of a technology scale on the area of a design.

between objects on the same layer, and relations between objects on different layers. We discuss each of them in sequence.

Layer Representation

The layer concept translates the intractable set of masks currently used in CMOS into a simple set of conceptual layout levels that are easier to visualize by the circuit designer. From a designer's viewpoint, all CMOS designs are based on the following entities:

- S*ubstrates* and/or *wells*, being *p*-type (for NMOS devices) and *n*-type (for PMOS)
- *Diffusion regions* (n^+ and p^+) defining the areas where transistors can be formed. These regions are often called the *active areas*. Diffusions of an inverse type are needed to implement contacts to the wells or to the substrate. These are called *select regions*.
- One or more *polysilicon* layers, which are used to form the gate electrodes of the transistors (but serve as interconnect layers as well).
- A number of *metal interconnect* layers.
- *Contact and via* layers to provide interlayer connections.

A layout consists of a combination of polygons, each of which is attached to a certain layer. The functionality of the circuit is determined by the choice of the layers, as well as the interplay between objects on different layers. For instance, an MOS transistor is formed by the cross section of the diffusion layer and the polysilicon layer. An interconnection between two metal layers is formed by a cross section between the two metal layers and an additional contact layer. To visualize these relations, each layer is assigned a standard color (or stipple pattern for a black-and-white representation). The different layers used in our CMOS process are represented in Colorplate 1 (color insert).

Intralayer Constraints

A first set of rules defines the minimum dimensions of objects on each layer, as well as the minimum spacings between objects on the same layer. All distances are expressed in μm. These constraints are presented in a pictorial fashion in Colorplate 2.

Interlayer Constraints

Interlayer rules tend to be more complex. The fact that multiple layers are involved makes it harder to visualize their meaning or functionality. Understanding layout requires the capability of translating the two-dimensional picture of the layout drawing into the three-dimensional reality of the actual device. This takes some practice.

We present these rules in a set of separate groupings.

1. *Transistor Rules* (Colorplate 3). A transistor is formed by the overlap of the active and the polysilicon layers. From the intralayer design rules, it is already clear that the minimum length of a transistor equals 0.24 μm (the minimum width of polysilicon), while its width is at least 0.3 μm (the minimum width of diffusion). Extra rules

include the spacing between the active area and the well boundary, the gate overlap of the active area, and the active overlap of the gate.

2. *Contact and Via Rules* (Colorplates 2 and 4). A contact (which forms an interconnection between metal and active or polysilicon) or a via (which connects two metal layers) is formed by overlapping the two interconnecting layers and providing a contact hole, filled with metal, between the two. In our process, the minimum size of the contact hole is 0.3 μm, while the polysilicon and diffusion layers have to extend at least over 0.14 μm beyond the area of the contact hole. This sets the minimum area of a contact to 0.44 μm × 0.44 μm. This is larger than the dimensions of a minimum-size transistor! Excessive changes between interconnect layers in routing are thus to be avoided. The figure, furthermore, points out the minimum spacings between contact and via holes, as well as their relationship with the surrounding layers.

Well and Substrate Contacts (Colorplate 5). For robust digital circuit design, it is important for the well and substrate regions to be adequately connected to the supply voltages. Failing to do so results in a resistive path between the substrate contact of the transistors and the supply rails, and can lead to possibly devastating parasitic effects, such as latchup. It is therefore advisable to provide numerous substrate (well) contacts spread over the complete region. To establish an ohmic contact between a supply rail, implemented in metal1, and a *p*-type material, a p^+ diffusion region must be provided. This is enabled by the *select* layer, which reverses the type of diffusion. A number of rules regarding the use of the *select layer* are illustrated in Colorplate 5.

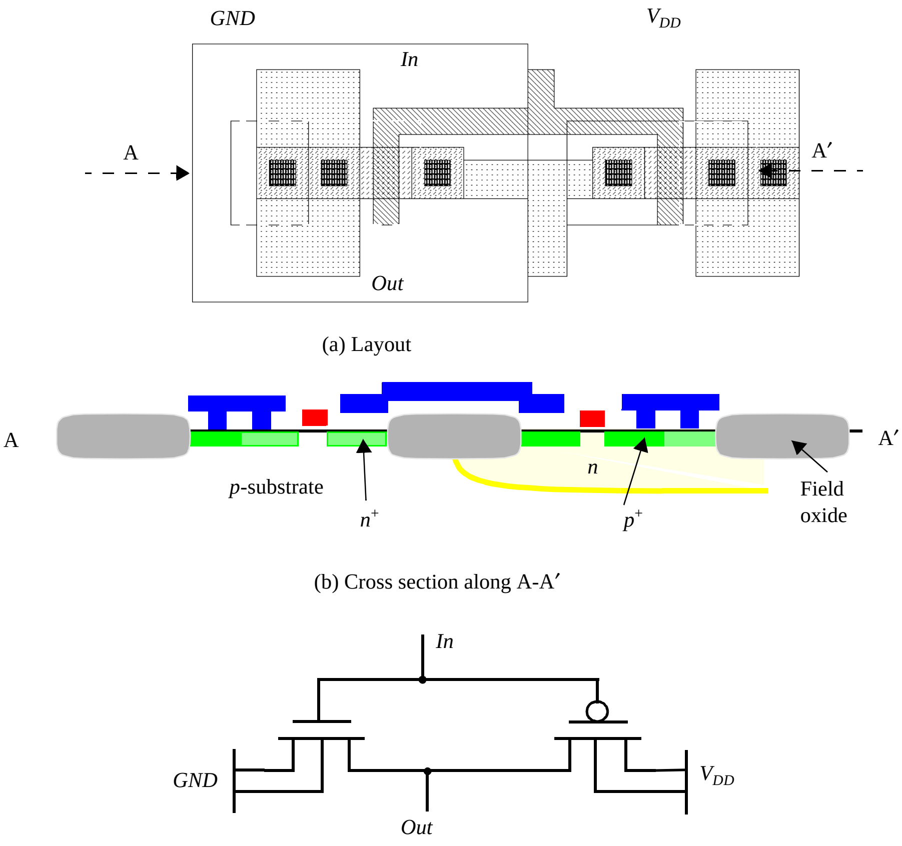
Consider an *n*-well process, which implements the PMOS transistors into an *n*-type well diffused in a *p*-type material. The nominal diffusion is p^+. To invert the polarity of the diffusion, an *n*-select layer is provided that helps to establish the n^+ diffusions for the well-contacts in the *n*-region as well as the n^+ source and drain regions for the NMOS transistors in the substrate.

Verifying the Layout

Ensuring that none of the design rules is violated is a fundamental requirement of the design process. Failing to do so will almost surely lead to a nonfunctional design. Doing so for a complex design that can contain millions of transistors is no sinecure, especially when taking into account the complexity of some design-rule sets. While design teams in the past used to spend numerous hours staring at room-size layout plots, most of this task is now done by computers. Computer-aided *Design-Rule Checking* (called *DRC*) is an integral part of the design cycle for virtually every chip produced today. A number of layout tools even perform *on-line DRC* and check the design in the background during the time of conception.

Example 2.1 Layout Example

An example of a complete layout containing an inverter is shown in Figure 2.9. To help the visualization process, a vertical cross section of the process along the design center is included as well as a circuit schematic.



(a) Layout

(b) Cross section along A-A′

(c) Circuit diagram

Figure 2.9 A detailed layout example, including vertical process cross section and circuit diagram.

It is left as an exercise for the reader to determine the sizes of both the NMOS and the PMOS transistors.

2.4 Packaging Integrated Circuits

The IC package plays a fundamental role in the operation and performance of a component. Besides providing a means of bringing signal and supply wires in and out of the silicon die, it also removes the heat generated by the circuit and provides mechanical support. Finally, its also protects the die against environmental conditions such as humidity.

The packaging technology furthermore has a major impact on the performance and power-dissipation of a microprocessor or signal processor. This influence is getting more pronounced as time progresses by the reduction in internal signal delays and on-chip capacitance as a result of technology scaling. Up to 50% of the delay of a high-performance computer is currently due to packaging delays, and this number is expected to rise.

The search for higher-performance packages with fewer inductive or capacitive parasitics has accelerated in recent years.

The increasing complexity of what can be integrated on a single die also translates into a need for ever more input-output pins, as the number of connections going off-chip tends to be roughly proportional to the complexity of the circuitry on the chip. This relationship was first observed by E. Rent of IBM (published in [Landman71]), who translated it into an empirical formula that is appropriately called *Rent's rule*. This formula relates the number of input/output pins to the complexity of the circuit, as measured by the number of gates.

$$P = K \times G^{\beta} \tag{2.1}$$

where K is the average number of I/Os per gate, G the number of gates, β the Rent exponent, and P the number of I/O pins to the chip. β varies between 0.1 and 0.7. Its value depends strongly upon the application area, architecture, and organization of the circuit, as demonstrated in Table 2.1. Clearly, microprocessors display a very different input/output behavior compared to memories.

Table 2.1 Rent's constant for various classes of systems ([Bakoglu90])

Application	β	K
Static memory	0.12	6
Microprocessor	0.45	0.82
Gate array	0.5	1.9
High-speed computer (chip)	0.63	1.4
High-speed computer (board)	0.25	82

The observed rate of pin-count increase for integrated circuits varies between 8% to 11% per year, and it has been projected that packages with more than 2000 pins will be required by the year 2010. For all these reasons, traditional dual-in-line, through-hole mounted packages have been replaced by other approaches such as surface-mount, ball-grid array, and multichip module techniques. It is useful for the circuit designer to be aware of the available options, and their pros and cons.

Due to its multi-functionality, a good package must comply with a large variety of requirements.

- **Electrical requirements**—Pins should exhibit low capacitance (both interwire and to the substrate), resistance, and inductance. A large characteristic impedance should be tuned to optimize transmission line behavior. Observe that intrinsic integrated-circuit impedances are high.
- **Mechanical and thermal properties**—The heat-removal rate should be as high as possible. Mechanical reliability requires a good matching between the thermal properties of the die and the chip carrier. Long-term reliability requires a strong connection from die to package as well as from package to board.

- **Low Cost—**Cost is always one of the more important properties. While ceramics have a superior performance over plastic packages, they are also substantially more expensive. Increasing the heat removal capacity of a package also tends to raise the package cost. The least expensive plastic packaging can dissipate up to 1 W. Somewhat more expensive, but still cheap, plastic packages can dissipate up to 2W. Higher dissipation requires more expensive ceramic packaging. Chips dissipating over 50 W require special heat sink attachments. Even more extreme techniques such as fans and blowers, liquid cooling hardware, or heat pipes, are needed for higher dissipation levels.
  Packing density is a major factor in reducing board cost. The increasing pin count either requires an increase in the package size or a reduction in the pitch between the pins. Both have a profound effect on the packaging economics.

Packages can be classified in many different ways —by their main material, the number of interconnection levels, and the means used to remove heat. In this short section, we can only glance briefly at each of those issues.

2.4.1 Package Materials

The most common materials used for the package body are ceramic and polymers (plastics). The latter have the advantage of being substantially cheaper, but suffer from inferior thermal properties. For instance, the ceramic Al_2O_3 (Alumina) conducts heat better than SiO_2 and the Polyimide plastic, by factors of 30 and 100 respectively. Furthermore, its thermal expansion coefficient is substantially closer to the typical interconnect metals. The disadvantage of alumina and other ceramics is their high dielectric constant, which results in large interconnect capacitances.

2.4.2 Interconnect Levels

The traditional packaging approach uses a two-level interconnection strategy. The die is first attached to an individual chip carrier or substrate. The package body contains an internal cavity where the chip is mounted. These cavities provide ample room for many connections to the chip leads (or pins). The leads compose the second interconnect level and connect the chip to the global interconnect medium, which is normally a PC board. Complex systems contain even more interconnect levels, since boards are connected together using backplanes or ribbon cables. The first two layers of the interconnect hierarchy are illustrated in the drawing of Figure 2.10. The following sections provide a brief overview of the interconnect techniques used at levels one and two of the interconnect hierarchy, followed by a short discussion of some more advanced packaging approaches.

Interconnect Level 1 —Die-to-Package-Substrate

For a long time, *wire bonding* was the technique of choice to provide an electrical connection between die and package. In this approach, the backside of the die is attached to the substrate using glue with a good thermal conductance. Next, the chip pads are individually connected to the lead frame with aluminum or gold wires. The wire-bonding machine use

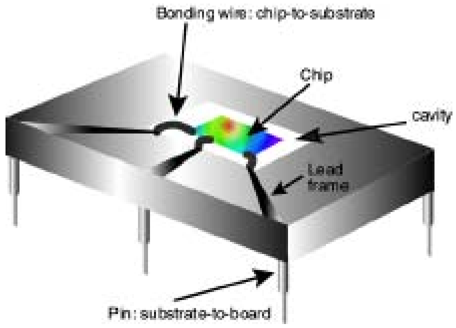


Figure 2.10 Interconnect hierarchy in traditional IC packaging.

for this purpose operates much like a sewing machine. An example of wire bonding is shown in Figure 2.11. Although the wire-bonding process is automated to a large degree, it has some major disadvantages.

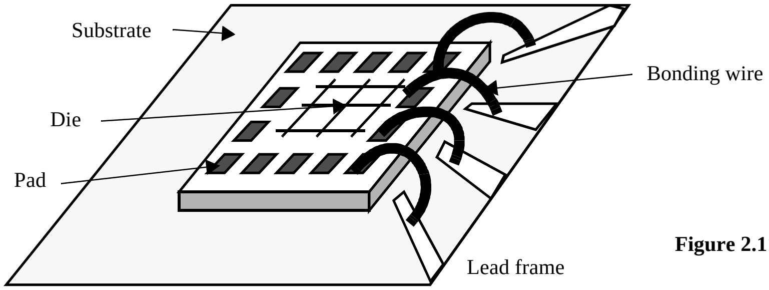


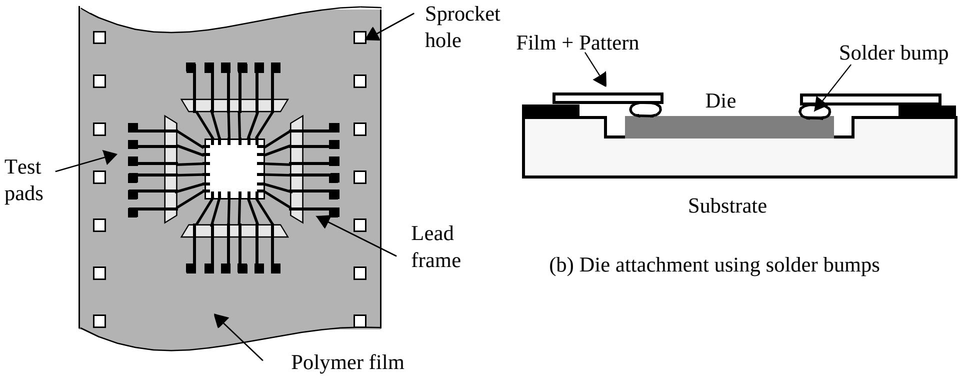
Figure 2.11 Wire bonding.

1. Wires must be attached serially, one after the other. This leads to longer manufacturing times with increasing pin counts.
2. Larger pin counts make it substantially more challenging to find bonding patterns that avoid shorts between the wires.

Bonding wires have inferior electrical properties, such as a high individual inductance (5 nH or more) and mutual inductance with neighboring signals. The inductance of a bonding wire is typically about 1 nH/mm, while the inductance per package pin ranges between 7 and 40 nH per pin depending on the type of package as well as the positioning of the pin on the package boundary [Steidel83].Typical values of the parasitic inductances and capacitances for a number of commonly used packages are summarized in Table 2.2.

3. The exact value of the parasitics is hard to predict because of the manufacturing approach and irregular outlay.

New attachment techniques are being explored as a result of these deficiencies. In one approach, called *Tape Automated Bonding* (or TAB), the die is attached to a metal lead frame that is printed on a polymer film (typically polyimide) (Figure 2.12a). The connection between chip pads and polymer film wires is made using solder bumps (Figure 2.12b). The tape can then be connected to the package body using a number of techniques. One possible approach is to use pressure connectors.



(a) Polymer tape with imprinted wiring pattern

(b) Die attachment using solder bumps

Figure 2.12 Tape-automated bonding (TAB).

Table 2.2 Typical capacitance and inductance values of package and bonding styles (from [Steidel83] and [Franzon93]).

Package Type	Capacitance (pF)	Inductance (nH)
68-pin plastic DIP	4	35
68-pin ceramic DIP	7	20
256-pin grid array	1–5	2–15
Wire bond	0.5–1	1–2
Solder bump	0.1–0.5	0.01–0.1

The advantage of the TAB process is that it is highly automated. The sprockets in the film are used for automatic transport. All connections are made simultaneously. The printed approach helps to reduce the wiring pitch, which results in higher lead counts. Elimination of the long bonding wires improves the electrical performance. For instance, for a two-conductor layer, 48 mm TAB Circuit, the following electrical parameters hold: $L \approx 0.3\text{–}0.5$ nH, $C \approx 0.2\text{–}0.3$ pF, and $R \approx 50\text{–}200\ \Omega$ [Doane93, p. 420].

Another approach is to flip the die upside-down and attach it directly to the substrate using solder bumps. This technique, called *flip-chip* mounting, has the advantage of a superior electrical performance (Figure 2.13). Instead of making all the I/O connections on the die boundary, pads can be placed at any position on the chip. This can help address the power- and clock-distribution problems, since the interconnect materials on the substrate (e.g., Cu or Au) are typically of a better quality than the Al on the chip.

Interconnect Level 2—Package Substrate to Board

When connecting the package to the PC board, *through-hole mounting* has been the packaging style of choice. A PC board is manufactured by stacking layers of copper and insu-

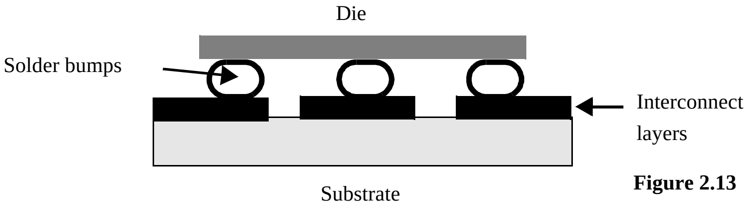


Figure 2.13 Flip-chip bonding.

lating epoxy glass. In the through-hole mounting approach, holes are drilled through the board and plated with copper. The package pins are inserted and electrical connection is made with solder (Figure 2.14a). The favored package in this class was the *dual-in-line* package or DIP (Figure 2.15a). The packaging density of the DIP degrades rapidly when the number of pins exceeds 64. This problem can be alleviated by using the *pin-grid-array* (PGA) package that has leads on the entire bottom surface instead of only on the periphery (Figure 2.15b). PGAs can extend to large pin counts (over 400 pins are possible).

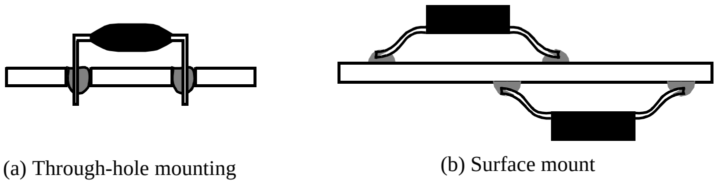
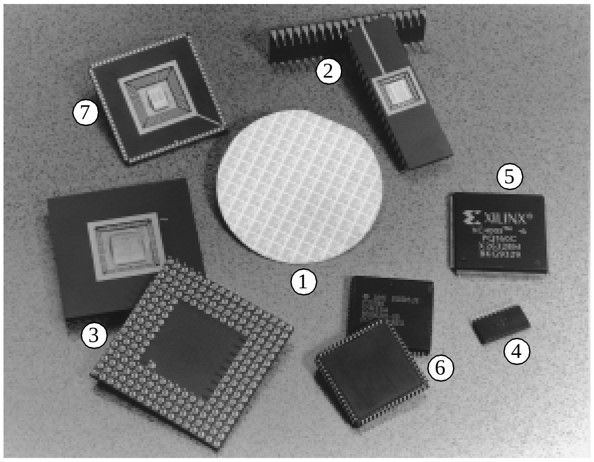


Figure 2.14 Board-mounting approaches.

The through-hole mounting approach offers a mechanically reliable and sturdy connection. However, this comes at the expense of packaging density. For mechanical reasons, a minimum pitch of 2.54 mm between the through-holes is required. Even under those circumstances, PGAs with large numbers of pins tend to substantially weaken the board. In addition, through-holes limit the board packing density by blocking lines that might otherwise have been routed below them, which results in longer interconnections. PGAs with large pin counts hence require extra routing layers to connect to the multitudes of pins. Finally, while the parasitic capacitance and inductance of the PGA are slightly lower than that of the DIP, their values are still substantial.

Many of the shortcomings of the through-hole mounting are solved by using the *surface-mount* technique. A chip is attached to the surface of the board with a solder connection without requiring any through-holes (Figure 2.14b). Packing density is increased for the following reasons: (1) through-holes are eliminated, which provides more wiring space; (2) the lead pitch is reduced; and (3) chips can be mounted on both sides of the board. In addition, the elimination of the through-holes improves the mechanical strength of the board. On the negative side, the on-the-surface connection makes the chip-board connection weaker. Not only is it cumbersome to mount a component on a board, but also more expensive equipment is needed, since a simple soldering iron will not do anymore. Finally, testing of the board is more complex, because the package pins are no longer accessible at the backside of the board. Signal probing becomes hard or even impossible.

A variety of surface-mount packages are currently in use with different pitch and pin-count parameters. Three of these packages are shown in Figure 2.15: the *small-outline package* with gull wings, the *plastic leaded package* (PLCC) with J-shaped leads, and the



1 Bare die
2 DIP
3 PGA
4 Small-outline IC
5 Quad flat pack
6 PLCC
7 Leadless carrier

Figure 2.15 An overview of commonly used package types.

leadless chip carrier. An overview of the most important parameters for a number of packages is given in Table 2.3.

Table 2.3 Parameters of various types of chip carriers.

Package type	**Lead spacing (Typical)**	**Lead count (Maximum)**
Dual-in-line	2.54 mm	64
Pin grid array	2.54 mm	> 300
Small-outline IC	1.27 mm	28
Leaded chip carrier (PLCC)	1.27 mm	124
Leadless chip carrier	0.75 mm	124

Even surface-mount packaging is unable to satisfy the quest for evermore higher pin-counts. This is worsened by the demand for power connections: today’s high performance chips, operating at low supply voltages, require as many power and ground pins as signal I/Os! When more than 300 I/O connections are needed, solder balls replace pins as the preferred interconnect medium between package and board. An example of such a packaging approach, called ceramic *ball grid array* (BGA), is shown in Figure 2.16. Solder bumps are used to connect both the die to the package substrate, and the package to the board. The area array interconnect of the BGA provides constant input/output density regardless of the number of total package I/O pins. A minimum pitch between solder balls of as low as 0.8 mm can be obtained, and packages with multiple 1000’s of I/O signals are feasible.

