Chapter 2

Design and Fabrication of VLSI Devices

VLSI chips are manufactured in a fabrication facility usually referred to as a “fab”. A fab is a collection of manufacturing facilities and “clean rooms”, where wafers are processed through a variety of cutting, sizing, polishing, deposition, etching and cleaning operations. Clean room is a term used to describe a closed environment where air quality must be strictly regulated. The number and size of dust particles allowed per unit volume is specified by the classification standard of the clean room. Usually space-suit like overalls and other dress gear is required for humans, so they do not contaminate the clean room. The cleanliness of air in a fab is a critical factor, since dust particles cause major damage to chips, and thereby affect the overall yield of the fabrication process. The key factor which describes the fab in terms of technology is the minimum feature size it is capable of manufacturing. For example, a fab which runs a 0.25 micron fabrication process is simply referred to as a 0.25 micron fab.

A chip consists of several layers of different materials on a silicon wafer. The shape, size and location of material in each layer must be accurately specified for proper fabrication. A *mask* is a specification of geometric shapes that need to be created on a certain layer. Several masks must be created, one for each layer. The actual fabrication process starts with the creation of a silicon wafer by crystal growth. The wafer is then processed for size and shape with proper tolerance. The wafer’s size is typically large enough to fabricate several dozen identical/different chips. Masks are used to create specific patterns of each material in a sequential manner, and create a complex pattern of several layers. The order in which each layer is defined, or ‘patterned’ is very important. Devices are formed by overlapping a material of certain shape in one layer by another material and shape in another layer. After patterning all the layers, the wafer is cut into individual chips and packaged. Thus, the VLSI physical design is a process of creating all the necessary masks that define the sizes and location of the various devices and the interconnections between them.

The complex process of creating the masks requires a good understanding of the functionality of the devices to be formed, and the rigid rules imposed by the fabrication process. The manufacturing tolerances in the VLSI fabrication process are so tight that misalignment of a shape in a layer by a few microns can render the entire chip useless. Therefore, shapes and sizes of all the materials on all the layers of a wafer must conform to strict design rules to ensure proper fabrication. These rules play a key role in defining the physical design problems, and they depend rather heavily on the materials, equipment used and maturity of the fabrication process. The understanding of limitations imposed by the fabrication process is very important in the development of efficient algorithms for VLSI physical design.

In this chapter we will study the basic properties of the materials used in the fabrication of VLSI chips, and details of the actual fabrication process. We will also discuss the layout of several elementary VLSI devices, and how such elementary layouts can be used to construct the layout of larger circuits.

2.1 Fabrication Materials

The electrical characteristics of a material depend on the number of 'available' electrons in its atoms. Within each atom electrons are organized in concentric shells, each capable of holding a certain number of electrons. In order to balance the nuclear charge, the inner shells are first filled by electrons and these electrons may become inaccessible. However, the outermost shell may or may not be complete, depending on the number of electrons available. Atoms organize themselves into molecules, crystals, or form other solids to completely fill their outermost shells by sharing electrons. When two or more atoms having incomplete outer shells approach close enough, their accessible outermost or valence electrons can be shared to complete all shells. This process leads to the formation of covalent bonds between atoms. Full removal of electrons from an atom leaves the atom with a net positive charge, of course, while the addition of electrons leaves it with a net negative charge. Such electrically unbalanced atoms are called *ions*.

The current carrying capacity of a material depends on the distribution of electrons within the material. In order to carry electrical current, some 'free' electrons must be available. The resistance to the flow of electricity is measured in terms of the amount of resistance in ohms (Ω) per unit length or resistivity. On the basis of resistivity, there are three types of materials, as described below:

1. **Insulators:** Materials which have high electrical resistance are called insulators. The high electric resistance is due to strong covalent bonds which do not permit free movement of electrons. The electrons can be set free only by large forces and generally only from the surface of the solid. Electrons within the solids cannot move and the surface of the stripped insulator remains charged until new electrons are reintroduced. Insulators have electrical resistivity greater than millions of GΩ-cm. The principle insulator used in VLSI fabrication is silicon dioxide. It is used to

electrically isolate different devices, and different parts of a single device to satisfy design requirements.