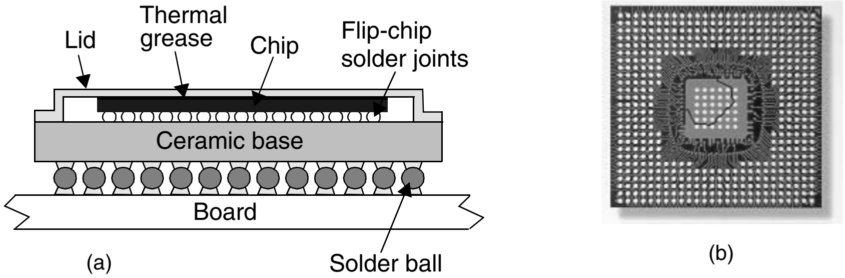


Figure 2.16 Ball grid array packaging; (a) cross-section, (b) photo of package bottom

Multi-Chip Modules—Die-to-Board

The deep hierarchy of interconnect levels in the package is becoming unacceptable in today's complex designs with their higher levels of integration, large signal counts, and increased performance requirements. The trend is toward reducing the number of levels. For the time being, attention is focused on the elimination of the first level in the packaging hierarchy. Eliminating one layer in the packaging hierarchy by mounting the die directly on the wiring backplanes—board or substrate—offers a substantial benefit when performance or density is a major issue. This packaging approach is called the multichip module technique (or MCM), and results in a substantial increase in packing density as well as improved performance.

A number of the previously mentioned die-mounting techniques can be adapted to mount dies directly on the substrate. This includes wire bonding, TAB, and flip-chip, although the latter two are preferable. The substrate itself can vary over a wide range of materials, depending upon the required mechanical, electrical, thermal, and economical requirements. Materials of choice are epoxy substrates (similar to PC boards), metal, ceramics, and silicon. Silicon has the advantage of presenting a perfect match in mechanical and thermal properties with respect to the die material.

The main advantages of the MCM approach are the increased packaging density and performance. An example of an MCM module implemented using a silicon substrate (commonly dubbed *silicon-on-silicon*) is shown in Figure 2.17. The module, which implements an avionics processor module and is fabricated by Rockwell International, contains 53 ICs and 40 discrete devices on a 2.2″ × 2.2″ substrate with aluminum polyimide interconnect. The interconnect wires are only an order of magnitude wider than what is typical for on-chip wires, since similar patterning approaches are used. The module itself has 180 I/O pins. Performance is improved by the elimination of the chip-carrier layer with its assorted parasitics, and through a reduction of the global wiring lengths on the die, a result of the increased packaging density. For instance, a solder bump has an assorted capacitance and inductance of only 0.1 pF and 0.01 nH respectively. The MCM technology can also reduce power consumption significantly, since large output drivers—and associated dissipation—become superfluous due to the reduced load capacitance of the output pads.

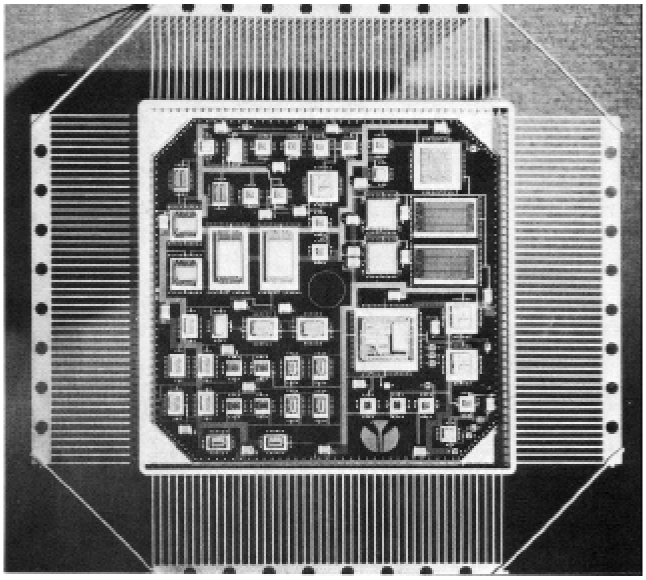


Figure 2.17 Avionics processor module. *Courtesy of Rockwell International*.

The dynamic power associated with the switching of the large load capacitances is simultaneously reduced.

While MCM technology offers some clear benefits, its main disadvantage is economic. This technology requires some advanced manufacturing steps that make the process expensive. The approach is only justifiable when either dense housing or extreme performance is essential. In the near future, this argument might become obsolete as MCM approaches proliferate.

2.4.3 Thermal Considerations in Packaging

As the power consumption of integrated circuits rises, it becomes increasingly important to efficiently remove the heat generated by the chips. A large number of failure mechanisms in ICs are accentuated by increased temperatures. Examples are leakage in reverse-biased diodes, electromigration, and hot-electron trapping. To prevent failure, the temperature of the die must be kept within certain ranges. The supported temperature range for commercial devices during operation equals 0° to 70°C. Military parts are more demanding and require a temperature range varying from –55° to 125°C.

The cooling effectiveness of a package depends upon the thermal conduction of the package material, which consists of the package substrate and body, the package composition, and the effectiveness of the heat transfer between package and cooling medium. Standard packaging approaches use still or circulating air as the cooling medium. The transfer efficiency can be improved by adding finned metal heat sinks to the package. More expensive packaging approaches, such as those used in mainframes or supercomputers, force air, liquids, or inert gases through tiny ducts in the package to achieve even greater cooling efficiencies.

<We may want to briefly introduce the thermal equation here.>

As an example, a 40-pin DIP has a thermal resistance of 38 °C/W and 25 °C/W for natural and forced convection of air. This means that a DIP can dissipate 2 watts (3 watts) of power with natural (forced) air convection, and still keep the temperature difference between the die and the environment below 75 °C. For comparison, the thermal resistance of a ceramic PGA ranges from 15 ° to 30 °C/W.

Since packaging approaches with decreased thermal resistance are prohibitively expensive, keeping the power dissipation of an integrated circuit within bounds is an economic necessity. The increasing integration levels and circuit performance make this task nontrivial. An interesting relationship in this context has been derived by Nagata [Nagata92]. It provides a bound on the integration complexity and performance as a function of the thermal parameters

$$\frac{N_G}{t_p} \le \frac{\Delta T}{\theta E} \tag{2.2}$$

where N_G is the number of gates on the chip, t_p the propagation delay, ΔT the maximum temperature difference between chip and environment, θ the thermal resistance between them, and E the switching energy of each gate.

Example 2.2 Thermal Bounds On Integration

For $\Delta T = 100$ °C, $\theta = 2.5$ °C/W and $E = 0.1$ pJ, this results in $N_G/t_p \le 4 \times 10^5$ (gates/nsec). In other words, the maximum number of gates on a chip, when all gates are operating simultaneously, must be less than 400,000 if the switching speed of each gate is 1 nsec. This is equivalent to a power dissipation of 40 W.

Fortunately, not all gates are operating simultaneously in real systems. The maximum number of gates can be substantially larger, based on the activity in the circuit. For instance, it was experimentally derived that the ratio between the average switching period and the propagation delay ranges from 20 to 200 in mini- and large-scale computers [Masaki92].

Nevertheless, Eq. (2.2) demonstrates that heat dissipation and thermal concerns present an important limitation on circuit integration. Design approaches for low power that reduce either E or the activity factor are rapidly gaining importance.

2.5 Perspective — Trends in Process Technology

Modern CMOS processes pretty much track the flow described in the previous sections although a number of the steps might be reversed, a single well approach might be followed, a grown field oxide instead of the trench approach might be used, or extra steps such as LDD (Lightly Doped Drain) might be introduced. Also, it is quite common to cover the polysilicon interconnections as well as the drain and source regions with a *silicide* such as $TiSi_2$ to improve the conductivity (see Figure 2.2). This extra operation is inserted between steps *i* and *j* of our process. Some important modifications or improvements to the technology are currently under way or are on the horizon, and deserve some attention. Beyond these, it is our belief that no dramatic changes, breaking away from the described CMOS technology, must be expected in the next decade.

2.5.1 Short-Term Developments

Copper and Low-k Dielectrics

A recurring theme in this text book will be the increasing impact of interconnect on the overall design performance. Process engineers are continuously evaluating alternative options for the traditional ‘Aluminum conductor—SiO_2 insulator’ combination that has been the norm for the last decades. In 1998, engineers at IBM introduced an approach that finally made the use of Copper as an interconnect material in a CMOS process viable and economical [IEEESpectrum98]. Copper has the advantage of have a resistivity that is substantially lower than Aluminum. Yet it has the disadvantage of easy diffusion into silicon, which degrades the characteristics of the devices. Coating the copper with a buffer material such as Titanium Nitride, preventing the diffusion, addresses this problem, but requires a special deposition process. The Dual Damascene process, introduced by IBM, (Figure 2.18) uses a metallization approach that fills trenches etched into the insulator, followed by a chemical-mechanical polishing step. This is in contrast with the traditional approach that first deposits a full metal layer, and removes the redundant material through etching.

In addition to the lower resistivity interconnections, insulator materials with a lower dielectric constant than SiO_2 —and hence lower capacitance— have also found their way into the production process starting with the 0.18 μm CMOS process generation.

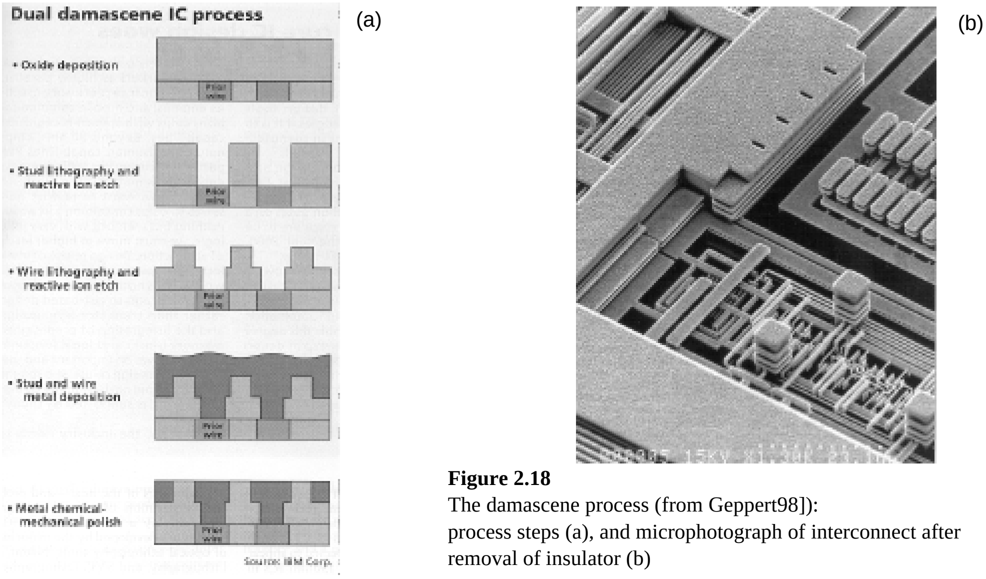
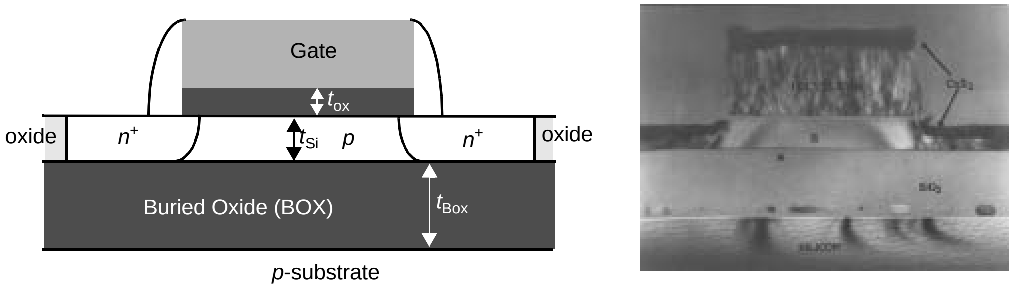

Dual damascene IC process
(a)
• Oxide deposition
Prior wire
• Stud lithography and reactive ion etch
• Wire lithography and reactive ion etch
• Stud and wire metal deposition
• Metal chemical-mechanical polish
Source: IBM Corp.
(b)

Figure 2.18
The damascene process (from Geppert98]):
process steps (a), and microphotograph of interconnect after removal of insulator (b)

Silicon-on-Insulator

While having been around for quite a long time, there seems to be a good chance that Silicon-on-Insulator (SOI) CMOS might replace the traditional CMOS process, described in the previous sections (also known as the *bulk CMOS process*). The main difference lies in the start material: the transistors are constructed in a very thin layer of silicon, deposited on top of a thick layer of insulating SiO_2 (Figure 2.19). The primary advantages of the SOI process are reduced parasitics and better transistor on-off characteristics. It has, for instance, been demonstrated by researchers at IBM that the porting of a design from a bulk CMOS to an SOI process —leaving all other design and process parameters such as channel length and oxide thickness identical— yields a performance improvement of 22% [Allen99]. Preparing a high quality SOI substrate at an economical cost was long the main hindrance against a large-scale introduction of the process. This picture has changed at the end of the nineties, and SOI is steadily moving into the mainstream.



(a) (b)

Figure 2.19 Silicon-on-insulator process— schematic diagram (a) and SEM cross-section (b).

2.5.2 In the Longer Term

Extending the life of CMOS technology beyond the next decade, and deeply below the 100 nm channel length region however will require re-engineering of both the process technology and the device structure. We already are witnessing the emergence of a wide range of new devices (such as organic transistors, molecular switches, and quantum devices). While projecting what approaches will dominate in that era equals resorting to crystal-ball gazing, one interesting development is worth mentioning.

Truly Three-Dimensional Integrated Circuits

Getting signals in and out of the computation elements in a timely fashion is one of the main challenges presented by the continued increase in integration density. One way to address this problem is to introduce extra active layers, and to sandwich them in-between the metal interconnect layers (Figure 2.20). This enables us to position high density memory on top of the logic processors implemented in the bulk CMOS, reducing the distance between computation and storage, and hence also the delay [Souri00]. In addition, devices with different voltage, performance, or substrate material requirements can be placed in different layers. For instance, the top active layer can be reserved for the realization of optical transceivers, which may help to address the input/output requirements, or MEMS (Micro Electro-Mechanical Systems) devices providing sensoring functions or radio-frequency (RF) interfaces.

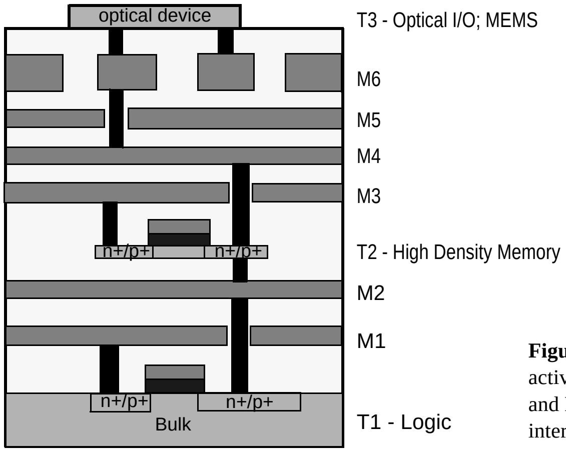


Figure 2.20 Example of true 3D integration. Extra active layers (T*), implementing high density memory and I/O, are sandwiched between the metal interconnect layers (M*).

While this approach may seem to be promising, a number of major challenges and hindrances have to be resolved to make it really viable. How to remove the dissipated heat is one of the compelling questions. Ensuring yield is another one. Yet, researchers are demonstrating major progress, and 3D integration might very well be on the horizon. Before the true solution arrives, we might have to rely on some intermediate approaches. One alternative, called 2.5D integration, is to bond two fully processed wafers, on which circuits are fabricated on the surface such that the chips completely overlap. Vias are etched to electrically connect both chips after metallization. The advantages of this technology lie in the similar electrical properties of devices on all active levels and the independence of processing temperature since all chips can be fabricated separately and later bonded. The major limitation of this technique is its lack of precision (best case alignment +/- 2 µm), which restricts the inter-chip communication to global metal lines.

One picture that strongly emerges from these futuristic devices is that the line between chip, substrate, package, and board is blurring, and that designers of these systems-on-a-die will have to consider all these aspects simultaneously.

2.6 Summary

This chapter has presented an a birds-eye view on issues regarding the manufacturing and packaging of CMOS integrated circuits.

- The manufacturing process of integrated circuits require a large number of steps, each of which consists of a sequence of basic operations. A number of these steps and/or operations, such as photolithograpical exposure and development, material deposition, and etching, are executed very repetitively in the course of the manufacturing process.

- The *optical masks* forms the central interface between the intrinsics of the manufacturing process and the design that the user wants to see transferred to the silicon fabric.
- The *design rules set* define the constraints n terms of minimum width and separation that the IC design has to adhere to if the resulting circuit is to be fully functional. This design rules acts as the contract between the circuit designer and the process engineer.
- The *package* forms the interface between the circuit implemented on the silicon die and the outside world, and as such has a major impact on the performance, reliability, longevity, and cost of the integrated circuit.

2.7 To Probe Further

Many textbooks on semiconductor manufacturing have been published over the last few decades. An excellent overview of the state-of-the-art in CMOS manufacturing can be found in the “Silicon VLSI Technology” book by J. Plummer, M. Deal, and P. Griffin [Plummer00]. A visual overview of the different steps in the manufacturing process can be found on the web at [Fullman99]. Other sources for information are the IEEE Transactions on Electron Devices, and the Technical Digest of the IEDM conference.

REFERENCES

[Allen99] D. Allen, et al., “A 0.2 μm 1.8 V SOI 550 MHz PowerPC Microprocessor with Copper Interconnects,” *Proceedings IEEE ISSCC Conference*, vol. XLII, pp. 438-439, February 1999.

[Bakoglu90] H. Bakoglu, *Circuits, Interconnections and Packaging for VLSI*, Addison-Wesley, 1990.

[Doane93] D. Doane, ed., *Multichip Module Technologies and Alternatives*, Van Nostrand-Reinhold, 1993.

[Franzon93] P. Franzon, “Electrical Design of Digital Multichip Modules,” in [Doane93], pp 525–568, 1993.

[Fullman99] Fullman Kinetics, “The Semiconductor Manufacturing Process”, *http://www.fullman-kinetics.com/semiconductors/semiconductors.html*, 1999.

[Geppert98] L. Geppert, “Technology—1998 Analysis and Forecast”, *IEEE Spectrum*, Vol. 35, No 1, pp. 23, January 1998.

[Landman71] B. Landman and R. Russo, “On a Pin versus Block Relationship for Partitions of Logic Graphs,” *IEEE Trans. on Computers*, vol. C-20, pp. 1469–1479, December 1971.

[Masaki92] A. Masaki, “Deep-Submicron CMOS Warms Up to High-Speed Logic,” *Circuits and Devices Magazine*, Nov. 1992.

[Mead80] C. Mead and L. Conway, *Introduction to VLSI Systems*, Addison-Wesley, 1980.

[Nagata92] M. Nagata, “Limitations, Innovations, and Challenges of Circuits and Devices into a Half Micrometer and Beyond,” *IEEE Journal of Solid State Circuits*, vol. 27, no. 4, pp. 465–472, April 1992.

[Plummer00] J. Plummer, M. Deal, and P. Griffin, *Silicon VLSI Technology*, Prentice Hall, 2000.

[Steidel83] C. Steidel, “*Assembly Techniques and Packaging*,” in [Sze83], pp. 551–598, 1983.

[Souri00] S. J. Souri, K. Banerjee, A. Mehrotra and K. C. Saraswat, “*Multiple Si Layer ICs: Motivation, Performance Analysis, and Design Implications*,” Proceedings 37th Design Automation Conference, pp. 213-220, June 2000.

CHAPTER 3

THE DEVICES

Qualitative understanding of MOS devices

n

Simple component models for manual analysis

n

Detailed component models for SPICE

n

Impact of process variations

3.1 Introduction

It is a well-known premise in engineering that the conception of a complex construction without a prior understanding of the underlying building blocks is a sure road to failure. This surely holds for digital circuit design as well. The basic building blocks in today's digital circuits are the silicon semiconductor devices, more specifically the MOS transistors and to a lesser degree the parasitic diodes, and the interconnect wires. The role of the semiconductor devices has been appreciated for a long time in the world of digital integrated circuits. On the other hand, interconnect wires have only recently started to play a dominant role as a result of the advanced scaling of the semiconductor technology.

Giving the reader the necessary *knowledge and understanding of these components* is the prime motivation for the next two chapters. It is not our intention to present an in-depth treatment of the physics of semiconductor devices and interconnect wires. We refer the reader to the many excellent textbooks on semiconductor devices for that purpose, some of which are referenced in the *To Probe Further* section at the end of the chapters. The goal is rather to describe the functional operation of the devices, to highlight the properties and parameters that are particularly important in the design of digital gates, and to introduce notational conventions.

Another important function of this chapter is the introduction of *models*. Taking all the physical aspects of each component into account when designing complex digital circuits leads to an unnecessary complexity that quickly becomes intractable. Such an approach is similar to considering the molecular structure of concrete when constructing a bridge. To deal with this issue, an abstraction of the component behavior called a *model* is typically employed. A range of models can be conceived for each component presenting a trade-off between accuracy and complexity. A simple first-order model is useful for manual analysis. It has limited accuracy but helps us to understand the operation of the circuit and its dominant parameters. When more accurate results are needed, complex, second- or higher-order models are employed in conjunction with computer-aided simulation. In this chapter, we present both first-order models for manual analysis as well as higher-order models for simulation for each component of interest.

Designers tend to take the component parameters offered in the models for granted. They should be aware, however, that these are only nominal values, and that the actual parameter values vary with operating temperature, over manufacturing runs, or even over a single wafer. To highlight this issue, a short discussion on *process variations* and their impact is included in the chapter.

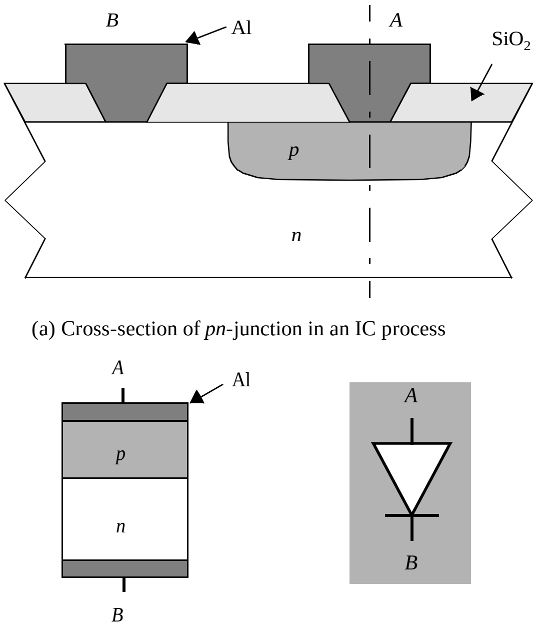
3.2 The Diode

Although diodes rarely occur directly in the schematic diagrams of present-day digital gates, they are still omnipresent. Each MOS transistor implicitly contains a number of reverse-biased diodes that directly influence the behavior of the device. Especially, the voltage-dependent capacitances contributed by these parasitic elements play an important role in the switching behavior of the MOS digital gate. Diodes are also used to protect the input devices of an IC against static charges. Therefore, a brief review of the basic properties and device equations of the diode is appropriate. Rather than being comprehensive, we choose to focus on those aspects that prove to be influential in the design of digital MOS circuits, this is the operation in reverse-biased mode.[1]

3.2.1 A First Glance at the Diode — The Depletion Region

The *pn*-junction diode is the simplest of the semiconductor devices. Figure 3.1a shows a cross-section of a typical *pn*-junction. It consists of two homogeneous regions of *p*- and *n*-type material, separated by a region of transition from one type of doping to another, which is assumed thin. Such a device is called a *step* or *abrupt junction*. The *p*-type material is doped with *acceptor* impurities (such as boron), which results in the presence of holes as the dominant or majority carriers. Similarly, the doping of silicon with *donor* impurities (such as phosphorus or arsenic) creates an *n*-type material, where electrons are the majority carriers. Aluminum contacts provide access to the *p*- and *n*-terminals of the device. The circuit symbol of the diode, as used in schematic diagrams, is introduced in Figure 3.1c.

To understand the behavior of the *pn*-junction diode, we often resort to a one-dimensional simplification of the device (Figure 3.1b). Bringing the *p*- and *n*-type materials together causes a large concentration gradient at the boundary. The electron concentration changes from a high value in the *n*-type material to a very small value in the *p*-type material. The reverse is true for the hole concentration. This gradient causes electrons to



(a) Cross-section of *pn*-junction in an IC process

(b) One-dimensional representation

(c) Diode symbol

Figure 3.1 Abrupt *pn*-junction diode and its schematic symbol.

[1] We refer the interested reader to the web-site of the textbook for a comprehensive description of the diode operation.

diffuse from *n* to *p* and holes to diffuse from *p* to *n*. When the holes leave the *p*-type material, they leave behind immobile acceptor ions, which are negatively charged. Consequently, the *p*-type material is negatively charged in the vicinity of the *pn*-boundary. Similarly, a positive charge builds up on the *n*-side of the boundary as the diffusing electrons leave behind the positively charged donor ions. The region at the junction, where the majority carriers have been removed, leaving the fixed acceptor and donor ions, is called the *depletion* or *space-charge region*. The charges create an electric field across the boundary, directed from the *n* to the *p*-region. This field counteracts the diffusion of holes and electrons, as it causes electrons to *drift* from *p* to *n* and holes to drift from *n* to *p*. Under equilibrium, the depletion charge sets up an electric field such that the drift currents are equal and opposite to the diffusion currents, resulting in a zero net flow.

The above analysis is summarized in Figure 3.2 that plots the current directions, the charge density, the electrical field, and the electrostatic field of the abrupt *pn*-junction under zero-bias conditions. In the device shown, the *p* material is more heavily doped than

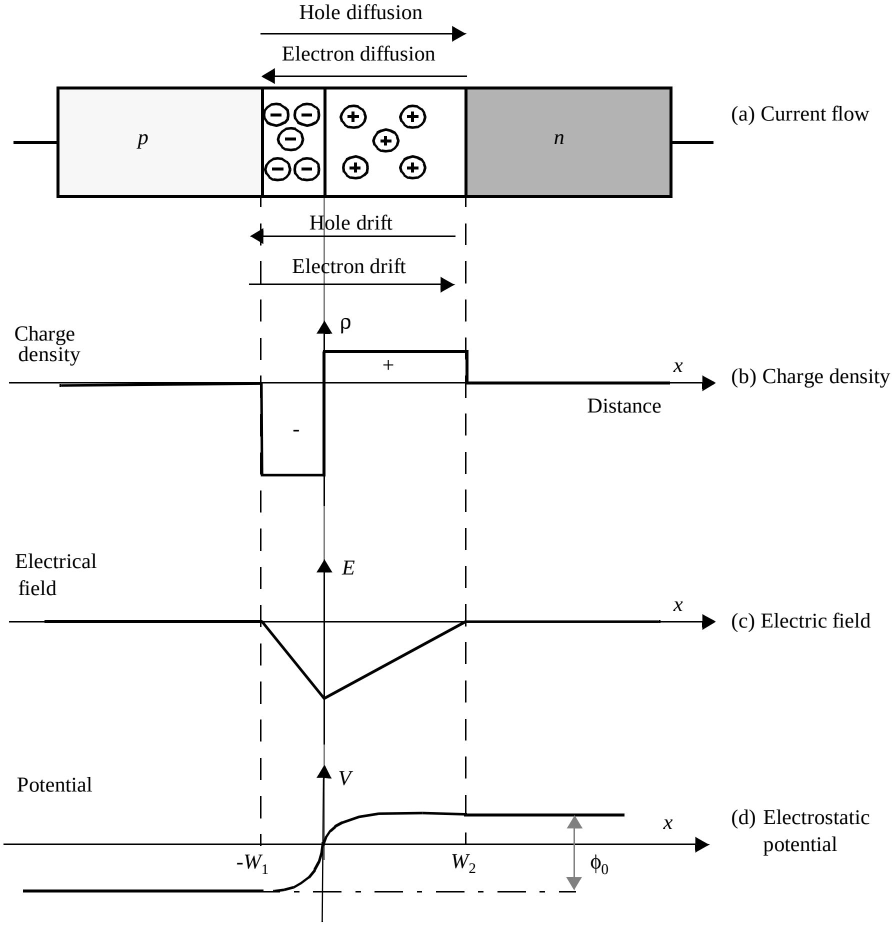


Figure 3.2 The abrupt *pn*-junction under equilibrium bias.

the n, or $N_A > N_D$, with N_A and N_D the acceptor and donor concentrations, respectively. Hence, the charge concentration in the depletion region is higher on the p-side of the junction. Figure 3.2 also shows that under zero bias, there exists a voltage ϕ_0 across the junction, called the *built-in potential*. This potential has the value

$$\phi_0 = \phi_T \ln\left[\frac{N_A N_D}{n_i^2}\right] \tag{3.1}$$

where ϕ_T is the *thermal voltage*

$$\phi_T = \frac{kT}{q} = 26\,\text{mV at 300 K} \tag{3.2}$$

The quantity n_i is the intrinsic carrier concentration in a pure sample of the semiconductor and equals approximately 1.5×10^{10} cm^{-3} at 300 K for silicon.

Example 3.1 Built-in Voltage of *pn*-junction

An abrupt junction has doping densities of $N_A = 10^{15}$ atoms/cm^3, and $N_D = 10^{16}$ atoms/cm^3. Calculate the built-in potential at 300 K.

From Eq. (3.1),

$$\phi_0 = 26\ln\left[\frac{10^{15} \times 10^{16}}{2.25 \times 10^{20}}\right] \text{mV} = 638 \text{ mV}$$

3.2.2 Static Behavior

The Ideal Diode Equation

Assume now that a forward voltage V_D is applied to the junction or, in other words, that the potential of the p-region is raised with respect to the n-zone. The applied potential lowers the potential barrier. Consequently, the flow of mobile carriers across the junction increases as the diffusion current dominates the drift component. These carriers traverse

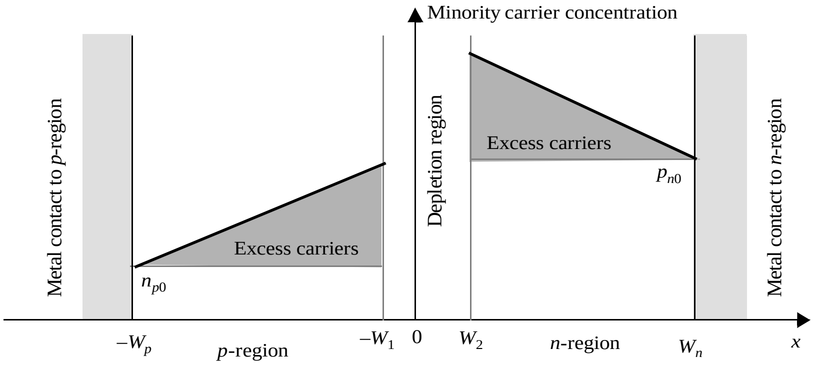


Figure 3.3 Minority carrier concentrations in the neutral region near an abrupt *p*n-junction under forward-bias conditions.

the depletion region and are injected into the neutral *n*- and *p*-regions, where they become minority carriers, as is illustrated in Figure 3.3. Under the assumption that no voltage gradient exists over the neutral regions, which is approximately the case for most modern devices, these minority carriers will diffuse through the region as a result of the concentration gradient until they get recombined with a majority carrier. The net result is a current flowing through the diode from the *p-r*egion to the *n*-region, and the diode is said to be in the *forward-bias* mode.

On the other hand, when a reverse voltage V_D is applied to the junction or when the potential of the *p*-region is lowered with respect to the *n*-region, the potential barrier is raised. This results in a reduction in the diffusion current, and the drift current becomes dominant. A current flows from the *n*-region to the *p*-region. Since the number of minority carriers in the neutral regions (electrons in the *p*-zone, holes in the *n*-region) is very small, this drift current component is virtually ignorable (Figure 3.4). It is fair to state that in the *reverse-bias* mode the diode operates as a nonconducting, or blocking, device. The diode thus acts as a one-way conductor.

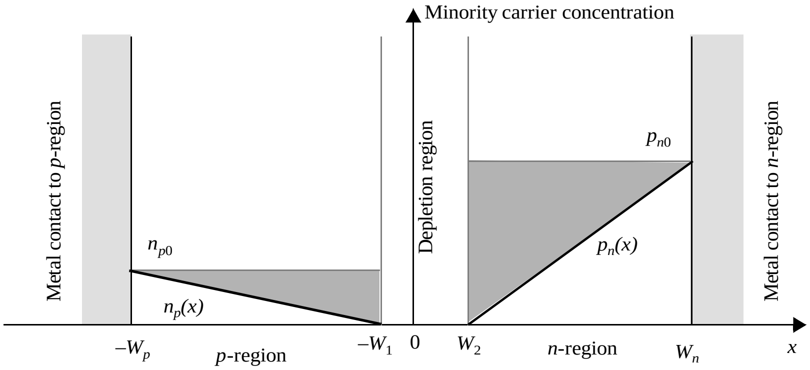


Figure 3.4 Minority carrier concentration in the neutral regions near the *pn*-junction under reverse-bias conditions.

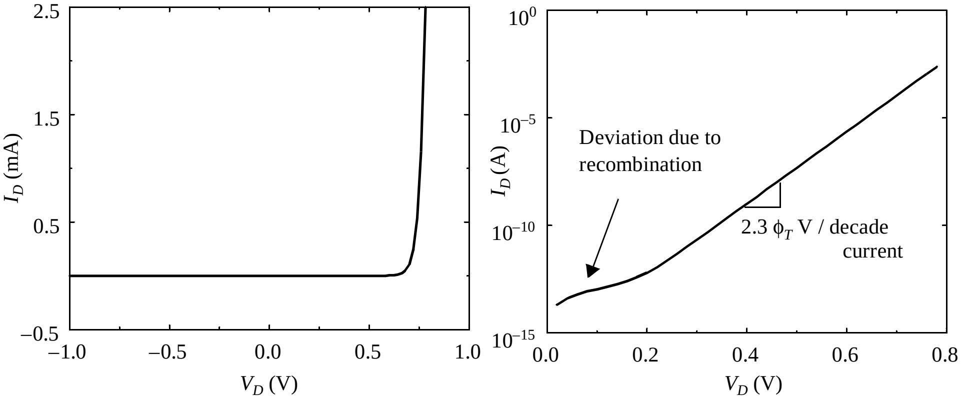
The most important property of the diode current is its *exponential dependence* upon the applied bias voltage. This is illustrated in Figure 3.5, which plots the diode current I_D as a function of the bias voltage V_D. The exponential behavior for positive-bias voltages is even more apparent in Figure 3.5b, where the current is plotted on a logarithmic scale. The current increases by a factor of 10 for every extra 60 mV (= 2.3 ϕ_T) of forward bias. At small voltage levels ($V_D < 0.15$ V), a deviation from the exponential dependence can be observed, which is due to the recombination of holes and electrons in the depletion region.

The behavior of the diode for both forward- and reverse bias conditions is best described by the well-known *ideal diode equation*, which relates the current through the diode I_D to the diode bias voltage V_D

$$I_D = I_S(e^{V_D/\phi_T} - 1) \tag{3.3}$$

Observe how Eq. (3.3) corresponds to the exponential behavior plotted in Figure 3.5. ϕ_T is the thermal voltage of Eq. (3.2) and is equal to 26 mV at room temperature.

I_S represents a constant value, called the *saturation current* of the diode. It is proportional to the area of the diode, and a function of the doping levels and widths of the neutral



(a) On a linear scale

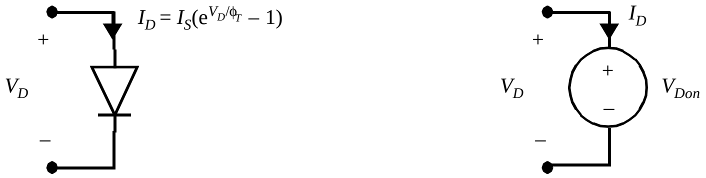
(b) On a logarithmic scale (forward bias only)

Figure 3.5 Diode current as a function of the bias voltage V_D.

regions. Most often, I_S is determined empirically.It is worth mentioning that in actual devices, the reverse currents are substantially larger than the saturation current I_S. This is due to the thermal generation of hole and electron pairs in the depletion region. The electric field present sweeps these carriers out of the region, causing an additional current component. For typical silicon junctions, the saturation current is nominally in the range of 10^{-17} A/μm^2, while the actual reverse currents are approximately three orders of magnitude higher. Actual device measurements are, therefore, necessary to determine realistic values for the reverse diode leakage currents.

Models for Manual Analysis

The derived current-voltage equations can be summarized in a set of simple models that are useful in the manual analysis of diode circuits. A first model, shown in Figure 3.6a, is based on the ideal diode equation Eq. (3.3). While this model yields accurate results, it has the disadvantage of being strongly nonlinear. This prohibits a fast, first-order analysis of the dc-operation conditions of a network. An often-used, simplified model is derived by inspecting the diode current plot of Figure 3.5. For a “fully conducting” diode, the voltage drop over the diode V_D lies in a narrow range, approximately between 0.6 and 0.8 V. To a



(a) Ideal diode model

(b) First-order diode model

Figure 3.6 Diode models.

first degree, it is reasonable to assume that a conducting diode has a fixed voltage drop V_{Don} over it. Although the value of V_{Don} depends upon I_S, a value of 0.7 V is typically assumed. This gives rise to the model of Figure 3.6b, where a conducting diode is replaced by a fixed voltage source.

Example 3.2 Analysis of Diode Network

Consider the simple network of Figure 3.7 and assume that V_S = 3 V, R_S = 10 kΩ and $I_S = 0.5 \times 10^{-16}$ A. The diode current and voltage are related by the following network equation

$$V_S - R_S I_D = V_D$$

Inserting the ideal diode equation and (painfully) solving the nonlinear equation using either numerical or iterative techniques yields the following solution: I_D = 0.224 mA, and V_D = 0.757 V. The simplified model with V_{Don} = 0.7 V produces similar results (V_D = 0.7 V, I_D = 0.23 A) with far less effort. It hence makes considerable sense to use this model when determining a first-order solution of a diode network.

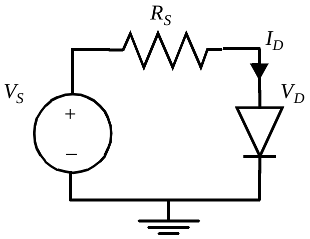


Figure 3.7 A simple diode circuit.

3.2.3 Dynamic, or Transient, Behavior

So far, we have mostly been concerned with the static, or steady-state, characteristics of the diode. Just as important in the design of digital circuits is the response of the device to changes in its bias conditions. The transient, or dynamic, response determines the maximum speed at which the device can be operated. Because the operation mode of the diode is a function of the amount of charge present in both the neutral and the space-charge regions, its dynamic behavior is strongly determined by how fast charge can be moved around.

While we could embark at this point onto an in-depth analysis of the switching behavior of the diode in the forward-biasing mode, it is our conviction that this would be besides the point and unnecessarily complicate the discussion. In fact, all diodes in an operational MOS digital integrated circuit are reverse-biased and are supposed to remain so under all circumstances. Only under exceptional conditions may forward-biasing occur. A signal over(under) shooting the supply rail is an example of such. Due to its detrimental impact on the overall circuit operation, this should be avoided under all circumstances.

Hence, we will devote our attention solely to what governs the dynamic response of the diode under reverse-biasing conditions, the depletion-region charge.

Depletion-Region Capacitance

In the ideal model, the depletion region is void of mobile carriers, and its charge is determined by the immobile donor and acceptor ions. The corresponding charge distribution under zero-bias conditions was plotted in Figure 3.2. This picture can be easily extended to incorporate the effects of biasing. At an intuitive level the following observations can be easily verified—under forward-bias conditions, the potential barrier is reduced, which means that less space charge is needed to produce the potential difference. This corresponds to a reduced depletion-region width. On the other hand, under reverse conditions, the potential barrier is increased corresponding to an increased space charge and a wider depletion region. These observations are confirmed by the well- known depletion-region expressions given below (a derivation of these expressions, which are valid for abrupt junctions, is either simple or can be found in any textbook on devices such as [Howe97]). One observation is crucial — due to the global charge neutrality requirement of the diode, the total acceptor and donor charges must be numerically equal.

1. Depletion-region charge (V_D is positive for forward bias).

$$Q_j = A_D \sqrt{\left(2\varepsilon_{si} q \frac{N_A N_D}{N_A + N_D}\right)(\phi_0 - V_D)} \tag{3.4}$$

2. Depletion-region width.

$$W_j = W_2 - W_1 = \sqrt{\left(\frac{2\varepsilon_{si}}{q}\frac{N_A + N_D}{N_A N_D}\right)(\phi_0 - V_D)} \tag{3.5}$$

3. Maximum electric field.

$$E_j = \sqrt{\left(\frac{2q}{\varepsilon_{si}}\frac{N_A N_D}{N_A + N_D}\right)(\phi_0 - V_D)} \tag{3.6}$$

In the preceding equations ε_{si} stands for the electrical permittivity of silicon and equals 11.7 times the permittivity of a vacuum, or 1.053×10^{-10} F/m. The ratio of the n- versus p-side of the depletion-region width is determined by the doping-level ratios: $W_2/(-W_1) = N_A/N_D$.

From an abstract point of view, it is possible to visualize the depletion region as a capacitance, albeit one with very special characteristics. Because the space-charge region contains few mobile carriers, it acts as an insulator with a dielectric constant ε_{si} of the semiconductor material. The n- and p-regions act as the capacitor plates. A small change in the voltage applied to the junction dV_D causes a change in the space charge dQ_j. Hence, a depletion-layer capacitance can be defined

$$\begin{aligned} C_j &= \frac{dQ_j}{dV_D} = A_D \sqrt{\left(\frac{\varepsilon_{si} q}{2}\frac{N_A N_D}{N_A + N_D}\right)(\phi_0 - V_D)^{-1}} \\ &= \frac{C_{j0}}{\sqrt{1 - V_D/\phi_0}} \end{aligned} \tag{3.7}$$

where C_{j0} is the capacitance under zero-bias conditions and is only a function of the physical parameters of the device.

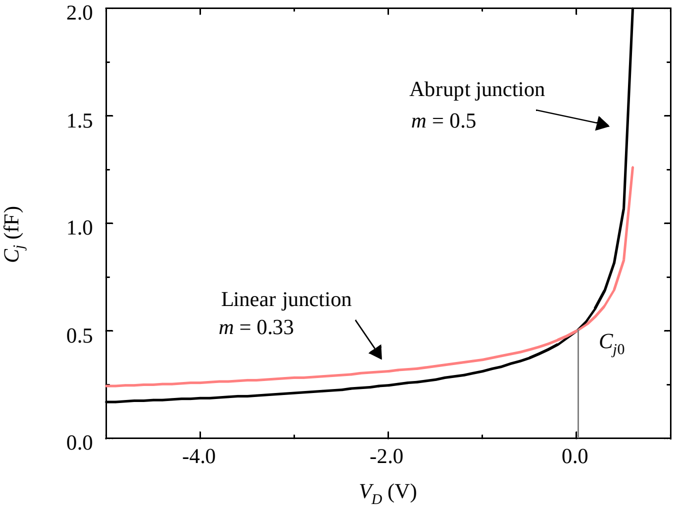


Figure 3.8 Junction capacitance (in fF/μm^2) as a function of the applied bias voltage.

$$C_{j0} = A_D\sqrt{\left(\frac{\varepsilon_{si}q}{2}\frac{N_AN_D}{N_A+N_D}\right)\phi_0^{-1}} \tag{3.8}$$

Notice that the same capacitance value is obtained when using the standard parallel-plate capacitor equation $C_j = \varepsilon_{si}\, A_D/W_j$ (with W_j given in Eq. (3.5)). Typically, the A_D factor is omitted, and C_j and C_{j0} are expressed as a capacitance/unit area.

The resulting junction capacitance is plotted in the function of the bias voltage in Figure 3.8 for a typical silicon diode found in MOS circuits. A strong *nonlinear dependence* can be observed. Note also that the capacitance decreases with an increasing reverse bias: a reverse bias of 5 V reduces the capacitance by more than a factor of two.

Example 3.3 Junction Capacitance

Consider the following silicon junction diode: $C_{j0} = 2 \times 10^{-3}$ F/m^2, $A_D = 0.5$ μm^2, and ϕ_0 = 0.64 V. A reverse bias of -2.5 V results in a junction capacitance of 0.9×10^{-3} F/m^2 (0.9 fF/μm$^{2)}$, or, for the total diode, a capacitance of 0.45 fF.

Equation (3.7) is only valid under the condition that the *pn*-junction is an *abrupt junction*, where the transition from *n* to *p* material is instantaneous. This is often not the case in actual integrated-circuit *pn*-junctions, where the transition from *n* to *p* material can be gradual. In those cases, a linear distribution of the impurities across the junction is a better approximation than the step function of the abrupt junction. An analysis of the *linearly-graded junction* shows that the junction capacitance equation of Eq. (3.7) still holds, but with a variation in order of the denominator. A more generic expression for the junction capacitance can be provided,

$$C_j = \frac{C_{j0}}{(1 - V_D/\phi_0)^m} \tag{3.9}$$

where m is called the *grading coefficient* and equals 1/2 for the abrupt junction and 1/3 for the linear or graded junction. Both cases are illustrated in Figure 3.8.

Large-Signal Depletion-Region Capacitance

Figure 3.8 raises awareness to the fact that the junction capacitance is a voltage-dependent parameter whose value varies widely between bias points. In digital circuits, operating voltages tend to move rapidly over a wide range. Under those circumstances, it is more attractive to replace the voltage-dependent, nonlinear capacitance C_j by an equivalent, linear capacitance C_{eq}. C_{eq} is defined such that, for a given voltage swing from voltages V_{high} to V_{low}, the same amount of charge is transferred as would be predicted by the nonlinear model

$$C_{eq} = \frac{\Delta Q_j}{\Delta V_D} = \frac{Q_j(V_{high}) - Q_j(V_{low})}{V_{high} - V_{low}} = K_{eq} C_{j0} \tag{3.10}$$

Combining Eq. (3.4) (extended to accommodate the grading coefficient m) and Eq. (3.10) yields the value of K_{eq}.

$$K_{eq} = \frac{-\phi_0^m}{(V_{high} - V_{low})(1-m)}[(\phi_0 - V_{high})^{1-m} - (\phi_0 - V_{low})^{1-m}] \tag{3.11}$$

Example 3.4 Average Junction Capacitance

The diode of Example 3.3 is switched between 0 and –2.5 V. Compute the average junction capacitance (m = 0.5).

For the defined voltage range and for $\phi_0 = 0.64$ V, K_{eq} evaluates to 0.622. The average capacitance hence equals 1.24 fF/µm^2.

3.2.4 The Actual Diode—Secondary Effects

In practice, the diode current is less than what is predicted by the ideal diode equation. Not all applied bias voltage appears directly across the junction, as there is always some voltage drop over the neutral regions. Fortunately, the resistivity of the neutral zones is generally small (between 1 and 100 Ω, depending upon the doping levels) and the voltage drop only becomes significant for large currents (>1 mA). This effect can be modeled by adding a resistor in series with the *n*- and *p*-region diode contacts.