2. **Conductors:** Materials with low electrical resistance are referred to as conductors. Low resistance in conductors is due to the existence of valence electrons. These electrons can be easily separated from their atoms. If electrons are separated from their atoms, they move freely at high speeds in all directions in the conductor, and frequently collide with each other. If some extra electrons are introduced into this conductor, they quickly disperse themselves throughout the material. If an escape path is provided by an electrical circuit, then electrons will move in the direction of the flow of electricity. The movement of electrons, in terms of the number of electrons pushed along per second, depends on how hard they are being pushed, the cross-sectional area of the conductive corridor, and finally the electron mobility factor of the conductor. Conductors can have resistivity as low as $1\ \mu\Omega\text{-cm}$, and are used to make connections between different devices on a chip. Examples of conductors used in VLSI fabrication include aluminum and gold. A material that has almost no resistance, i.e., close to zero resistance, is called a *superconductor.* Several materials have been shown to act as superconductors and promise faster VLSI chips. Unfortunately, all existing superconductors work at very low temperatures, and therefore cannot be used for VLSI chips without specialized refrigeration equipment.

3. **Semiconductors:** Materials with electrical resistivity at room temperature ranging from $10\ \text{m}\Omega\text{-cm}$ to $1\ \text{G}\Omega\text{-cm}$ are called semiconductors. The most important property of a semiconductor is its mode of carrying electric current. Current conduction in semiconductors occurs due to two types of carriers, namely, *holes* and *free electrons.* Let us explain these concepts by using the example of semiconductor silicon, which is widely used in VLSI fabrication. A silicon atom has four valence electrons which can be readily bonded with four neighboring atoms. At room temperatures the bonds in silicon atoms break randomly and release electrons, which are called free electrons. These electrons make bonds with bond deficient ionized sites. These bond deficiencies are known as holes. Since the breaking of any bond releases exactly one hole and one free electron, while the opposite process involves the capture of one free electron by one hole, the number of holes is always equal to number of free electrons in pure silicon crystals (see Figure 2.1). Holes move about and repel one another, just as electrons do, and each moving hole momentarily defines a positive ion which inhibits the intrusion of other holes into its vicinity. In silicon crystals, the mobility of such holes is about one third that of free electrons, but charge can be 'carried' by either or both. Since these charge carriers are very few in number, 'pure' silicon crystal will conduct weakly. Although there is no such thing as completely pure crystalline silicon, it appears that, as pure crystals, semiconductors seem to have no electrical properties of great utility.

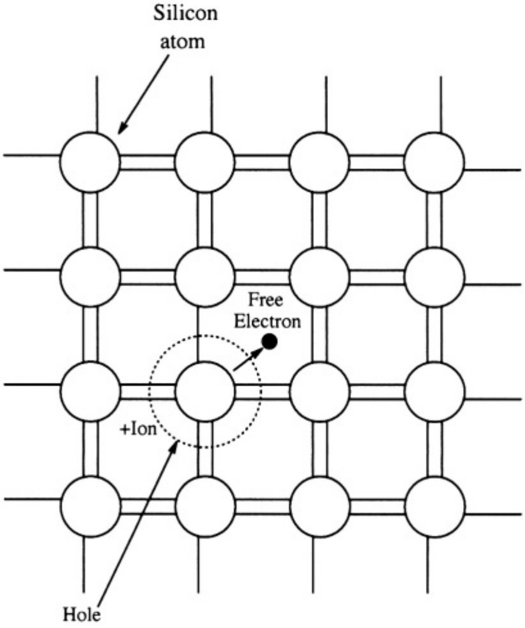


Figure 2.1: Electrons and Holes.