In the discussion above, it was further assumed that under sufficient reverse bias, the reverse current reaches a constant value, which is essentially zero. When the reverse bias exceeds a certain level, called the *breakdown voltage*, the reverse current shows a dramatic increase as shown in Figure 3.9. In the diodes found in typical CMOS processes, this increase is caused by the *avalanche breakdown*. The increasing reverse bias heightens the magnitude of the electrical field across the junction. Consequently, carriers crossing the depletion region are accelerated to high velocity. At a critical field E_{crit}, the carriers

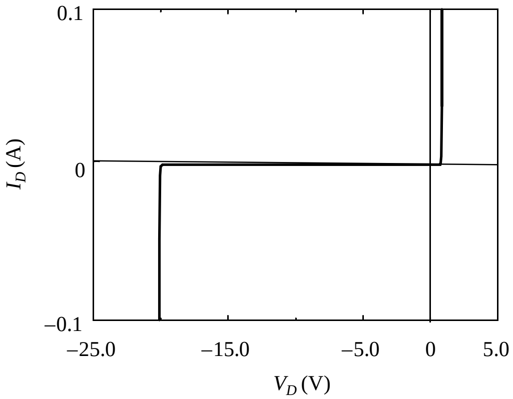


Figure 3.9 *I-V* characteristic of junction diode, showing breakdown under reverse-bias conditions (Breakdown voltage = 20 V).

reach a high -enough energy level that electron-hole pairs are created on collision with immobile silicon atoms. These carriers create, in turn, more carriers before leaving the depletion region. The value of E_{crit} is approximately 2×10^5 V/cm for impurity concentrations of the order of 10^{16} cm^{-3}. While avalanche breakdown in itself is not destructive and its effects disappear after the reverse bias is removed, maintaining a diode for a long time in avalanche conditions is not recommended as the high current levels and the associated heat dissipation might cause permanent damage to the structure. Observe that avalanche breakdown is not the only breakdown mechanism encountered in diodes. For highly doped diodes, another mechanism, called Zener breakdown, can occur. Discussion of this phenomenon is beyond the scope of this text.

Finally, it is worth mentioning that the diode current is affected by the operating *temperature* in a dual way:

1. The thermal voltage ϕ_T, which appears in the exponent of the current equation, is linearly dependent upon the temperature. An increase in ϕ_T causes the current to drop.
2. The saturation current I_S is also temperature-dependent, as the thermal equilibrium carrier concentrations increase with increasing temperature. Theoretically, the saturation current approximately doubles every 5 °C. Experimentally, the reverse current has been measured to double every 8 °C.

This dual dependence has a significant impact on the operation of a digital circuit. First of all, current levels (and hence power consumption) can increase substantially. For instance, for a forward bias of 0.7 V at 300 K, the current increases approximately 6%/°C, and doubles every 12 °C. Secondly, integrated circuits rely heavily on reverse-biased diodes as isolators. Increasing the temperature causes the leakage current to increase and decreases the isolation quality.

3.2.5 The SPICE Diode Model

In the preceding sections, we have presented a model for manual analysis of a diode circuit. For more complex circuits, or when a more accurate modeling of the diode that takes

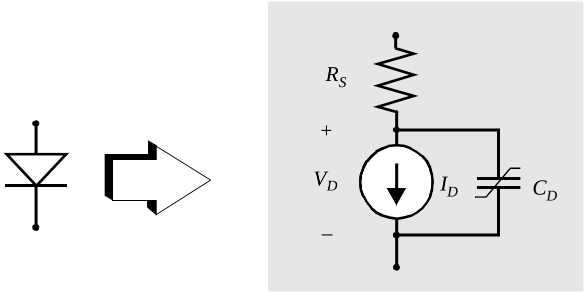


Figure 3.10 SPICE diode model.

into account second-order effects is required, manual circuit evaluation becomes intractable, and computer-aided simulation is necessary. While different circuit simulators have been developed over the last decades, the SPICE program, developed at the University of California at Berkeley, is definitely the most successful [Nagel75]. Simulating an integrated circuit containing active devices requires a mathematical model for those devices (which is called the *SPICE model* in the rest of the text). The accuracy of the simulation depends directly upon the quality of this model. For instance, one cannot expect to see the result of a second-order effect in the simulation if this effect is not present in the device model. Creating accurate and computation-efficient SPICE models has been a long process and is by no means finished. Every major semiconductor company has developed their own proprietary models, which it claims have either better accuracy or computational efficiency and robustness.

The standard SPICE model for a diode is simple, as shown in Figure 3.10. The steady-state characteristic of the diode is modeled by the nonlinear current source I_D, which is a modified version of the ideal diode equation

$$I_D = I_S(e^{V_D / n\phi_T} - 1) \tag{3.12}$$

The extra parameter n is called the *emission coefficient*. It equals 1 for most common diodes but can be somewhat higher than 1 for others. The resistor R_s models the series resistance contributed by the neutral regions on both sides of the junction. For higher current levels, this resistance causes the internal diode V_D to differ from the externally applied voltage, hence causing the current to be lower than what would be expected from the ideal diode equation.

The dynamic behavior of the diode is modeled by the nonlinear capacitance C_D, which combines the two different charge-storage effects in the diode: the space (or depletion-region) charge, and the excess minority carrier charge. Only the former was discussed in this chapter, as the latter is only an issue under forward-biasing conditions.

$$C_D = \frac{C_{j0}}{(1 - V_D / \phi_0)^m} + \frac{\tau_T I_S}{\phi_T} e^{V_D / n\phi_T} \tag{3.13}$$

A listing of the parameters used in the diode model is given in Table 3.1. Besides the parameter name, symbol, and SPICE name, the table contains also the default value used by SPICE in case the parameter is left undefined. Observe that this table is by no means complete. Other parameters are available to govern second-order effects such as breakdown, high-level injection, and noise. To be concise, we chose to limit the listing to the parameters of direct interest to this text. For a complete description of the device models (as well as the usage of SPICE), we refer to the numerous textbooks devoted to SPICE (e.g., [Banhzaf92], [Thorpe92]).

Table 3.1 First-order SPICE diode model parameters.

Parameter Name	Symbol	SPICE Name	Units	Default Value
Saturation current	I_S	IS	A	1.0 E–14
Emission coefficient	n	N	–	1
Series resistance	R_S	RS	Ω	0
Transit time	τ_T	TT	s	0
Zero-bias junction capacitance	C_{j0}	CJ0	F	0
Grading coefficient	m	M	–	0.5
Junction potential	ϕ_0	VJ	V	1

3.3 The MOS(FET) Transistor

The metal-oxide-semiconductor field-effect transistor (MOSFET or MOS, for short) is certainly the workhorse of contemporary digital design. Its major asset from a digital perspective is that the device performs very well as a switch, and introduces little parasitic effects. Other important advantages are its integration density combined with a relatively "simple" manufacturing process, which make it possible to produce large and complex circuits in an economical way.

Following the approach we took for the diode, we restrict ourselves in this section to a general overview of the transistor and its parameters. After a generic overview of the device, we present an analytical description of the transistor from a static (steady-state) and dynamic (transient) viewpoint. The discussion concludes with an enumeration of some second-order effects and the introduction of the SPICE MOS transistor models.

3.3.1 A First Glance at the Device

The MOSFET is a four terminal device. The voltage applied to the *gate* terminal determines if and how much current flows between the *source* and the *drain* ports. The *body* represents the fourth terminal of the transistor. Its function is secondary as it only serves to modulate the device characteristics and parameters.

At the most superficial level, the transistor can be considered to be a switch. When a voltage is applied to the gate that is larger than a given value called the *threshold voltage* V_T, a conducting channel is formed between drain and source. In the presence of a voltage difference between the latter two, current flows between them. The conductivity of the channel is modulated by the gate voltage—the larger the voltage difference between gate and source, the smaller the resistance of the conducting channel and the larger the current.

When the gate voltage is lower than the threshold, no such channel exists, and the switch is considered open.

Two types of MOSFET devices can be identified. The NMOS transistor consists of n^+ drain and source regions, embedded in a *p*-type substrate. The current is carried by electrons moving through an *n*-type channel between source and drain. This is in contrast with the *pn*-junction diode, where current is carried by both holes and electrons. MOS devices can also be made by using an *n*-type substrate and p^+ drain and source regions. In such a transistor, current is carried by holes moving through a *p*-type channel. The device is called a *p*-channel MOS, or PMOS transistor. In a complementary MOS technology (CMOS), both devices are present. The cross-section of a contemporary dual-well CMOS process was presented in Chapter 2, and is repeated here for convenience (Figure 3.11).

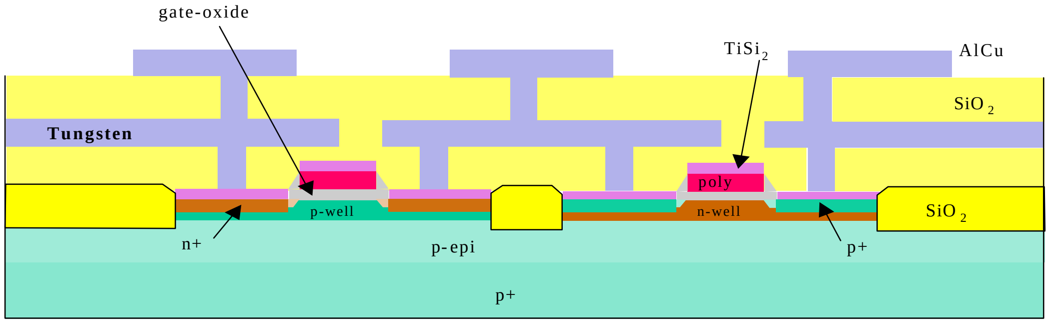
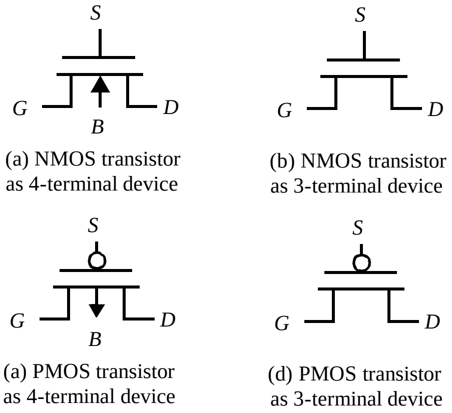


Figure 3.11 Cross-section of contemporary dual-well CMOS process.

Circuit symbols for the various MOS transistors are shown in Figure 3.12. As mentioned earlier, the transistor is a four-port device with gate, source, drain, and body terminals (Figures a and c). Since the body is generally connected to a dc supply that is identical for all devices of the same type (GND for NMOS, V_{dd} for PMOS), it is most often not shown on the schematics (Figures b and d). **If the fourth terminal is not shown, it is assumed that the body is connected to the appropriate supply.**



(a) NMOS transistor as 4-terminal device

(b) NMOS transistor as 3-terminal device

(a) PMOS transistor as 4-terminal device

(d) PMOS transistor as 3-terminal device

Figure 3.12 Circuit symbols for MOS transistors.

3.3.2 The MOS Transistor under Static Conditions

In the derivation of the static model of the MOS transistor, we concentrate on the NMOS device. All the arguments made are valid for PMOS devices as well as will be discussed at the end of the section.

The Threshold Voltage

Consider first the case where $V_{GS} = 0$ and drain, source, and bulk are connected to ground. The drain and source are connected by back-to-back *pn*-junctions (substrate-source and substrate-drain). Under the mentioned conditions, both junctions have a 0 V bias and can be considered off, which results in an extremely high resistance between drain and source.

Assume now that a positive voltage is applied to the gate (with respect to the source), as shown in Figure 3.13. The gate and substrate form the plates of a capacitor with the gate oxide as the dielectric. The positive gate voltage causes positive charge to accumulate on the gate electrode and negative charge on the substrate side. The latter manifests itself initially by repelling mobile holes. Hence, a depletion region is formed below the gate. This depletion region is similar to the one occurring in a *pn*-junction diode. Consequently, similar expressions hold for the width and the space charge per unit area. Compare these expressions to Eq. (3.4) and Eq. (3.5).

$$W_d = \sqrt{\frac{2\varepsilon_{si}\phi}{qN_A}} \tag{3.14}$$

and

$$Q_d = \sqrt{2qN_A\varepsilon_{si}\phi} \tag{3.15}$$

with N_A the substrate doping and ϕ the voltage across the depletion layer (i.e., the potential at the oxide-silicon boundary).

As the gate voltage increases, the potential at the silicon surface at some point reaches a critical value, where the semiconductor surface inverts to *n*-type material. This

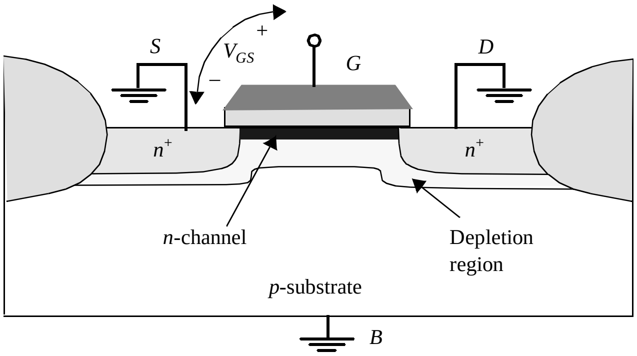


Figure 3.13 NMOS transistor for positive V_{GS}, showing depletion region and induced channel.

point marks the onset of a phenomenon known as *strong inversion* and occurs at a voltage equal to twice the *Fermi Potential* (Eq. (3.16)) ($\phi_F \approx -0.3$ V for typical *p*-type silicon substrates):

$$\phi_F = -\phi_T \ln(\frac{N_A}{n_i}) \qquad (3.16)$$

Further increases in the gate voltage produce no further changes in the depletion-layer width, but result in additional electrons in the thin inversion layer directly under the oxide. These are drawn into the inversion layer from the heavily doped *n*+ source region. Hence, a continuous *n*-type channel is formed between the source and drain regions, the conductivity of which is modulated by the gate-source voltage.

In the presence of an inversion layer, the charge stored in the depletion region is fixed and equals

$$Q_{B0} = \sqrt{2qN_A\varepsilon_{si}|-2\phi_F|} \qquad (3.17)$$

This picture changes somewhat in case a substrate bias voltage V_{SB} is applied (V_{SB} is normally positive for *n*-channel devices). This causes the surface potential required for strong inversion to increase and to become $|-2\phi_F + V_{SB}|$. The charge stored in the depletion region now is expressed by Eq. (3.18)

$$Q_B = \sqrt{2qN_A\varepsilon_{si}(|-2\phi_F + V_{SB}|)} \qquad (3.18)$$

The value of V_{GS} where strong inversion occurs is called the *threshold voltage* V_T. V_T is a function of several components, most of which are material constants such as the difference in work-function between gate and substrate material, the oxide thickness, the Fermi voltage, the charge of impurities trapped at the surface between channel and gate oxide, and the dosage of ions implanted for threshold adjustment. From the above arguments, it has become clear that the source-bulk voltage V_{SB} has an impact on the threshold. as well. Rather than relying on a complex (and hardly accurate) analytical expression for the threshold, we rely on an empirical parameter called V_{T0}, which is the threshold voltage for $V_{SB} = 0$, and is mostly a function of the manufacturing process. The threshold voltage under different body-biasing conditions can then be determined in the following manner,

$$V_T = V_{T0} + \gamma(\sqrt{|-2\phi_F + V_{SB}|} - \sqrt{|-2\phi_F|}) \qquad (3.19)$$

The parameter γ (gamma) is called the *body-effect coefficient*, and expresses the impact of changes in V_{SB}. Observe that the threshold voltage has a **positive** value for a typical **NMOS** device, while it is **negative** for a normal **PMOS** transistor.

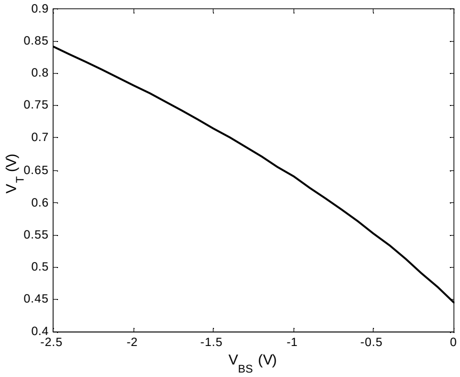


Figure 3.14 Effect of body-bias on theshold.

The effect of the well bias on the threshold voltage of an NMOS transistor is plotted in for typical values of $|-2\phi_F|$ = 0.6 V and γ = 0.4 $V^{0.5}$. A negative bias on the well or substrate causes the threshold to increase from 0.45 V to 0.85 V. Note also that V_{SB} always has to be larger than -0.6 V in an NMOS. If not, the source-body diode becomes forward biased, which deteriorates the transistor operation.

Example 3.5 Threshold Voltage of a PMOS Transistor

An PMOS transistor has a threshold voltage of -0.4 V, while the body-effect coefficient equals -0.4. Compute the threshold voltage for V_{SB} = -2.5 V. $2\phi_F$ = 0.6 V.

Using Eq. (3.19), we obtain $V_T(-2.5\text{ V}) = -0.4 - 0.4 \times ((2.5+0.6)^{0.5} - 0.6^{0.5})$ V = -0.79 V, which is twice the threshold under zero-bias conditions!

Resistive Operation

Assume now that $V_{GS} > V_T$ and that a small voltage, V_{DS}, is applied between drain and source. The voltage difference causes a current I_D to flow from drain to source (Figure 3.15). Using a simple analysis, a first-order expression of the current as a function of V_{GS} and V_{DS} can be obtained.

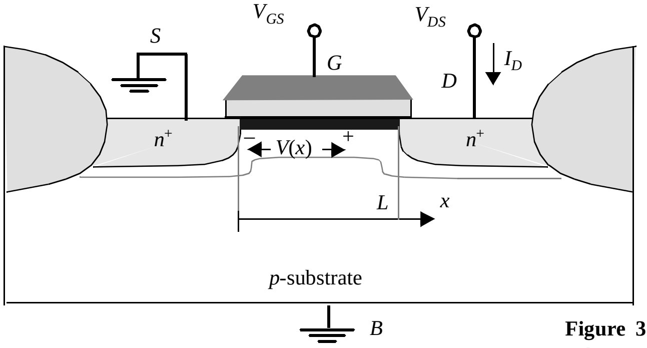


Figure 3.15 NMOS transistor with bias voltages.

At a point x along the channel, the voltage is $V(x)$, and the gate-to-channel voltage at that point equals $V_{GS} - V(x)$. Under the assumption that this voltage exceeds the threshold voltage all along the channel, the induced channel charge per unit area at point x can be computed.

$$Q_i(x) = -C_{ox}[V_{GS} - V(x) - V_T] \tag{3.20}$$

C_{ox} stands for the capacitance per unit area presented by the gate oxide, and equals

$$C_{ox} = \frac{\varepsilon_{ox}}{t_{ox}} \tag{3.21}$$

with $\varepsilon_{ox} = 3.97 \times \varepsilon_o = 3.5 \times 10^{-11}$ F/m the oxide permittivity, and t_{ox} is the thickness of the oxide. The latter which is 10 nm (= 100 Å) or smaller for contemporary processes. For an oxide thickness of 5 nm, this translates into an oxide capacitance of 7 fF/μm^2.

The current is given as the product of the drift velocity of the carriers υ_n and the available charge. Due to charge conservation, it is a constant over the length of the channel. *W* is the width of the channel in a direction perpendicular to the current flow.

$$I_D = -\upsilon_n(x)Q_i(x)W \tag{3.22}$$

The electron velocity is related to the electric field through a parameter called the *mobility* μ_n (expressed in m^2/V·s). The mobility is a complex function of crystal structure, and local electrical field. In general, an empirical value is used.

$$\upsilon_n = -\mu_n\xi(x) = \mu_n\frac{dV}{dx} \tag{3.23}$$

Combining Eq. (3.20) – Eq. (3.23) yields

$$I_D dx = \mu_n C_{ox} W(V_{GS} - V - V_T)dV \tag{3.24}$$

Integrating the equation over the length of the channel *L* yields the voltage-current relation of the transistor.

$$I_D = k'_n\frac{W}{L}\left[(V_{GS} - V_T)V_{DS} - \frac{V_{DS}^2}{2}\right] = k_n\left[(V_{GS} - V_T)V_{DS} - \frac{V_{DS}^2}{2}\right] \tag{3.25}$$

k'_n is called the *process transconductance parameter* and equals

$$k'_n = \mu_n C_{ox} = \frac{\mu_n\varepsilon_{ox}}{t_{ox}} \tag{3.26}$$

The product of the process transconductance k'_n and the (*W/L*) ratio of an (NMOS) transistor is called the *gain factor* k_n of the device. For smaller values of V_{DS}, the quadratic factor in Eq. (3.25) can be ignored, and we observe a linear dependence between V_{DS} and I_D. The operation region where Eq. (3.25) holds is hence called the *resistive* or *linear* region. One of its main properties is that it displays a continuous conductive channel between source and drain regions.

NOTICE: The *W* and *L* parameters in Eq. (3.25) represent the *effective channel width and length* of the transistor. These values differ from the dimensions *drawn* on the layout due to effects such as lateral diffusion of the source and drain regions (*L*), and the encroachment of the isolating field oxide (*W*). In the remainder of the text, *W* and *L* will always stand for the effective dimensions, while a *d* subscript will be used to indicate the drawn size. The following expressions related the two parameters, with ΔW and ΔL parameters of the manufacturing process:

$$\begin{aligned} W &= W_d - \Delta W \\ L &= L_d - \Delta L \end{aligned} \tag{3.27}$$

The Saturation Region

As the value of the drain-source voltage is further increased, the assumption that the channel voltage is larger than the threshold all along the channel ceases to hold. This happens when $V_{GS} - V(x) < V_T$. At that point, the induced charge is zero, and the conducting channel disappears or is *pinched off*. This is illustrated in Figure 3.16, which shows (in an exaggerated fashion) how the channel thickness gradually is reduced from source to drain until pinch-off occurs. No channel exists in the vicinity of the drain region. Obviously, for this phenomenon to occur, it is essential that the pinch-off condition be met at the drain region, or

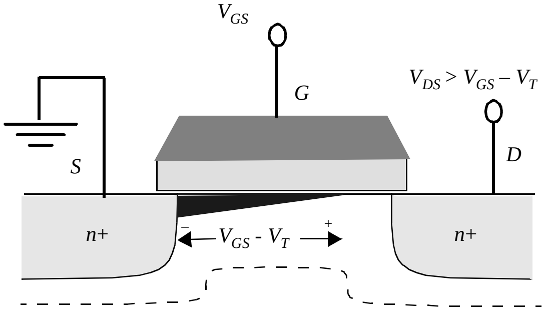


Figure 3.16 NMOS transistor under pinch-off conditions.

$$V_{GS} - V_{DS} \leq V_T. \tag{3.28}$$

Under those circumstances, the transistor is in the *saturation* region, and Eq. (3.25) no longer holds. The voltage difference over the induced channel (from the pinch-off point to the source) remains fixed at $V_{GS} - V_T$, and consequently, the current remains constant (or saturates). Replacing V_{DS} by $V_{GS} - V_T$ in Eq. (3.25) yields the drain current for the saturation mode. It is worth observing that, to a first agree, the current is no longer a function of V_{DS}. Notice also the *squared dependency* of the drain current with respect to the control voltage V_{GS}.

$$I_D = \frac{k'_n}{2}\frac{W}{L}(V_{GS} - V_T)^2 \tag{3.29}$$

Channel-Length Modulation

The latter equation seems to suggest that the transistor in the saturation mode acts as a perfect current source — or that the current between drain and source terminal is a constant, independent of the applied voltage over the terminals. This not entirely correct. The effective length of the conductive channel is actually modulated by the applied V_{DS}: increasing V_{DS} causes the depletion region at the drain junction to grow, reducing the length of the effective channel. As can be observed from Eq. (3.29), the current increases when the length factor L is decreased. A more accurate description of the current of the MOS transistor is therefore given in Eq. (3.30).

$$I_D = I_D'(1 + \lambda V_{DS}) \tag{3.30}$$

with I_D' the current expressions derived earlier, and λ an empirical parameter, called the *channel-length modulation*. Analytical expressions for λ have proven to be complex and inaccurate. λ varies roughly with the inverse of the channel length. In shorter transistors, the drain-junction depletion region presents a larger fraction of the channel, and the channel-modulation effect is more pronounced. It is therefore advisable to resort to long-channel transistors if a high-impedance current source is needed.

Velocity Saturation

The behavior of transistors with very short channel lengths (called *short-channel devices*) deviates considerably from the resistive and saturated models, presented in the previous paragraphs. The main culprit for this deficiency is the *velocity saturation* effect. Eq. (3.23) states that the velocity of the carriers is proportional to the electrical field, independent of the value of that field. In other words, the carrier mobility is a constant. However, at high field strengths, the carriers fail to follow this linear model. In fact, when the electrical field along the channel reaches a critical value ξ_c, the velocity of the carriers tends to saturate due to scattering effects (collisions suffered by the carriers). This is illustrated in Figure 3.17.

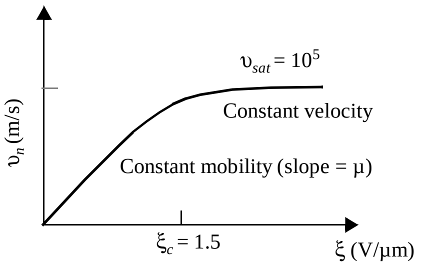


Figure 3.17 Velocity-saturation effect.

For *p*-type silicon, the critical field at which electron saturation occurs is around 1.5×10^6 V/m (or 1.5 V/µm), and the saturation velocity υ_{sat} approximately equals 10^5 m/s. This means that in an NMOS device with a channel length of 1 µm, only a couple of volts between drain and source are needed to reach the saturation point. This condition is easily met in current short-channel devices. Holes in a *n*-type silicon saturate at the same velocity, although a higher electrical field is needed to achieve saturation. Velocity-saturation effects are hence less pronounced in PMOS transistors.

This effect has a profound impact on the operation of the transistor. We will illustrate this with a first-order derivation of the device characteristics under velocity-saturating conditions [Ko89]. The velocity as a function of the electrical field, plotted in Figure 3.17, can be roughly approximated by the following expression:

$$\begin{aligned} \upsilon &= \frac{\mu_n \xi}{1 + \xi / \xi_c} \quad \text{for} \quad \xi \leq \xi_c \\ &= \upsilon_{sat} \qquad \text{for} \quad \xi \geq \xi_c \end{aligned} \tag{3.31}$$

The continuity requirement between the two regions dictates that $\xi_c = 2\upsilon_{sat}/\mu_n$. Re-evaluation of Eq. (3.20) and Eq. (3.22) in light of revised velocity formula leads to a modified expression of the drain current in the resistive region:

$$I_D = \kappa(V_{DS}) \mu_n C_{ox} \frac{W}{L} \left[(V_{GS} - V_T) V_{DS} - \frac{V_{DS}^2}{2} \right] \tag{3.32}$$

with

$$\kappa(V) = \frac{1}{1 + (V / \xi_c L)} \tag{3.33}$$

κ is a measure of the degree of velocity saturation, since V_{DS}/L can be interpreted as the average field in the channel. In case of long-channel devices (large values of L) or small values of V_{DS}, κ approaches 1 and Eq. (3.32) simplifies to the traditional current equation for the resistive operation mode. For short-channel devices, κ is smaller than 1, which means that the delivered current is smaller than what would be normally expected.

When increasing the drain-source voltage, the electrical field in the channel will ultimately reach the critical value, and the carriers at the drain become velocity saturated. The saturation drain voltage V_{DSAT} can be calculated by equating the current at the drain to the current given by Eq. (3.32) for $V_{DS} = V_{DSAT}$. The former is derived from Eq. (3.22), assuming that the drift velocity is saturated and equals υ_{sat}.

$$\begin{aligned} I_{DSAT} &= \upsilon_{sat} C_{ox} W (V_{GT} - V_{DSAT}) \\ &= \kappa(V_{DSAT}) \mu_n C_{ox} \frac{W}{L} \left[V_{GT} V_{DSAT} - \frac{V_{DSAT}^2}{2} \right] \end{aligned} \tag{3.34}$$

V_{GT} is a shorthand notation for V_{GS} - V_T. After some algebra, we obtain

$$V_{DSAT} = \kappa(V_{GT}) V_{GT} \tag{3.35}$$

Further increasing the drain-source voltage does not yield more current (to a first degree) and the transistor current saturates at I_{DSAT}. This leads to some interesting observations:

- For a short-channel device and for large enough values of V_{GT}, $\kappa(V_{GT})$ is substantially smaller than 1, hence $V_{DSAT} < V_{GT}$. The device enters saturation before V_{DS} reaches V_{GS} - V_T. Short-channel devices therefore experience an extended saturation

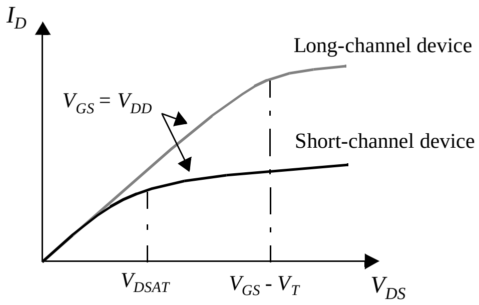


Figure 3.18 Short-channel devices display an extended saturation region due to velocity-saturation.

region, and tend to operate more often in saturation conditions than their long-channel counterparts, as is illustrated in Figure 3.18.

- The saturation current I_{DSAT} displays a *linear dependence* with respect to the gate-source voltage V_{GS}, which is in contrast with the squared dependence in the long-channel device. This reduces the amount of current a transistor can deliver for a given control voltage. On the other hand, reducing the operating voltage does not have such a significant effect in submicron devices as it would have in a long-channel transistor.

The equations above ignore that a larger portion of the channel becomes velocity-saturated with a further increase of V_{DS}. From a modeling perspective, it appears as though the effective channel is shortening with increasing V_{DS}, similar in effect to the channel-length modulation. The resulting increase in current is easily accommodated by introducing an extra $(1 + \lambda \times V_{DS})$ multiplier.

Thus far we have only considered the effects of the tangential field along the channel due to the V_{DS}, when considering velocity-saturation effects. However, there also exists a normal (vertical) field originating from the gate voltage that further inhibits channel carrier mobility. This effect, which is called *mobility degradation*, reduces the surface mobility with respect to the bulk mobility. Eq. (3.36) provides a simple estimation of the mobility reduction,

$$\mu_{n,\,eff} = \frac{\mu_{n0}}{1 + \eta(V_{GS} - V_T)} \tag{3.36}$$

with μ_{n0} the bulk mobility and η an empirical parameter. A typical approach is to use derive the actual value of μ for a given field strength from tables or empirical charts.

Readers interested in a more in-depth perspective on the short-channel effects in MOS transistors are referred to the excellent reference works on this topic, such as [Ko89].

Velocity Saturation — Revisited

Unfortunately, the drain-current equations Eq. (3.32) and Eq. (3.33) are complex expressions of V_{GS} and V_{DS}, which makes them rather unwieldy for a first-order manual analysis. A substantially simpler model can be obtained by making two assumptions:

1. The velocity saturates abruptly at ξ_c, and is approximated by the following expression:

$$\begin{aligned} \upsilon &= \mu_n \xi \qquad\qquad \text{for} \quad \xi \le \xi_c \\ &= \upsilon_{sat} = \mu_n \xi_c \quad \text{for} \quad \xi \ge \xi_c \end{aligned} \tag{3.37}$$

2. The drain-source voltage V_{DSAT} at which the critical electrical field is reached and velocity saturation comes into play is *constant* and is approximated by Eq. (3.38). From Eq. (3.35), it can be observed that this assumption is reasonable for larger values of V_{GT} (>> $\xi_c L$).

$$V_{DSAT} = L\xi_c = \frac{L\upsilon_{sat}}{\mu_n} \tag{3.38}$$

Under these circumstances, the current equations for the resistive region remain unchanged from the long-channel model. Once V_{DSAT} is reached, the current abruptly saturates. The value for I_{DSAT} at that point can be derived by plugging the saturation voltage into the current equation for the resistive region (Eq. (3.25)).

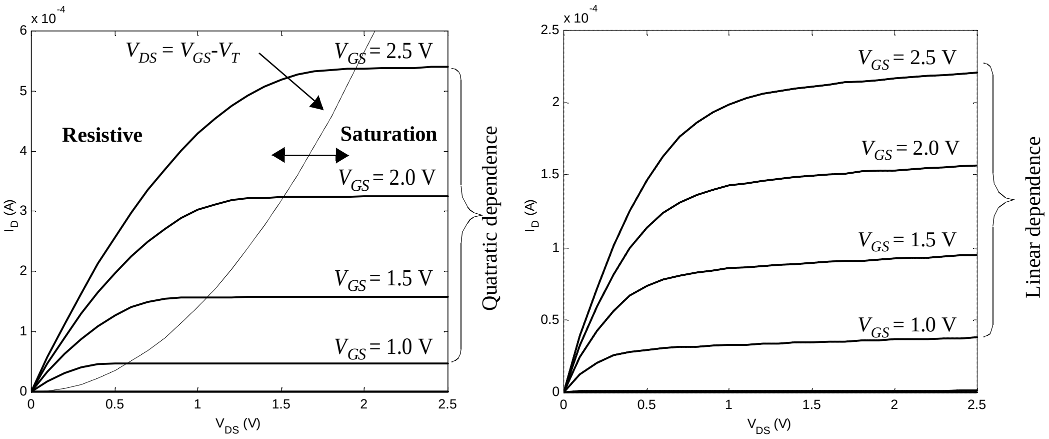
$$\begin{aligned} I_{DSAT} &= I_D(V_{DS} = V_{DSAT}) \\ &= \mu_n C_{ox} \frac{W}{L}\left((V_{GS} - V_T)V_{DSAT} - \frac{V_{DSAT}^2}{2}\right) \\ &= \upsilon_{sat} C_{ox} W\left(V_{GS} - V_T - \frac{V_{DSAT}}{2}\right) \end{aligned} \tag{3.39}$$

This model is truly first-order and empirical. The simplified velocity model causes substantial deviations in the transition zone between linear and velocity-saturated regions. Yet, by carefully choosing the model parameters, decent matching can be obtained with empirical data in the other operation regions, as will be shown in one of the following sections. Most importantly, the equations are coherent with the familiar long-channel equations, and provide the digital designer with a much needed tool for intuitive understanding and interpretation.

Drain Current versus Voltage Charts

The behavior for the MOS transistor in the different operation regions is best understood by analyzing its I_D-V_{DS} curves, which plot I_D versus V_{DS} with V_{GS} as a parameter. Figure 3.19 shows these charts for two NMOS transistors, implemented in the same technology and with the same W/L ratio. One would hence expect both devices to display identical *I-V* characteristics, The main difference however is that the first device has a long channel length (L_d = 10 μm), while the second transistor is a short channel device (L_d = 0.25 μm), and experiences velocity saturation.

Consider first the long-channel device. In the resistive region, the transistor behaves like a voltage-controlled resistor, while in the saturation region, it acts as a voltage-controlled current source (when the channel-length modulation effect is ignored). The transi-

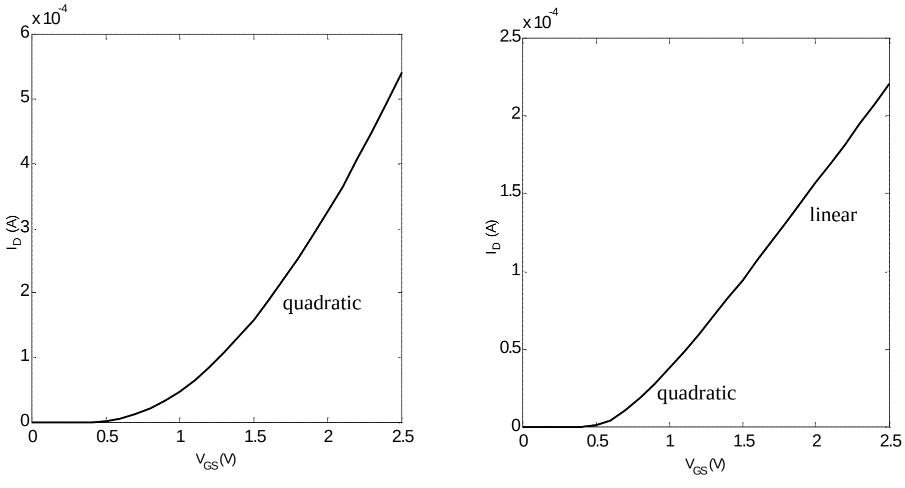


(a) Long-channel transistor (L_d = 10 µm)

(b) Short-channel transistor (L_d = 0.25 µm)

Figure 3.19 *I-V* characteristics of long- and a short-channel NMOS transistors in a 0.25 µm CMOS technology. The (*W*/*L*) ration of both transistors is identical and equals 1.5

tion between both regions is delineated by the V_{DS} = V_{GS} - V_T curve. The squared dependence of I_D as a function of V_{GS} in the saturation region — typical for a long channel device — is clearly observable from the spacing between the different curves. The linear dependence of the saturation current with respect to VGS is apparent in the short-channel device of b. Notice also how velocity-saturation causes the device to saturate for substantially smaller values of V_{DS}. This results in a substantial drop in current drive for high voltage levels. For instance, at (V_{GS} = 2.5 V, V_{DS} = 2.5 V), the drain current of the short transistor is only 40% of the corresponding value of the longer device (220 µA versus 540 µA).



(a) Long-channel device (L_d = 10 µm)

(b) Short-channel device (L_d = 0.25 µm)

Figure 3.20 NMOS transistor I_D-V_{GS} characteristic for long and short-channel devices (0.25 µm CMOS technology). *W*/*L* = 1.5 for both transistors and V_{DS} = 2.5 V.

The difference in dependence upon V_{GS} between long- and short-channel devices is even more pronounced in another set of simulated charts that plot I_D as a function of V_{GS} for a fixed value of V_{DS} ($\geq V_{GS}$ — hence ensuring saturation) (Figure 3.20). A quadratic versus linear dependence is apparent for larger values of V_{GS}.

All the derived equations hold for the PMOS transistor as well. The only difference is that **for PMOS devices, the polarities of all voltages and currents are reversed**. This is illustrated in Figure 3.21, which plots the I_D-V_{DS} characteristics of a minimum-size PMOS transistor in our generic 0.25 μm CMOS process. The curves are in the third quadrant as I_D, V_{DS}, and V_{GS} are all negative. Interesting to observe is also that the effects of velocity saturation are less pronounced than in the CMOS devices. This can be attributed to the higher value of the critical electrical field, resulting from the smaller mobility of holes versus electrons.

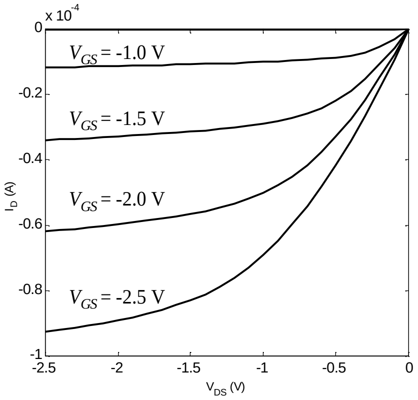


Figure 3.21 *I-V* characteristics of (W_d=0.375 μm, L_d=0.25 μm) PMOS transistor in 0.25 μm CMOS process. Due to the smaller mobility, the maximum current is only 42% of what is achieved by a similar NMOS transistor.

Subthreshold Conduction

A closer inspection of the I_D-V_{GS} curves of Figure 3.20 reveals that the current does not drop abruptly to 0 at $V_{GS} = V_T$. It becomes apparent that the MOS transistor is already partially conducting for voltages below the threshold voltage. This effect is called *subthreshold* or *weak-inversion* conduction. The onset of strong inversion means that ample carriers are available for conduction, but by no means implies that no current at all can flow for gate-source voltages below V_T, although the current levels are small under those conditions. The transition from the on- to the off-condition is thus not abrupt, but gradual.

To study this effect in somewhat more detail, we redraw the I_D versus V_{GS} curve of Figure 3.20b on a logarithmic scale as shown in Figure 3.22. This confirms that the current does not drop to zero immediately for $V_{GS} < V_T$, but actually decays in an exponential fashion, similar to the operation of a bipolar transistor.[2] In the absence of a conducting channel, the n^+ (source) - p (bulk) - n^+ (drain) terminals actually form a parasitic bipolar transistor. The current in this region can be approximated by the expression

[2] Discussion of the operation of bipolar transistors is out of the scope of this textbook. We refer to various textbooks on semiconductor devices, or to the additional information that is available on the web-site of this book.

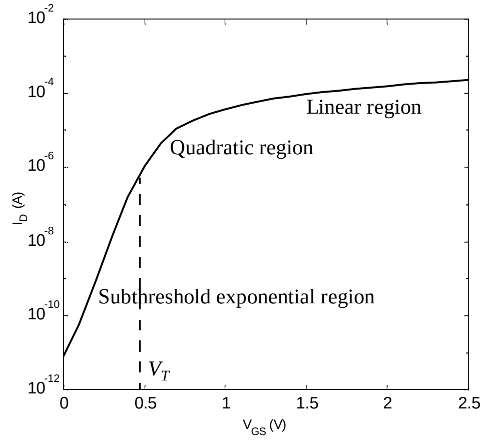


Figure 3.22 I_D current versus V_{GS} (on logarithmic scale), showing the exponential characteristic of the subthreshold region.

$$I_D = I_S \mathrm{e}^{\frac{V_{GS}}{nkT/q}}\left(1 - \mathrm{e}^{-\frac{V_{DS}}{kT/q}}\right) \tag{3.40}$$

where I_S and n are empirical parameters, with $n \geq 1$ and typically ranging around 1.5.

In most digital applications, the presence of subthreshold current is undesirable as it detracts from the ideal switch-like behavior that we like to assume for the MOS transistor. We would rather have the current drop as fast as possible once the gate-source voltage falls below V_T. The (inverse) rate of decline of the current with respect to V_{GS} below V_T hence is a quality measure of a device. It is often quantified by the *slope factor S*, which measures by how much V_{GS} has to be reduced for the drain current to drop by a factor of 10. From Eq. (3.40), we find

$$S = n\left(\frac{kT}{q}\right)\ln(10) \tag{3.41}$$

with S is expressed in mV/decade. For an ideal transistor with the sharpest possible roll-off, $n = 1$ and $(kT/q)\ln(10)$ evaluates to 60 mV/decade at room temperature, which means that the subthreshold current drops by a factor of 10 for a reduction in V_{GS} of 60 mV. Unfortunately, n is larger than 1 for actual devices and the current falls at a reduced rate (90 mV/decade for $n = 1.5$). The current roll-off is further affected in a negative sense by an increase in the operating temperature (most integrated circuits operate at temperatures considerably beyond room temperature). The value of n is determined by the intrinsic device topology and structure. Reducing its value hence requires a different process technology, such as silicon-on-insulator.

Subthreshold current has some important repercussions. In general, we want the current through the transistor to be as close as possible to zero at $V_{GS} = 0$. This is especially important in the so-called *dynamic circuits*, which rely on the storage of charge on a capacitor and whose operation can be severely degraded by subthreshold leakage. Achieving this in the presence of subthreshold current requires a firm lower bound on the value of the threshold voltage of the devices.

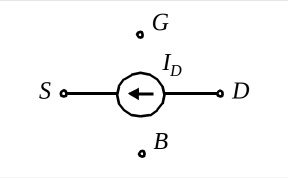
Example 3.6 Subthreshold Slope

For the example of Figure 3.22, a slope of 89.5 mV/decade is observed (between 0.2 and 0.4 V). This is equivalent to an *n*-factor of 1.49.

In Summary – Models for Manual Analysis

The preceding discussions made it clear that the deep-submicron transistor is a complex device. Its behavior is heavily non-linear and is influenced by a large number of second-order effects. Fortunately, accurate circuit-simulation models have been developed that make it possible to predict the behavior of a device with amazing precision over a large range of device sizes, shapes, and operation modes, as we will discuss later in this chapter. While excellent from an accuracy perspective, these models fail in providing a designer with an intuitive insight in the behavior of a circuit and its dominant design parameters. Such an understanding is necessary in the design analysis and optimization process. A designer who misses a clear vision on what drives and governs the circuit operation by necessity resorts on a lengthy trial by error optimization process, that most often leads to an inferior solution.

The obvious question is now how to abstract the behavior of our MOS transistor into a simple and tangible analytical model that does not lead to hopelessly complex equations, yet captures the essentials of the device. It turns out that the first-order expressions, derived earlier in the chapter, can be combined into a single expression that meets these goals. The model presents the transistor as a single current source (Figure 3.23), the value of which is given defined in the Figure. The reader can verify that, depending upon the operating condition, the model simplifies into either Eq. (3.25), Eq. (3.29), or Eq. (3.39) (corrected for channel-length modulation), depending upon operating conditions..



$$I_D = 0 \text{ for } V_{GT} \le 0$$

$$I_D = k'\frac{W}{L}\left(V_{GT}V_{min} - \frac{V_{min}^2}{2}\right)(1 + \lambda V_{DS}) \text{ for } V_{GT} \ge 0$$

$$\text{with } V_{min} = \min(V_{GT}, V_{DS}, V_{DSAT}),$$

$$V_{GT} = V_{GS} - V_T,$$

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

Figure 3.23 A unified MOS model for manual analysis.

Besides being a function of the voltages at the four terminals of the transistor, the model employs a set of five parameters: V_{TO}, γ, V_{DSAT}, k', and λ. In principle, it would be possible to determine these parameters from the process technology and from the device physics equations. The complexity of the device makes this a precarious task. A more rewarding approach is to choose the values such that a good matching with the actual device characteristics is obtained. More significantly, the model should match the best in the regions that matter the most. In digital circuits, this in the region of high V_{GS} and V_{DS}. The performance of an MOS digital circuit is primarily determined by the maximum available current (i.e., the current obtained for V_{GS} = V_{DS} = supply voltage). A good matching in this region is therefore essential.

Example 3.7 Manual Analysis Model for 0.25 μm CMOS Process[3]

Based on the simulated I_D-V_{DS} and I_D-V_{GS} plots of a (W_d = 0.375 μm, L_d = 0.25 μm) transistor, implemented in our generic 0.25 micron CMOS process (Figure 3.19, Figure 3.20), we have derived a set of device parameters to match well in the (V_{DS} = 2.5 V, V_{GS} = 2.5 V) region — 2.5 V being the typical supply voltage for this process. The resulting characteristics are plotted in Figure 3.24 for the NMOS transistor, and compared to the simulated values. Overall, a good correspondence can be observed with the exception of the transition region between resistive and velocity-saturation. This discrepancy, which is due the simple velocity model of Eq. (3.37) as explained earlier, is acceptable as it occurs in the lower value-range of V_{DS}. It demonstrates that our model, while simple, manages to give a fair indication of the overall behavior of the device.

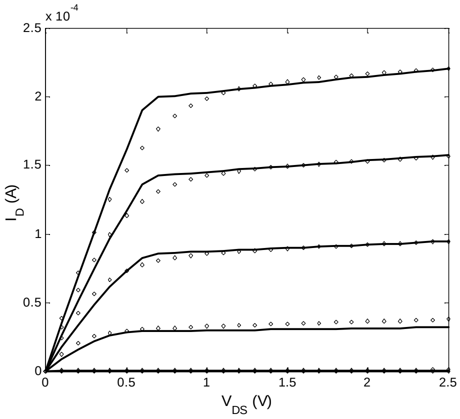


Figure 3.24 Correspondence between simple model (solid line) and SPICE simulation (dotted) for minimum-size NMOS transistor (W_d=0.375 μm, L_d=0.25 μm). Observe the discrepancy in the transition zone between resistive and velocity saturation.

Design Data — Transistor Model for Manual Analysis

Table 3.2 tabulates the obtained parameter values for the minimum-sized NMOS and a similarly sized PMOS device in our generic 0.25 μm CMOS process. These values will be used as generic model-parameters in later chapters.

Table 3.2 Parameters for manual model of generic 0.25 μm CMOS process (minimum length device).

	V_{T0} (V)	γ ($V^{0.5}$)	V_{DSAT} (V)	k' (A/V^2)	λ (V^{-1})
NMOS	0.43	0.4	0.63	115×10^{-6}	0.06
PMOS	–0.4	-0.4	-1	-30×10^{-6}	-0.1

[3] A MATLAB implementation of the model is available on the web-site of the textbook.

A word of caution — The model presented here is derived from the characteristics of a single device with a minimum channel-length and width. Trying to extrapolate this behavior to devices with substantially different values of W and L will probably lead to sizable errors. Fortunately, digital circuits typically use only minimum-length devices as these lead to the smallest implementation area. Matching for these transistors will typically be acceptable. It is however advisable to use a different set of model parameters for devices with dramatically different size- and shape-factors.

The presented current-source model will prove to be very useful in the analysis of the basic properties and metrics of a simple digital gate, yet its non-linearity makes it intractable for anything that is somewhat more complex. We therefore introduce an even more simplified model that has the advantage of being linear and straightforward. It is based on the underlying assumption in most digital designs that the transistor is nothing more than a switch with an infinite off-resistance, and a finite on-resistance R_{on}.

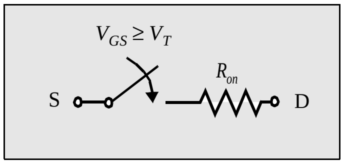


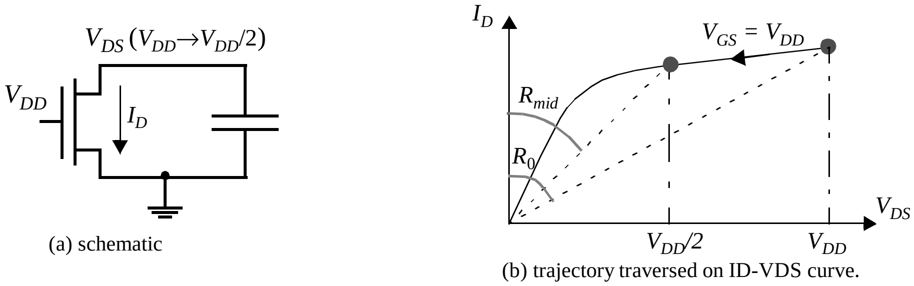
Figure 3.25 NMOS transistor modeled as a switch.

The main problem with this model is that R_{on} is still time-variant, non-linear and depending upon the operation point of the transistor. When studying digital circuits in the transient mode — which means while switching between different logic states — it is attractive to assume R_{on} as a constant and linear resistance R_{eq}, chosen so that the final result is similar to what would be obtained with the original transistor. A reasonable approach in that respect is to use the average value of the resistance over the operation region of interest, or even simpler, the average value of the resistances at the end-points of the transition. The latter assumption works well if the resistance does not experience any strong non-linearities over the range of the averaging interval.

$$\begin{aligned} R_{eq} &= \text{average}_{t = t_1 \ldots t_2}(R_{on}(t)) = \frac{1}{t_2 - t_1}\int_{t_1}^{t_2} R_{on}(t)dt = \frac{1}{t_2 - t_1}\int_{t_1}^{t_2} \frac{V_{DS}(t)}{I_D(t)}dt \\ &\approx \frac{1}{2}(R_{on}(t_1) + R_{on}(t_2)) \end{aligned} \tag{3.42}$$

Example 3.8 Equivalent resistance when (dis)charging a capacitor

One of the most common scenario's in contemporary digital circuits is the discharging of a capacitor from V_{DD} to GND through an NMOS transistor with its gate voltage set to V_{DD}, or vice-versa the charging of the capacitor to V_{DD} through a PMOS with its gate at GND. Of special interest is the point where the voltage on the capacitor reaches the mid-point ($V_{DD}/2$) — this is by virtue of the definition of the propagation delay as introduced in Chapter 2. Assuming that the supply voltage is substantially larger than the velocity-saturation voltage V_{DSAT} of the transistor, it is fair to state that the transistor stays in velocity saturation for the entire duration of the transition. This scenario is plotted in for the case of an NMOS discharging a capacitor from V_{DD} to $V_{DD}/2$.



(a) schematic

(b) trajectory traversed on ID-VDS curve.

Figure 3.26 Discharging a capacitor through an NMOS transistor: Schematic (a) and *I-V* trajectory (b). The instantianous resistance of the transistor equals (V_{DS}/I_D) and is visualized by the angle with respect to the *y*-axis.

With the aid of Eq. (3.42) and Eq. , we can derive the value of the equivalent resistance, which averages the resistance of the device over the interval.

$$R_{eq} = \frac{1}{-V_{DD}/2}\int_{V_{DD}}^{V_{DD}/2} \frac{V}{I_{DSAT}(1+\lambda V)}dV \approx \frac{3}{4}\frac{V_{DD}}{I_{DSAT}}\left(1-\frac{7}{9}\lambda V_{DD}\right)$$
$$\text{with} \quad I_{DSAT} = k'\frac{W}{L}\left((V_{DD}-V_T)V_{DSAT}-\frac{V_{DSAT}^2}{2}\right) \tag{3.43}$$

A similar result can be obtained by just averaging the values of the resistance at the end points (and simplifying the result using a Taylor expansion):

$$R_{eq} = \frac{1}{2}\left(\frac{V_{DD}}{I_{DSAT}(1+\lambda V_{DD})}+\frac{V_{DD}/2}{I_{DSAT}(1+\lambda V_{DD}/2)}\right) \approx \frac{3}{4}\frac{V_{DD}}{I_{DSAT}}\left(1-\frac{5}{6}\lambda V_{DD}\right) \tag{3.44}$$

A number of conclusions are worth drawing from the above expressions:

- The resistance is inversely proportional to the (*W*/*L*) ratio of the device. Doubling the transistor width halves the resistance.
- For $V_{DD} >> V_T + V_{DSAT}/2$, the resistance becomes virtually independent of the supply voltage. This is confirmed in halves, which plots the simulated equivalent resistance as a function of the supply voltage V_{DD}. Only a minor improvement in resistance, attributable to the channel-length modulation, can be observed when raising the supply voltage.
- Once the supply voltage approaches V_T, a dramatic increase in resistance can be observed.

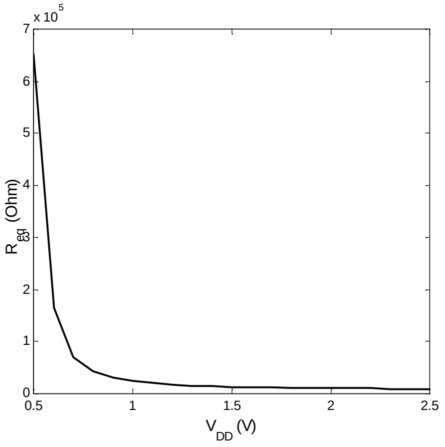


Figure 3.27 Simulated equivalent resistance of a minimum size NMOS transistor in 0.25 μm CMOS process as a function of V_{DD} ($V_{GS} = V_{DD}$, $V_{DS} = V_{DD} \rightarrow V_{DD}/2$).

Design Data — Equivalent Resistance Model

Table 3.3 enumerates the equivalent resistances obtained by simulation of our generic 0.25 μm CMOS process. These values will come in handy when analyzing the performance of CMOS gates in later chapters.

Table 3.3 Equivalent resistance R_{eq} (W/L= 1) of NMOS and PMOS transistors in 0.25 μm CMOS process (with $L = L_{min}$). For larger devices, divide R_{eq} by W/L.

V_{DD} **(V)**	1	1.5	2	2.5
NMOS (kΩ)	35	19	15	13
PMOS (kΩ)	115	55	38	31

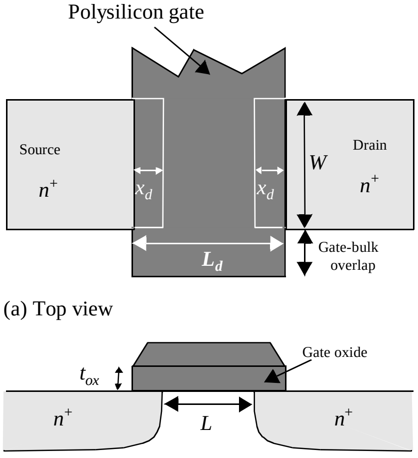
3.3.3 Dynamic Behavior

The dynamic response of a MOSFET transistor is a sole function of the time it takes to (dis)charge the parasitic capacitances that are intrinsic to the device, and the extra capacitance introduced by the interconnecting lines (and are the subject of Chapter 4). A profound understanding of the nature and the behavior of these intrinsic capacitances is essential for the designer of high-quality digital integrated circuits. They originate from three sources: the basic MOS structure, the channel charge, and the depletion regions of the reverse-biased *pn*-junctions of drain and source. Aside from the MOS structure capacitances, all capacitors are nonlinear and vary with the applied voltage, which makes their analysis hard. We discuss each of the components in turn.

MOS Structure Capacitances

The gate of the MOS transistor is isolated from the conducting channel by the gate oxide that has a capacitance per unit area equal to $C_{ox} = \varepsilon_{ox} / t_{ox}$. We learned earlier that from a *I-V* perspective it is useful to have C_{ox} as large as possible, or to keep the oxide thickness very thin. The total value of this capacitance is called the *gate capacitance* C_g and can be decomposed into two elements, each with a different behavior. Obviously, one part of C_g contributes to the channel charge, and is discussed in a subsequent section. Another part is solely due to the topological structure of the transistor. This component is the subject of the remainder of this section.

Consider the transistor structure of Figure 3.28. Ideally, the source and drain diffusion should end right at the edge of the gate oxide. In reality, both source and drain tend to extend somewhat below the oxide by an amount x_d, called the *lateral diffusion*. Hence, the effective channel of the transistor L becomes shorter than the drawn length L_d (or the length the transistor was originally designed for) by a factor of $\Delta L = 2x_d$. It also gives rise to a parasitic capacitance between gate and source (drain) that is called the *overlap capacitance.* This capacitance is strictly linear and has a fixed value



(a) Top view

(b) Cross section

Figure 3.28 MOSFET overlap capacitance.

$$C_{GSO} = C_{GDO} = C_{ox}x_dW = C_oW \tag{3.45}$$

Since x_d is a technology-determined parameter, it is customary to combine it with the oxide capacitance to yield the overlap capacitance per unit transistor width C_o (more specifically, C_{gso} and C_{gdo}).

Channel Capacitance

Perhaps the most significant MOS parasitic circuit element, the gate-to-channel capacitance C_{GC} varies in both magnitude and in its division into three components C_{GCS}, C_{GCD}, and C_{GCB} (being the gate-to-source, gate-to-drain, and gate-to-body capacitances, respectively), depending upon the operation region and terminal voltages. This varying distribution is best explained with the simple diagrams of Figure 3.29. When the transistor is in cut-off (a), no channel exists, and the total capacitance C_{GC} appears between gate and body. In the resistive region (b), an inversion layer is formed, which acts as a conductor between source and drain. Consequently, $C_{GCB} = 0$ as the body electrode is shielded from the gate by the channel. Symmetry dictates that the capacitance distributes evenly between source and drain. Finally, in the saturation mode (c), the channel is pinched off. The capacitance between gate and drain is approximately zero, and so is the gate-body capacitance. All the capacitance hence is between gate and source.

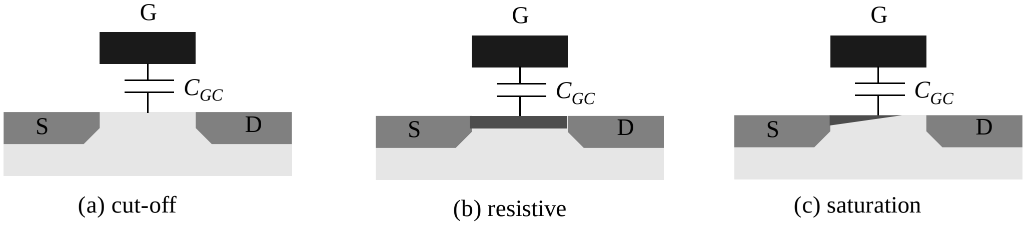


Figure 3.29 The gate-to-channel capacitance and how the operation region influences is distribution over the three other device terminals.

To actual value of the total gate-channel capacitance and its distribution over the three components is best understood with the aid of a number of charts. The first plot (Figure 3.30a) captures the evolution of the capacitance as a function of V_{GS} for $V_{DS} = 0$. For $V_{GS} = 0$, the transistor is off, no channel is present and the total capacitance, equal to WLC_{ox}, appears between gate and body. When increasing V_{GS}, a depletion region forms under the gate. This seemingly causes the thickness of the gate dielectric to increase, which means a reduction in capacitance. Once the transistor turns on ($V_{GS} = V_T$), a channel is formed and C_{GCB} drops to 0. With $V_{DS} = 0$, the device operates in the resistive mode and the capacitance divides equally between source and drain, or $C_{GCS} = C_{GCD} = WLC_{ox}/2$. The large fluctuation of the channel capacitance around $V_{GS}=V_T$ is worth remembering. A designer looking for a well-behaved linear capacitance should avoid operation in this region.

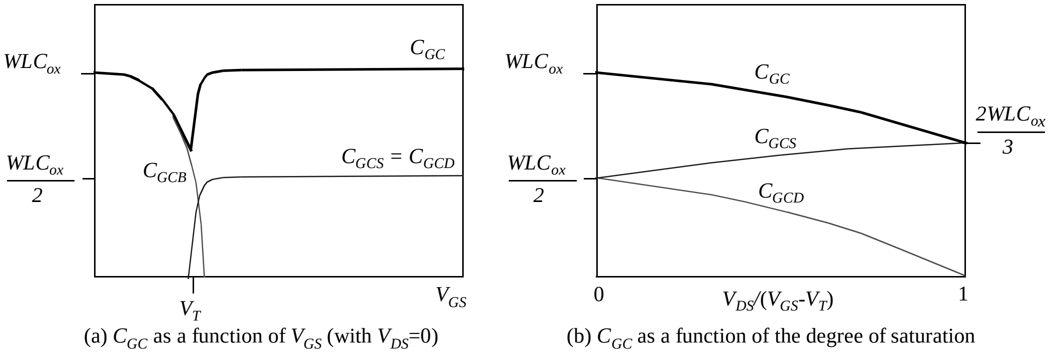


Figure 3.30 Distribution of the gate-channel capacitance as a function of V_{GS} and V_{DS} (from [Dally98]).

Once the transistor is on, the distribution of its gate capacitance depends upon the degree of saturation, measured by the $V_{DS}/(V_{GS}-V_T)$ ratio. As illustrated in Figure 3.30b,

C_{GCD} gradually drops to 0 for increasing levels of saturation, while C_{GCS} increases to 2/3 $C_{ox}WL$. This also means that the total gate capacitance is getting smaller with an increased level of saturation.

From the above, it becomes clear that the gate-capacitance components are nonlinear and varying with the operating voltages. To make a first-order analysis possible, we will use a simplified model with a constant capacitance value in each region of operation in the remainder of the text. The assumed values are summarized in Table 3.4.

Table 3.4 Average distribution of channel capacitance of MOS transistor for different operation regions.

Operation Region	C_{GCB}	C_{GCS}	C_{GCD}	C_{GC}	C_G
Cutoff	$C_{ox}WL$	0	0	$C_{ox}WL$	$C_{ox}WL+2C_oW$
Resistive	0	$C_{ox}WL$ / 2	$C_{ox}WL$ / 2	$C_{ox}WL$	$C_{ox}WL+2C_oW$
Saturation	0	(2/3)$C_{ox}WL$	0	(2/3)$C_{ox}WL$	(2/3)$C_{ox}WL+2C_oW$

Example 3.9 Using a circuit simulator to extract capacitance

Determining the value of the parasitic capacitances of an MOS transistor for a given operation mode is a labor-intensive task, and requires the knowledge of a number of technology parameters that are often not explicitly available. Fortunately, once a SPICE model of the transistor is attained, a simple simulation can give you the data you are interested in. Assume we would like to know the value of the total gate capacitance of a transistor in a given technology as a function of V_{GS} (for V_{DS} = 0). A simulation of the circuit of Figure 3.31a will give us exactly this information. In fact, the following relation is valid:

$$I = C_G(V_{GS})\frac{dV_{GS}}{dt}$$

which can be rewritten to yield an expression for C_G.

$$C_G(V_{GS}) = I/\left(\frac{dV_{GS}}{dt}\right)$$

A transient simulation gives us V_{GS} as a function of time, which can be translated into the capacitance with the aid of some simple mathematical manipulations. This is demonstrated in Figure 3.31b, which plots the simulated gate capacitance of a minimum size 0.25 μm NMOS transistor as a function of V_{GS}. The graphs clearly shows the drop of the capacitance when V_{GS} approaches V_T and the discontinuity at V_T, predicted in Figure 3.30.

Junction Capacitances

A final capacitive component is contributed by the reverse-biased source-body and drain-body *pn*-junctions. The depletion-region capacitance is nonlinear and decreases when the reverse bias is raised as discussed earlier. To understand the components of the junction capacitance (often called the *diffusion capacitance*), we must look at the source (drain)

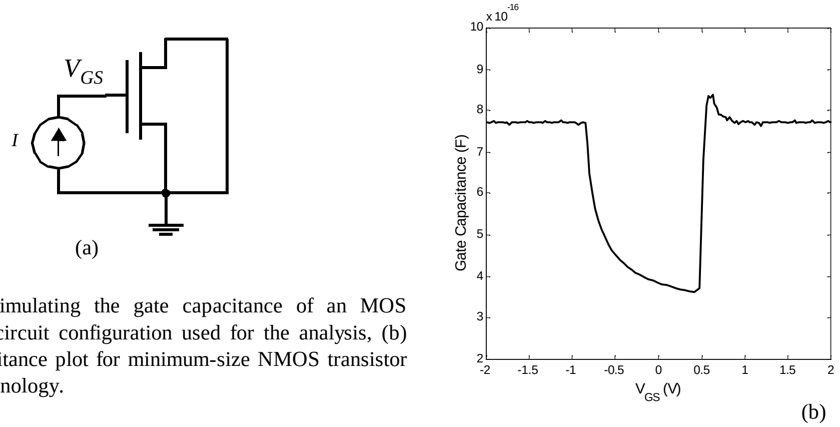


Figure 3.31 Simulating the gate capacitance of an MOS transistor; (a) circuit configuration used for the analysis, (b) resulting capacitance plot for minimum-size NMOS transistor in 0.25 μm technology.

region and its surroundings. The detailed picture, shown in Figure 3.32, shows that the junction consists of two components:

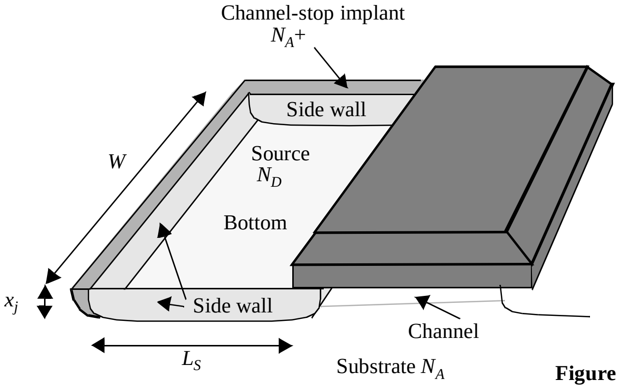


Figure 3.32 Detailed view of source junction.

- The *bottom-plate* junction, which is formed by the source region (with doping N_D) and the substrate with doping N_A. The total depletion region capacitance for this component equals $C_{bottom} = C_j W L_S$, with C_j the junction capacitance per unit area as given by Eq. (3.9). As the bottom-plate junction is typically of the abrupt type, the grading coefficient m approaches 0.5.
- The *side-wall* junction, formed by the source region with doping N_D and the p^+ channel-stop implant with doping level $N_A{}^+$. The doping level of the stopper is usually larger than that of the substrate, resulting in a larger capacitance per unit area. The side-wall junction is typically graded, and its grading coefficient varies from 0.33 to 0.5. Its capacitance value equals $C_{sw} = C'_{jsw} x_j\ (W + 2 \times L_s)$. Notice that no side-wall capacitance is counted for the fourth side of the source region, as this represents the conductive channel.[4]

  Since x_j, the junction depth, is a technology parameter, it is normally combined with C'_{jsw} into a capacitance per unit perimeter $C_{jsw} = C'_{jsw} x_j$. An expression for the total junction capacitance can then be derived,

$$\begin{aligned} C_{diff} &= C_{bottom} + C_{sw} = C_j \times AREA + C_{jsw} \times PERIMETER \\ &= C_j L_S W + C_{jsw}(2L_S + W) \end{aligned} \tag{3.46}$$

Since all these capacitances are small-signal capacitances, we normally linearize them and use average capacitances along the lines of Eq. (3.10).

Problem 3.1 Using a circuit simulator to determine the drain capacitance

Derive a simple circuit that would help you to derive the drain capacitance of an NMOS transistor in the different operation modes using circuit simulation (in the style of Figure 3.31).

Capacitive Device Model

All the above contributions can be combined in a single capacitive model for the MOS transistor, which is shown Figure 3.33. Its components are readily identified on the basis of the preceding discussions.

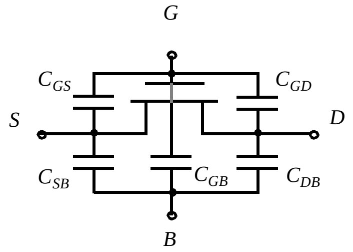


Figure 3.33 MOSFET capacitance model.

$$C_{GS} = C_{GCS} + C_{GSO};\ C_{GD} = C_{GCD} + C_{GDO};\ C_{GB} = C_{GCB}$$

$$C_{SB} = C_{Sdiff};\ C_{DB} = C_{Ddiff} \tag{3.47}$$

It is essential for the designers of high-performance and low-energy circuits to be very familiar with this model as well as to have an intuitive feeling of the relative values of its components.

Example 3.10 MOS Transistor Capacitances

Consider an NMOS transistor with the following parameters: t_{ox} = 6 nm, L = 0.24 µm, W = 0.36 µm, $L_D = L_S$ = 0.625 µm, $C_O = 3 \times 10^{-10}$ F/m, $C_{j0} = 2 \times 10^{-3}$ F/m^2, $C_{jsw0} = 2.75 \times 10^{-10}$ F/m. Determine the zero-bias value of all relevant capacitances.

The gate capacitance per unit area is easily derived as $(\varepsilon_{ox} / t_{ox})$ and equals 5.7 fF/µm^2. The gate-to-channel C_{GC} then equals WLC_{ox} = 0.49 fF. To find the total gate capacitance, we have to add the source and drain overlap capacitors, each of which equals WC_O = 0.105 fF. This leads to a total gate capacitance of 0.7 fF.

[4] To be entirely correct, we should take the diffusion capacitance of the source(drain)-to-channel junction into account. Due to the doping conditions and the small area, this component can virtually always be ignored in a first-order analysis. Detailed SPICE models most often include a factor C_{JSWG} to account for this junction.

The diffusion capacitance consists of the bottom and the side-wall capacitances. The former is equal to $C_{j0}\ L_D W$ = 0.45 fF, while the side-wall capacitance under zero-bias conditions evaluates to $C_{jsw0}\ (2L_D + W)$ = 0.44 fF. This results in a total drain(source)-to-bulk capacitance of 0.89 fF.

The diffusion capacitance seems to dominate the gate capacitance. This is a worst-case condition, however. When increasing the value of the reverse bias over the junction — as is the normal operation mode in MOS circuits —, the diffusion capacitance is substantially reduced. Also, clever design can help to reduce the value of L_D (L_S). In general, it can be stated that the contribution of diffusion capacitances is at most equal, and very often substantially smaller than the gate capacitance.

Design Data — MOS Transistor Capacitances

Table 3.5 summarizes the parameters needed to estimate the parasitic capacitances of the MOS transistors in our generic 0.25 μm CMOS process.

Table 3.5 Capacitance parameters of NMOS and PMOS transistors in 0.25 μm CMOS process.

	C_{ox} (fF/μm^2)	C_O (fF/μm)	C_j (fF/μm^2)	m_j	ϕ_b (V)	C_{jsw} (fF/μm)	m_{jsw}	ϕ_{bsw} (V)
NMOS	6	0.31	2	0.5	0.9	0.28	0.44	0.9
PMOS	6	0.27	1.9	0.48	0.9	0.22	0.32	0.9

Source-Drain Resistance

The performance of a CMOS circuit may further be affected by another set of parasitic elements, being the resistances in series with the drain and source regions, as shown in Figure 3.34a. This effect become more pronounced when transistors are scaled down, as this leads to shallower junctions and smaller contact openings become smaller. The resistance of the drain (source) region can be expressed as

$$R_{S,D} = \frac{L_{S,D}}{W} R_{\square} + R_C \tag{3.48}$$

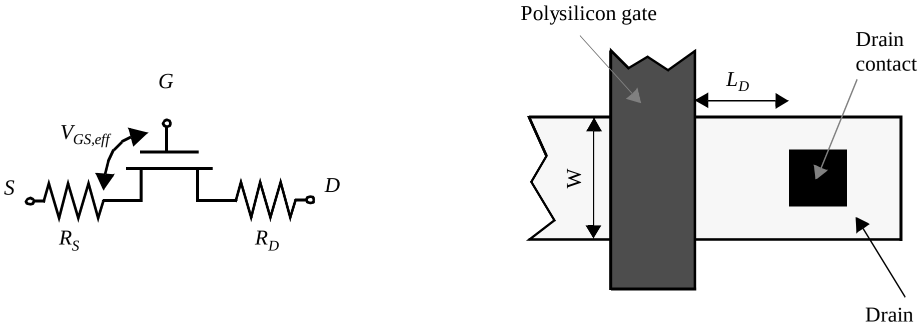
with R_C the contact resistance, W the width of the transistor, and $L_{S,D}$ the length of the source or drain region (Figure 3.34b). $R_{\square}$ is the *sheet resistance* per square of the drain-source diffusion, and ranges from 20 to 100 Ω/□. Observe that the resistance of a square of material is constant, independent of its size (see also Chapter 4).

The series resistance causes a deterioration in the device performance, as it reduces the drain current for a given control voltage. Keeping its value as small as possible is thus an important design goal for both the device and the circuit engineer. One option, popular in most contemporary processes, is to cover the drain and source regions with a low-resistivity material such as titanium or tungsten. This process is called *silicidation* and effectively reduces the sheet resistance to values in the range from 1 to 4 $\Omega/\square$.[5] Making the transistor wider than needed is another possibility as should be obvious from Eq. (3.48). With a process that includes silicidation and proper attention to layout, parasitic resistance is not important. However, the reader should be aware that careless layout may lead to resistances that severely degrade the device performance.

3.3.4 The Actual MOS Transistor—Some Secondary Effects

The operation of a contemporary transistor may show some important deviations from the model we have presented so far. These divergences become especially pronounced once the dimensions of the transistor reach the deep sub-micron realm. At that point, the assumption that the operation of a transistor is adequately described by a one-dimensional model, where it is assumed that all current flows on the surface of the silicon and the electrical fields are oriented along that plane, is not longer valid. Two- or even three-dimensional models are more appropriate. An example of such was already given in Section 3.2.2 when we discussed the mobility degradation.

The understanding of some of these second-order effects and their impact on the device behavior is essential in the design of today's digital circuits and therefore merits some discussion. One word of warning, though. Trying to take all those effects into account in a manual, first-order analysis results in intractable and opaque circuit models. It is therefore advisable to analyze and design MOS circuits first using the ideal model. The impact of the non-idealities can be studied in a second round using computer-aided simulation tools with more precise transistor models.

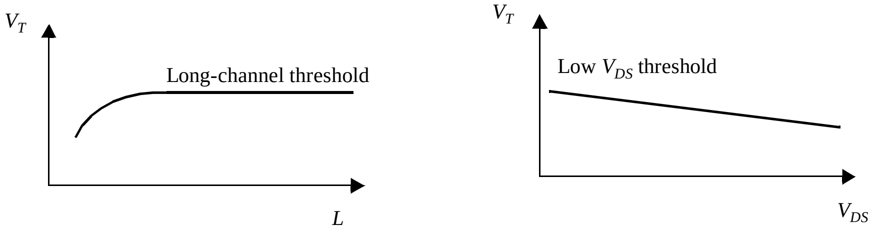


(a) Modeling the series resistance

(b) Parameters of the series resistance

Figure 3.34 Series drain and source resistance.

[5] Silicidation is also used to reduce the resistance of the polysilicon gate, as will be discussed in Chapter 4.



(a) Threshold as a function of the length (for low V_{DS})

(b) Drain-induced barrier lowering (for low L)

Figure 3.35 Threshold variations.

Threshold Variations

Eq. (3.19) states that the threshold voltage is only a function of the manufacturing technology and the applied body bias V_{SB}. The threshold can therefore be considered as a constant over all NMOS (PMOS) transistors in a design. As the device dimensions are reduced, this model becomes inaccurate, and the threshold potential becomes a function of L, W, and V_{DS}. Two-dimensional second-order effects that were ignorable for long-channel devices suddenly become significant.

In the traditional derivation of the V_{TO}, for instance, it is assumed that the channel depletion region is solely due to the applied gate voltage and that all depletion charge beneath the gate originates from the MOS field effects. This ignores the depletion regions of the source and reverse-biased drain junction, which become relatively more important with shrinking channel lengths. Since a part of the region below the gate is already depleted (by the source and drain fields), a smaller threshold voltage suffices to cause strong inversion. In other words, V_{T0} decreases with L for short-channel devices (Figure 3.35a). A similar effect can be obtained by raising the drain-source (bulk) voltage, as this increases the width of the drain-junction depletion region. Consequently, the threshold decreases with increasing $V_{DS.}$ This effect, called the *drain-induced barrier lowering*, or *DIBL*, causes the threshold potential to be a function of the operating voltages (Figure 3.35b). For high enough values of the drain voltage, the source and drain regions can even be shorted together, and normal transistor operation ceases to exist. The sharp increase in current that results from this effect, which is called *punch-through*, may cause permanent damage to the device and should be avoided. Punch-through hence sets an upper bound on the drain-source voltage of the transistor.

Since the majority of the transistors in a digital circuit are designed at the minimum channel length, the variation of the threshold voltage as a function of the length is almost uniform over the complete design, and is therefore not much of an issue except for the increased sub-threshold leakage currents. More troublesome is the DIBL, as this effect varies with the operating voltage. This is, for instance, a problem in dynamic memories, where the leakage current of a cell (being the subthreshold current of the access transistor) becomes a function of the voltage on the data-line, which depends upon the applied data patterns. From the cell perspective, DIBL manifests itself as a data-dependent noise source.

Worth mentioning is that the threshold of the MOS transistor is also subject to *narrow-channel* effects. The depletion region of the channel does not stop abruptly at the edges of the transistor, but extends somewhat under the isolating field-oxide. The gate voltage must support this extra depletion charge to establish a conducting channel. This effect is ignorable for wide transistors, but becomes significant for small values of W, where it results in an increase of the threshold voltage. For small geometry transistors, with small values of L and W, the effects of short- and narrow channels may tend to cancel each other out.

Hot-Carrier Effects

Besides varying over a design, threshold voltages in short-channel devices also have the tendency to *drift over time*. This is the result of the *hot-carrier* effect [Hu92]. Over the last decades, device dimensions have been scaled down continuously, while the power supply and the operating voltages were kept constant. The resulting increase in the electrical field strength causes an increasing velocity of the electrons, which can leave the silicon and tunnel into the gate oxide upon reaching a high-enough energy level. Electrons trapped in the oxide change the threshold voltage, typically increasing the thresholds of NMOS devices, while decreasing the V_T of PMOS transistors. For an electron to become hot, an electrical field of at least 10^4 V/cm is necessary. This condition is easily met in devices with channel lengths around or below 1 μm. The hot-electron phenomenon can lead to a long-term reliability problem, where a circuit might degrade or fail after being in use for a while. This is illustrated in Figure 3.36, which shows the degradation in the *I-V* characteristics of an NMOS transistor after it has been subjected to extensive operation. State-of-the-art MOSFET technologies therefore use specially-engineered drain and source regions to ensure that the peaks in the electrical fields are bounded, hence preventing carriers to reach the critical values necessary to become hot. The reduced supply voltage that is typical for deep sub-micron technologies can in part be attributed to the necessity to keep hot-carrier effects under control.

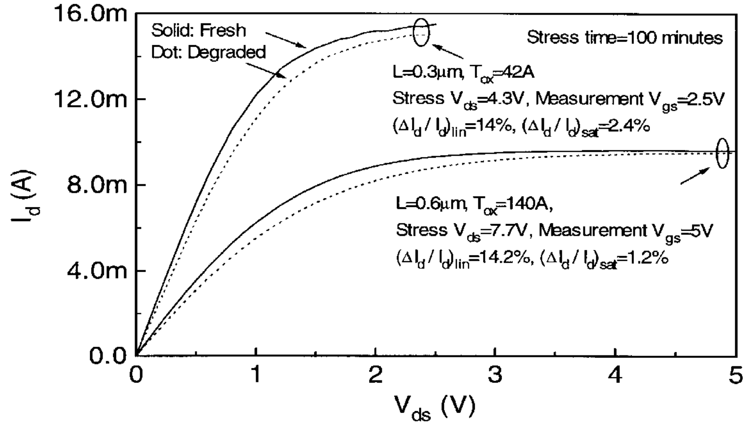
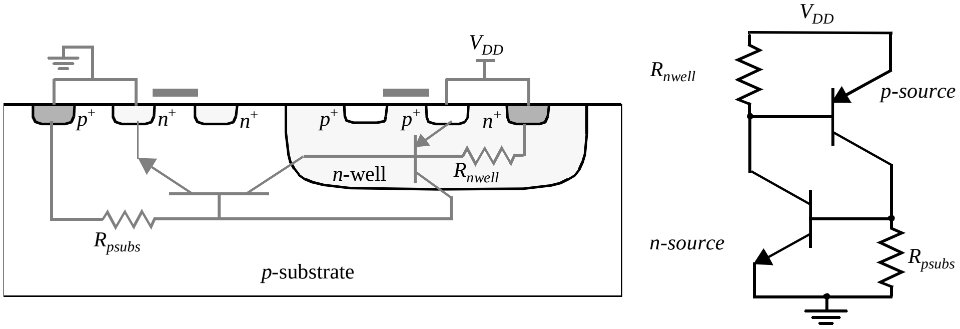


Figure 3.36 Hot-carrier effects cause the *I-V* characteristics of an NMOS transistor to degrade from extensive usage (from [McGaughy98]).

CMOS Latchup

The MOS technology contains a number of intrinsic bipolar transistors. These are especially troublesome in CMOS processes, where the combination of wells and substrates results in the formation of parasitic *n-p-n-p* structures. Triggering these thyristor-like devices leads to a shorting of the V_{DD} and V_{SS} lines, usually resulting in a destruction of the chip, or at best a system failure that can only be resolved by power-down.

Consider the *n*-well structure of Figure 3.37a. The *n-p-n-p* structure is formed by the source of the NMOS, the *p*-substrate, the *n*-well and the source of the PMOS. A circuit equivalent is shown in Figure 3.37b. When one of the two bipolar transistors gets forward biased (e.g., due to current flowing through the well, or substrate), it feeds the base of the other transistor. This positive feedback increases the current until the circuit fails or burns out.



(a) Origin of latchup

(b) Equivalent circuit

Figure 3.37 CMOS latchup.

From the above analysis the message to the designer is clear—to avoid latchup, the resistances R_{nwell} and R_{psubs} should be minimized. This can be achieved by providing numerous well and substrate contacts, placed close to the source connections of the NMOS/PMOS devices. Devices carrying a lot of current (such as transistors in the I/O drivers) should be surrounded by *guard rings*. These circular well/substrate contacts, positioned around the transistor, reduce the resistance even further and reduce the gain of the parasitic bipolars. For an extensive discussion on how to avoid latchup, please refer to [Weste93]. The latchup effect was especially critical in early CMOS processes. In recent years, process innovations and improved design techniques have all but eliminated the risks for latchup.

3.3.5 SPICE Models for the MOS Transistor

The complexity of the behavior of the short-channel MOS transistor and its many parasitic effects has led to the development of a wealth of models for varying degrees of accuracy and computing efficiency. In general, more accuracy also means more complexity and, hence, an increased run time. In this section, we briefly discuss the characteristics of the more popular MOSFET models, and describe how to instantiate a MOS transistor in a circuit description.

SPICE Models

SPICE has three built-in MOSFET models, selected by the LEVEL parameter in the model card. Unfortunately, all these models have been rendered obsolete by the progression to short-channel devices. They should only be used for first-order analysis, and we therefore limit ourselves to a short discussion of their main properties.

- The LEVEL 1 SPICE model implements the *Shichman-Hodges model*, which is based on the square law long-channel expressions, derived earlier in this chapter. It does not handle short-channel effects.

- The LEVEL 2 model is a geometry-based model, which uses detailed device physics to define its equations. It handles effects such as velocity saturation, mobility degradation, and drain-induced barrier lowering. Unfortunately, including all 3D-effects of an advanced submicron process in a pure physics-based model becomes complex and inaccurate.

- LEVEL 3 is a semi-empirical model. It relies on a mixture of analytical and empirical expressions, and uses measured device data to determine its main parameters. It works quite well for channel lengths down to 1 µm.

In response to the inadequacy of the built-in models, SPICE vendors and semi-conductor manufacturers have introduced a wide range of accurate, but proprietary models. A complete description of all those would take the remainder of this book, which is, obviously, not the goal. We refer the interested reader to the extensive literature on this topic [e.g. VladimirescuXX].

The BSIM3V3 SPICE Model

The confusing situation of having to use a different model for each manufacturer has fortunately been partially resolved by the adoption of the BSIM3v3 model as an industry-wide standard for the modeling of deep-submicron MOSFET transistors. The **B**erkeley **S**hort-Channel **I**GFET **M**odel (or BSIM in short) provides a model that is analytically simple and is based on a 'small' number of parameters, which are normally extracted from experimental data. Its popularity and accuracy make it the natural choice for all the simulations presented in this book.

A full-fledged BSIM3v3 model (denoted as LEVEL 49) contains over 200 parameters, the majority of which are related to the modeling of second-order effects. Fortunately, understanding the intricacies of all these parameters is not a requirement for the digital designer. We therefore only present an overview of the parameter categories (Table 3.6). The *Bin* category deserves some extra attention. Providing a single set of parameters that is acceptable over all possible device dimensions is deemed to be next to impossible. So, a set of models is provided, each of which is valid for a limited region delineated by

LMIN, LMAX, WMIN, and WMAX (called a bin). It is typically left to the user to select the correct bin for a particular transistor.

Table 3.6 BSIM3-V3 model parameter categories, and some important parameters.

Parameter Category	Description
Control	Selection of level and models for mobility, capacitance, and noise LEVEL, MOBMOD, CAPMOD
DC	Parameters for threshold and current calculations VTH0, K1, U0, VSAT, RSH,
AC & Capacitance	Parameters for capacitance computations CGS(D)O, CJ, MJ, CJSW, MJSW
dW and dL	Derivation of effective channel length and width
Process	Process parameters such as oxide thickness and doping concentrations TOX, XJ, GAMMA1, NCH, NSUB
Temperature	Nominal temperature and temperature coefficients for various device parameters TNOM
Bin	Bounds on device dimensions for which model is valid LMIN, LMAX, WMIN, WMAX
Flicker Noise	Noise model parameters

We refer the interested reader to the BSIM3v3 documentation provided on the website of the textbook (REFERENCE) for a complete description of the model parameters and equations. The LEVEL-49 models for our generic 0.25 μm CMOS process can be found at the same location.

Transistor Instantiation

The parameters that can be specified for an individual transistor are enumerated in Table 3.7. Not all these parameters have to be defined for each transistor. SPICE assumes default values (which are often zero!) for the missing factors.

WARNING: It is hard to expect accuracy from a simulator, when the circuit description provided by the designer does not contain the necessary details. For instance, you must accurately specify the area and the perimeter of the source and drain regions of the devices when performing a performance analysis. Lacking this information, which is used for the computation of the parasitic capacitances, your transient simulation will be next to useless. Similarly, it is often necessary to painstakingly define the value of the drain and source resistance. The NRS and NRD values multiply the sheet resistance specified in the transistor model for an accurate representation of the parasitic series source and drain resistance of each transistor.

Table 3.7 SPICE transistor parameters.

Parameter Name	Symbol	SPICE Name	Units	Default Value
Drawn Length	*L*	L	m	–
Effective Width	*W*	W	m	–
Source Area	*AREA*	AS	m^2	0
Drain Area	*AREA*	AD	m^2	0
Source Perimeter	*PERIM*	PS	m	0
Drain Perimeter	*PERIM*	PD	m	0
Squares of Source Diffusion		NRS	–	1
Squares of Drain Diffusion		NRD	–	1

Example 3.11 SPICE description of a CMOS inverter

An example of a SPICE description of a CMOS inverter, consisting of an NMOS and a PMOS transistor, is given below. Transistor M1 is an NMOS device of model-type (and bin) *nmos*.1 with its drain, gate, source, and body terminals connected to nodes *nvout*, *nvin*, 0, and 0, respectively. Its gate length is the minimum allowed in this technology (0.25 μm). The '+' character at the start of line 2 indicates that this line is a continuation of the previous one.

The PMOS device of type *pmos*.1, connected between nodes *nvout*, *nvin*, *nvdd*, and *nvdd* (D, G, S, and B, respectively), is three times wider, which reduces the series resistance, but increases the parasitic diffusion capacitances as the area and perimeter of the drain and source regions go up.

Finally, the *.lib* line refers to the file that contains the transistor models.

```
M1 nvout nvin 0 0 nmos.1 W=0.375U L=0.25U
+AD=0.24P PD=1.625U AS=0.24P PS=1.625U NRS=1 NRD=1
M2 nvout nvin nvdd nvdd pmos.1 W=1.125U L=0.25U
+AD=0.7P PD=2.375U AS=0.7P PS=2.375U NRS=0.33 NRD=0.33
.lib 'c:\Design\Models\cmos025.l'
```

3.4 A Word on Process Variations

The preceding discussions have assumed that a device is adequately modeled by a single set of parameters. In reality, the parameters of a transistor vary from wafer to wafer, or even between transistors on the same die, depending upon the position. This observed random distribution between supposedly identical devices is primarily the result of two factors:

1. Variations in the process parameters, such as impurity concentration densities, oxide thicknesses, and diffusion depths, caused by nonuniform conditions during the deposition and/or the diffusion of the impurities. These result in diverging values for sheet resistances, and transistor parameters such as the threshold voltage.

2. Variations in the dimensions of the devices, mainly resulting from the limited resolution of the photolithographic process. This causes deviations in the (W/L) ratios of MOS transistors and the widths of interconnect wires.

Observe that quite a number of these deviations are totally uncorrelated. For instance, variations in the length of an MOS transistor are unrelated to variations in the threshold voltage as both are set by different process steps. Below we examine the impact on some of the parameters that determine the transistor current.

- The *threshold voltage* V_T can vary for numerous reasons: changes in oxide thickness, substrate, poly and implant impurity levels, and the surface charge. Accurate control of the threshold voltage is an important goal for many reasons. Where in the past thresholds could vary by as much as 50%, state-of-the-art digital processes manage to control the thresholds to within 25-50 mV.
- k'_n: The main cause for variations in the process transconductance is changes in oxide thickness. Variations can also occur in the mobility but to a lesser degree.
- Variations in W and L. These are mainly caused by the lithographic process. Observe that variations in W and L are totally uncorrelated since the first is determined in the field-oxide step, while the second is defined by the polysilicon definition and the source and drain diffusion processes.

The measurable impact of the process variations may be a substantial deviation of the circuit behavior from the nominal or expected response, and this could be in either positive or negative directions. This poses the designer for an important economic dilemma. Assume, for instance, that you are supposed to design a microprocessor running at a clock frequency of 500 MHz. It is economically important that the majority of the manufactured dies meet that performance requirement. One way to achieve that goal is to design the circuit assuming worst-case values for all possible device parameters. While safe, this approach is prohibitively conservative and results in severely overdesigned and hence uneconomical circuits.

To help the designer make a decision on how much margin to provide, the device manufacturer commonly provides fast and slow device models in addition to the nominal ones. These result in larger or smaller currents than expected, respectively.

Example 3.12 MOS Transistor Process Variations

To illustrate the possible impact of process variations on the performance of an MOS device, consider a minimum-size NMOS device in our generic 0.25 μm CMOS process. A later chapter will establish that the speed of the device is proportional to the drain current that can be delivered.

Assume initially that $V_{GS} = V_{DS} = 2.5$ V. From earlier simulations, we know that this produces a drain current of 220 μA. The nominal model is now replaced by the fast and slow models, that modify the length and width (±10%), threshold (±60 mV), and oxide thickness (±5%) parameters of the device. Simulations produce the following data:

Fast: $I_d = 265$ μA: +20%
Slow: $I_d = 182$ μA: -17%

Let us now proceed one step further. The supply voltage delivered to a circuit is by no means a constant either. For instance, the voltage delivered by a battery can drop off substantially towards the end of its lifetime. In practice, a variation in 10% of the supply voltage may well be expected.

Fast + V_{dd} = 2.75 V: I_d = 302 µA: +37%
Slow + V_{dd} = 2.25 V: I_d = 155 µA: -30%

The current levels and the associated circuit performance can thus vary by almost 100% between the extreme cases. To guarantee that the fabricated circuits meet the performance requirements under all circumstances, it is necessary to make the transistor 42% (=220µA/155µA) wider then would be required in the nominal case. This translates into a severe area penalty.

Fortunately, these worst- (or best-) case conditions occur only very rarely in reality. The probability that all parameters assume their worst-case values simultaneously is very low, and most designs will display a performance centered around the nominal design. The art of the *design for manufacturability* is to center the nominal design so that the majority of the fabricated circuits (e.g., 98%) will meet the performance specifications, while keeping the area overhead minimal.

Specialized design tools to help meet this goal are available. For instance, the Monte Carlo analysis approach [Jensen91] simulates a circuit over a wide range of randomly chosen values for the device parameters. The result is a distribution plot of design parameters (such as the speed or the sensitivity to noise) that can help to determine if the nominal design is economically viable. Examples of such distribution plots, showing the impact of variations in the effective transistor channel length and the PMOS transistor thresholds on the speed of an adder cell, are shown in Figure 3.38. As can be observed, technology variations can have a substantial impact on the performance parameters of a design.

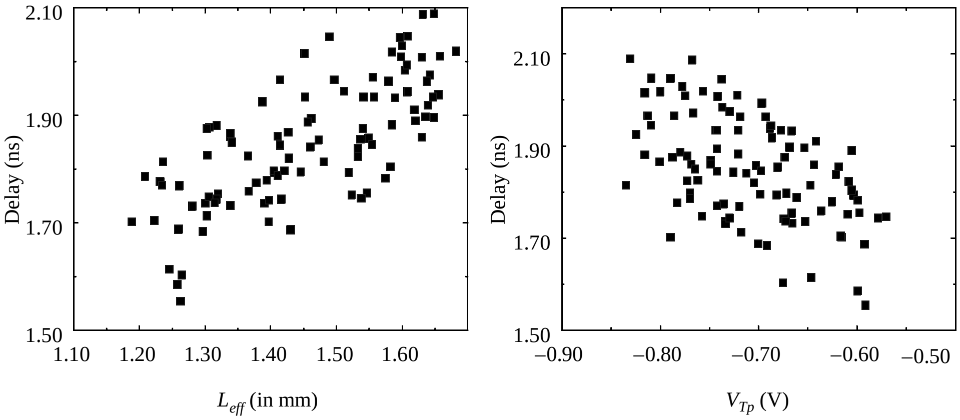


Figure 3.38 Distribution plots of speed of adder circuit as a function of varying device parameters, as obtained by a Monte Carlo analysis. The circuit is implemented in a 2 µm (nominal) CMOS technology (*courtesy of Eric Boskin, UCB, and ATMEL corp.*).

One important conclusion from the above discussion is that SPICE simulations should be treated with care. The device parameters presented in a model represent average values, measured over a batch of manufactured wafers. Actual implementations are bound to differ from the simulation results, and for reasons other than imperfections in the modeling approach. Be furthermore aware that temperature variations on the die can present another source for parameter deviations. Optimizing an MOS circuit with SPICE to a resolution level of a picosecond or a microvolt is clearly a waste of effort.

3.5 Perspective: Technology Scaling

Over the last decades, we have observed a spectacular increase in integration density and computational complexity of digital integrated circuits. As already argued in the introduction, applications that were considered implausible yesterday are already forgotten today. Underlying this revolution are the advances in device manufacturing technology that allow for a steady reduction of the minimum feature size such as the minimum transistor channel length realizable on a chip. To illustrate this point, we have plotted in Figure 3.39 the evolution of the (average) minimum device dimensions starting from the 1960s and projecting into the 21st century. We observe a reduction rate of approximately 13% per year, halving every 5 years. Another interesting observation is that no real sign of a slowdown is in sight, and that the breathtaking pace will continue in the foreseeable future.

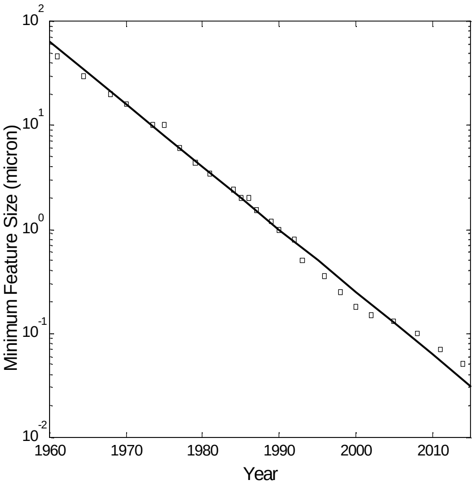


Figure 3.39 Evolution of (average) minimum channel length of MOS transistors over time. Dots represent observed or projected (2000 and beyond) values. The continuous line represents a scaling scenario that reduces the minimum feature with a factor 2 every 5 years.

A pertinent question is how this continued reduction in feature size influences the operating characteristics and properties of the MOS transistor, and indirectly the critical digital design metrics such as switching frequency and power dissipation. A first-order projection of this behavior is called a *scaling analysis,* and is the topic of this section. In addition to the minimum device dimension, we have to consider the supply voltage as a second independent variable in such a study. Different scaling scenarios result based on how these two independent variables are varied with respect to each other [Dennard74, Baccarani84].

Three different models are studied in Table 3.8. To make the results tractable, it is assumed that all device dimensions scale by the same factor *S* (with $S > 1$ for a reduction in size). This includes the width and length of the transistor, the oxide thickness, and the junction depths. Similarly, we assume that all voltages, including the supply voltage and the threshold voltages, scale by a same ratio U. The relations governing the scaling behavior of the dependent variables are tabulated in column 2. Observe that this analysis only considers short-channel devices with a linear dependence between control voltage and saturation current (as expressed by Eq. (3.39)). We discuss each scenario in turn.

Table 3.8 Scaling scenarios for short-channel devices.

Parameter	Relation	Full Scaling	General Scaling	Fixed-Voltage Scaling
W, L, t_{ox}		$1/S$	$1/S$	$1/S$
V_{DD}, V_T		$1/S$	$1/U$	1
N_{SUB}	$V/W_{depl}{}^2$	S	S^2/U	S^2
Area/Device	WL	$1/S^2$	$1/S^2$	$1/S^2$
C_{ox}	$1/t_{ox}$	S	S	S
C_{gate}	$C_{ox}WL$	$1/S$	$1/S$	$1/S$
k_n, k_p	$C_{ox}W/L$	S	S	S
I_{sat}	$C_{ox}WV$	$1/S$	$1/U$	1
Current Density	$I_{sat}/Area$	S	S^2/U	S^2
Ron	V/I_{sat}	1	1	1
Intrinsic Delay	$R_{on}C_{gate}$	$1/S$	$1/S$	$1/S$
P	$I_{sat}V$	$1/S^2$	$1/U^2$	1
Power Density	$P/Area$	1	S^2/U^2	S^2

Full Scaling (Constant Electrical Field Scaling)

In this ideal model, voltages and dimensions are scaled by the same factor S. The goal is to keep the electrical field patterns in the scaled device identical to those in the original device. Keeping the electrical fields constant ensures the physical integrity of the device and avoids breakdown or other secondary effects. This scaling leads to greater device density (*Area*), higher performance (*Intrinsic Delay*), and reduced power consumption (P). The effects of full scaling on the device and circuit parameters are summarized in the third column of Table 3.8. We use the intrinsic time constant, which is the product of the gate capacitance and the on-resistance, as a measure for the performance. The analysis shows that the on-resistance remains constant due to the simultaneous scaling of voltage swing and current level. The performance improved is solely due to the reduced capacitance. The results clearly demonstrate the beneficial effects of scaling—the speed of the circuit increases in a linear fashion, while the power/gate scales down quadratically![6]

[6] Some assumptions were made when deriving this table:

1. It is assumed that the carrier mobilities are not affected by the scaling.
2. The substrate doping N_{sub} is scaled so that the maximum depletion-layer width is reduced by a factor S.
3. It is furthermore assumed that the delay of the device is mainly determined by the intrinsic capacitance (the gate capacitance) and that other device capacitances, such as the diffusion capacitances, scale appropriately. This assumption is approximately true for the full-scaling case, but not for fixed-voltage scaling, where C_{diff} scales as $1/\sqrt{S}$.

Fixed-Voltage Scaling

In reality, full scaling is not a feasible option. First of all, to keep new devices compatible with existing components, voltages cannot be scaled arbitrarily. Having to provide for multiple supply voltages adds considerably to the cost of a system. As a result, voltages have not been scaled down along with feature sizes, and designers adhere to well-defined standards for supply voltages and signal levels. As is illustrated in Figure 3.40, 5 V was the de facto standard for all digital components up to the early 1990s, and a *fixed-voltage scaling model* was followed.

Only with the introduction of the 0.5 μm CMOS technology did new standards such as 3.3 V and 2.5 V make an inroad. Today, a closer tracking between voltage and device dimension can be observed.The reason for this change in operation model can partially be explained with the aid of the fixed-voltage scaling model, summarized in the fifth column of Table 3.8. In a velocity-saturated device, keeping the voltage constant while scaling the device dimensions does not give a performance advantage over the full-scaling model, but instead comes with a major power penalty. The gain of an increased current is simply offset by the higher voltage level, and only hurts the power dissipation. This scenario is very different from the situation that existed when transistors were operating in the long-channel mode, and the current was a quadratic function of the voltage (as per Eq. (3.29)). Keeping the voltage constant under these circumstances gives a distinct performance advantage, as it causes a net reduction in on-resistance.

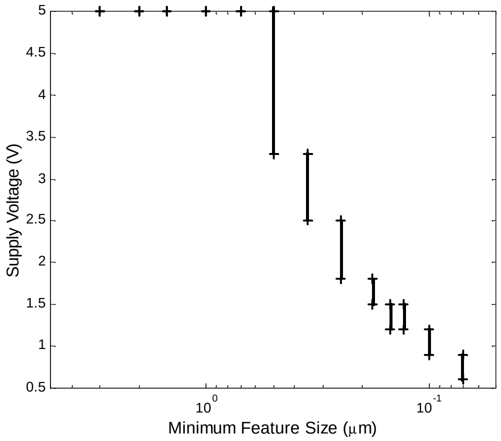


Figure 3.40 Evolution of min and max supply-voltage in digital integrated circuits as a function of feature size. All values for 0.15 micron and below are projected.

While the above argumentation offers ample reason to scale the supply voltages with the technology, other physical phenomena such as the hot-carrier effect and oxide breakdown also contributed to making the fixed-voltage scaling model unsustainable.

Problem 3.2 Scaling of Long-channel Devices

Demonstrate that for a long-channel transistor, full-voltage scaling results in a reduction of the intrinsic delay with a factor S^2, while increasing the power dissipation/device by S.

Reconstruct Table 3.8 assuming that the current is a quadratic function of the voltage (Eq. (3.29)).

WARNING: The picture painted in the previous section represents a first-order model. Increasing the supply voltage still offers somewhat of a performance benefit for short-channel transistors. This is apparent in Figure 3.27 and Table 3.3, which show a reduction of the equivalent on-resistance with increasing supply voltage — even for the high voltage range. Yet, this effect, which is mostly due to the channel-length modulation, is secondary and is far smaller than what would be obtained in case of long-channel devices.

The reader should keep this warning in the back of his mind throughout this scaling study. The goal is to discover first-order trends. This implies ignoring second-order effects such as mobility-degradation, series resistance, etc.

General Scaling

We observe in Figure 3.40 that the supply voltages, while moving downwards, are not scaling as fast as the technology. For instance, for the technology scaling from 0.5 μm to 0.1 μm, the maximum supply-voltage only reduces from 5 V to 1.5 V. The obvious question is why not to stick to the full-scaling model, when keeping the voltage higher does not yield any convincing benefits? This departure is motivated by the following argumentation:

- Some of the intrinsic device voltages such as the silicon bandgap and the built-in junction potential, are material parameters and cannot be scaled.
- The scaling potential of the transistor threshold voltage is limited. Making the threshold too low makes it difficult to turn off the device completely. This is aggravated by the large process variation of the value of the threshold, even on the same wafer.

Therefore, a more general scaling model is needed, where dimensions and voltages are scaled independently. This general scaling model is shown in the fourth column of Table 3.8. Here, device dimensions are scaled by a factor S, while voltages are reduced by a factor U. When the voltage is held constant, $U = 1$, and the scaling model reduces to the fixed-voltage model. Note that the general-scaling model offers a performance scenario identical to the full- and the fixed scaling, while its power dissipation lies between the two models (for $S > U > 1$).

Verifying the Model

To summarize this discussion on scaling, we have combined in Table 3.9 the characteristics of some of the most recent CMOS processes and projections on some future ones. Observe how the operating voltages are being continuously reduced with diminishing device dimensions, while threshold voltages remain virtually constant. As predicted by the scaling model, the maximum drive current remains approximately constant.

Table 3.9 MOSFET technology projection (from [SIA97]).

Year of Introduction	1997	1999	2001	2003	2006	2009
Channel length (µm)	0.25	0.18	0.15	0.13	0.1	0.07
Gate oxide (nm)	4-5	3-4	2-3	2-3	1.5-2	< 1.5
V_{DD} (V)	1.8-2.5	1.5-1.8	1.2-1.5	1.2-1.5	0.9-1.2	0.6-0.9
V_T (V)	0.4	0.4	0.4	0.4	0.4	0.4
NMOS/PMOS I_{Dsat} (nA/µm)	600/280	600/280	600/280	600/280	600/280	600/280

From the above, it is reasonable to conclude that both integration density and performance will continue to increase. The obvious question is for how long? Experimental 25 nm CMOS devices have proven to be operational in the laboratories and to display current characteristics that are surprisingly close to present-day transistors. These transistors, while working along similar concepts as the current MOS devices, look very different from the structures we are familiar with, and require some substantial *device engineering*. For instance, Figure 3.41 shows a potential transistor structure, the folded channel dual-gated transistor, which has proven to be operational up to very small channel lengths.

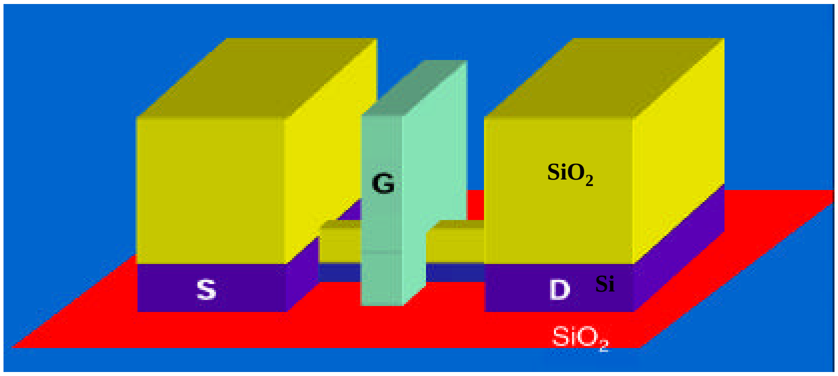


Figure 3.41 Folded dual-gated transistor with 25 nm channel length [Hu99].

Integrated circuits integrating more then one billion transistors clocked at speeds of multiple GHz's hence seem to be well under way. Whether this will actually happen is an open question. Even though it might be technologically feasible, other parameters have an equal impact on the feasibility of such an undertaking. A first doubt is if such a part can be manufactured in an economical way. Current semiconductor plants cost over $5 billion, and this price is expected to rise substantially with smaller feature sizes. Design considerations also play a role. Power consumption of such a component might be prohibitive. The growing role of interconnect parasitics might put an upper bound on performance. Finally, system considerations might determine what level of integration is really desirable. All in all, it is obvious that the design of semiconductor circuits still faces an exciting future.

3.6 Summary

In this chapter, we have presented a a comprehensive overview of the operation of the MOSFET transistor, the semiconductor device at the core of virtually all contemporary digital integrated circuits. Besides an intuitive understanding of its behavior, we have presented a variety of modeling approaches ranging from simple models, useful for a first-order manual analysis of the circuit operation, to complex SPICE models. These models will be used extensively in later chapters, where we look at the fundamental building blocks of digital circuits. We started off with a short discussion of the semiconductor diode, one of the most dominant parasitic circuit elements in CMOS designs.

- The static behavior of the junction diode is well described by the ideal diode equation that states that the current is an *exponential function of the applied voltage bias*.
- In reverse-biased mode, the depletion-region space charge of the diode can be modeled as a non-linear voltage-dependent capacitance. This is particularly important as the omnipresent source-body and drain-body junctions of the MOS transistors all operate in this mode. A linearized large-scale model for the depletion capacitance was introduced for manual analysis.
- The MOS(FET) transistor is a voltage-controlled device, where the controlling gate terminal is insulated from the conducting channel by a SiO_2 capacitor. Based on the value of the gate-source voltage with respect to a threshold voltage V_T, three operation regions have been identified: *cut-off, linear*, and *saturation*. One of the most enticing properties of the MOS transistor, which makes it particularly amenable to digital design, is that it approximates a voltage-controlled switch: when the control voltage is low, the switch is nonconducting (open); for a high control voltage, a conducting channel is formed, and the switch can be considered closed. This two-state operation matches the concepts of binary digital logic.
- The continuing reduction of the device dimensions to the submicron range has introduced some substantial deviations from the traditional long-channel MOS transistor model. The most important one is the *velocity saturation* effect, which changes the dependence of the transistor current with respect to the controlling voltage from *quadratic to linear*. Models for this effect as well as other second-order parasitics have been introduced. One particular effect that is gaining in importance is the *subthreshold conduction*, which causes devices to conduct current even when the control voltage drops below the threshold.
- The dynamic operation of the MOS transistor is dominated by the *device capacitors*. The main contributors are the gate capacitance and the capacitance formed by the depletion regions of the source and drain junctions. The minimization of these capacitances is the prime requirement in high-performance MOS design.
- SPICE models and their parameters have been introduced for all devices. It was observed that these models represent an average behavior and can vary over a single wafer or die.

- The MOS transistor is expected to dominate the digital integrated circuit scene for the next decade. Continued scaling will lead to device sizes of approximately 0.07 micron by the year 2010, and logic circuits integrating more than 1 billion transistors on a die.

3.7 To Probe Further

Semiconductor devices have been discussed in numerous books, reprint volumes, tutorials, and journal articles. The *IEEE Journal on Electron Devices* is one of the prime journals, where most of the state-of-the-art devices and their modeling are discussed. The books and journals referenced below contain excellent discussions of the semiconductor devices of interest or refer to specific topics brought up in the course of this chapter.

REFERENCES

[Antognetti88] P. Antognetti and G. Masobrio (eds.), *Semiconductor Device Modeling with SPICE*, McGraw-Hill, 1988.

[Banzhaf92] W. Bhanzhaf, *Computer Aided Analysis Using PSPICE*, 2nd ed., Prentice Hall, 1992.

[Chen90] J. Chen, *CMOS Devices and Technology for VLSI*, Prentice Hall, 1990.

[Gray69] P. Gray and C. Searle, *Electronic Principles*, John Wiley and Sons, 1969.

[Gray93] P. Gray and R. Meyer, *Analysis and Design of Analog Integrated Circuits*, 3rd ed., John Wiley and Sons, 1993.

[Haznedar91] H. Haznedar, *Digital Microelectronics*, Benjamin/Cummings, 1991.

[Hodges88] D. Hodges and H. Jackson, *Analysis and Design of Digital Integrated Circuits*, 2nd ed., McGraw-Hill, 1988.

[Howe95] R. Howe and S. Sodini, *Microelectronics: An Integrated Approach*, Prentice Hall, 1995.

[Hu92] C. Hu, “IC Reliability Simulation,” *IEEE Journal of Solid State Circuits*, vol. 27, no. 3, pp. 241–246, March 1992.

[Hu93] C. Hu, “Future CMOS Scaling and Reliability,” *IEEE Proceedings*, vol. 81, no. 5, May 1993.

[Jensen91] G. Jensen et al., “Monte Carlo Simulation of Semiconductor Devices,” *Computer Physics Communications*, 67, pp. 1–61, August 1991.

[Ko89] P. Ko, “Approaches to Scaling,” in *VLSI Electronics: Microstructure Science*, vol. 18, chapter 1, pp. 1–37, Academic Press, 1989.

[Muller86] R. Muller and T. Kamins, *Device Electronics for Integrated Circuits*, 2nd ed., John Wiley and Sons, 1986.

[Nagel75] L. Nagel, “SPICE2: a Computer Program to Simulate Semiconductor Circuits,” Memo ERL-M520, Dept. Elect. and Computer Science, University of California at Berkeley, 1975.

[Sedra87] A. Sedra and K. Smith, *Microelectronic Circuits*, 2nd ed., Holt, Rinehart and Winston, 1987.

[Sheu87] B. Sheu, D. Scharfetter, P. Ko, and M. Jeng, “BSIM: Berkeley Short-Channel IGFET Model for MOS Transistors,” *IEEE Journal of Solid-State Circuits,* vol. SC-22, no. 4, pp. 558–565, August 1987.

[Sze81] S. Sze, *Physics of Semiconductor Devices*, 2nd ed., John Wiley and Sons, 1981.

[Thorpe92] T. Thorpe, *Computerized Circuit Analysis with SPICE*, John Wiley and Sons, 1992.

[Toh88] K. Toh, P. Koh, and R. Meyer, "An Engineering Model for Short-Channel MOS Devices," *IEEE Journal of Solid-Sate Circuits*, vol. 23. no. 4, pp 950–957, August 1988.

[Tsividis87] Y. Tsividis, *Operation and Modeling of the MOS Transistor*, McGraw-Hill, 1987.

[Yamaguchi88] T. Yamaguchi et al., "Process and Device Performance of a High-Speed Double Poly-Si Bipolar Technology Using Borsenic-Poly Process with Coupling-Base Implant," *IEEE Trans. Electron. Devices*, vol. 35, no 8, pp. 1247–1255, August 1988.

[Weste93] N. Weste and K. Eshragian, *Principles of CMOS VLSI Design: A Systems Perspective*, Addison-Wesley, 1993.

3.8 Exercises and Design Problems

For all problems, use the device parameters provided in Chapter 3 (XXX) and the inside back book cover, unless otherwise mentioned. Also assume T = 300 K by default.

1. [E,SPICE,2.23]

a. Consider the circuit of Figure 3.42. Using the simple model, with V_{Don} = 0.7 V, solve for I_D.

b. Find I_D and V_D using the ideal diode equation. Use $I_s = 10^{-14}$ A and T = 300 K.

c. Solve for V_{D1}, V_{D2}, and I_D using SPICE.

d. Repeat parts *b* and *c* using $I_S = 10^{-16}$ A, T = 300K, and $I_S = 10^{-14}$A, T = 350 K.

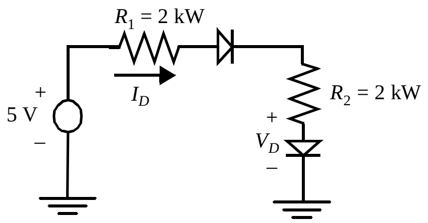


Figure 3.42 Resistor diode circuit.

2. [M, None, 2.2.3] For the circuit in Figure 3.43, V_s = 3.3 V. Assume A_D = 12 μm^2, ϕ_0 = 0.65 V, and m = 0.5. N_A = 2.5 E16 and N_D = 5 E15.

a. Find I_D and V_D.

b. Is the diode forward- or reverse-biased?

c. Find the depletion region width, W_j, of the diode.

d. Use the parallel-plate model to find the junction capacitance, C_j.

e. Set V_s = 1.5 V. Again using the parallel-plate model, explain qualitatively why C_j increases.

3. [E, SPICE, 2.3.2] Figure 3.44 shows NMOS and PMOS devices with drains, source, and gate ports annotated. Determine the mode of operation (saturation, triode, or cutoff) and drain current I_D for each of the biasing configurations given below. Verify with SPICE. Use the following transistor data: NMOS: k_n' = 60 μA/V^2, V_{T0} = 0.7 V, λ = 0.1 V^{-1}, PMOS: k_p' = 20 μA/V^2, V_{T0} = –0.8 V, λ = 0.1 V^{-1}. Assume (W/L) = 1.

a. NMOS: V_{GS} = 3.3 V, V_{DS} = 3.3 V. PMOS: V_{GS} = –0.5 V, V_{DS} = –1.5 V.