Semiconductor crystals can be enriched either in holes or electrons by embedding some atoms of certain other elements. This fact makes it possible to build useful devices. An atom of phosphorus has five valence electrons, whereas an atom of boron has three valence electrons. Atoms of either kind can be locked into the silicon lattice, and even a few atoms can make a dominant contribution. Once placed into a silicon lattice, the fifth valence electron of each phosphorus atom is promptly freed. On the other hand, if a boron atom is placed into a silicon lattice, the covalence deficit of a boron atom is no less promptly covered by a neighborly silicon atom, which takes a hole in exchange and passes it on. With enough phosphorus (or boron) atoms per crystal, the number of free electrons or holes in general circulation can be increased a million fold.

This process of substituting other atoms for some of the semiconductor atoms is called *doping.* Semiconductors doped with electron donors such as phosphorus are said to be of the *n-type,* while boron doping, which results in extra holes, produces *p-type* semiconductors. Though the doping elements give semiconductors desirable characteristics, they are referred to as impurities. Doping of silicon is easily accomplished by adding just the right amount of the doping element to molten silicon and allowing the result to cool and crystallize. Silicon is also doped by diffusing the dopant as a vapor through the surface of the crystalline solid at high temperature. At such temperatures all atoms are vibrating significantly in all directions. As a result, the dopant atoms can find accommodations in minor lattice defects without greatly upsetting the overall structure. The conductivity is directly related to the level of doping. The heavily doped material is referred to as n^+ or p^+. Heavier doping leads to higher conductivity of the semiconductor.

In VLSI fabrication, both silicon and germanium are used as semiconductors. However, silicon is the dominant semiconductor due to its ease of handling and large availability. A significant processing advantage of silicon lies in its capability of forming a thermally grown silicon dioxide layer which is used to isolate devices from metal layers.

2.2 Transistor Fundamentals

In digital circuits, a 'transistor' primarily means a 'switch'- a device that either conducts charge to the best of its ability or does not conduct any charge at all, depending on whether it is 'on' or 'off'. Transistors can be built in a variety of ways, exploiting different phenomenon. Each transistor type gives rise to a circuit family. There are many different circuit families. A partial list would include TTL (Transistor-Transistor Logic), MOS (Metal-Oxide-Semiconductor), and CMOS (Complimentary MOS) families, as well as the CCD (Charge-Coupled Device), ECL (Emitter-Coupled Logic), and I^2L (Integrated Injection Logic) families. Some of these families come in either *p* or *n* flavor (in CMOS both at once), and some in both high-power and low-power versions. In addition, some families are also available in both high and low speed versions. We restrict our discussion to TTL and MOS (and CMOS), and start with basic device structures for these types of transistors.

2.2.1 Basic Semiconductor Junction

If two blocks, one of n-type and another of p-type semiconductor are joined together to form a semiconductor junction, electrons and holes immediately start moving across the interface. Electrons from the n-region leave behind a region which is rich in positively charged immobile phosphorus ions. On the other hand, holes entering the interface from the p-region leave behind a region with a high concentration of uncompensated negative boron ions. Thus we have three different regions as shown in Figure 2.2. These regions establish a device with a remarkable one-way-flow property. Electrons cannot be introduced in the p-region, due to its strong repulsion by the negatively charged ions. Similarly, holes cannot be introduced in the n-region. Thus, no flow of electrons is possible in the p-to-n direction. On the other hand, if electrons are introduced in the n-region, they are passed along towards the middle region. Similarly, holes introduced from the other side flow towards the middle, thus establishing a p-to-p flow of holes.

The one-way-flow property of a semiconductor junction is the principle of

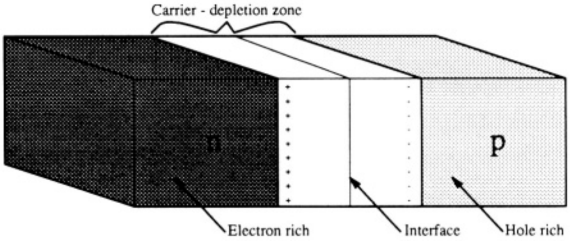


Figure 2.2: The three regions in a n-p junction.

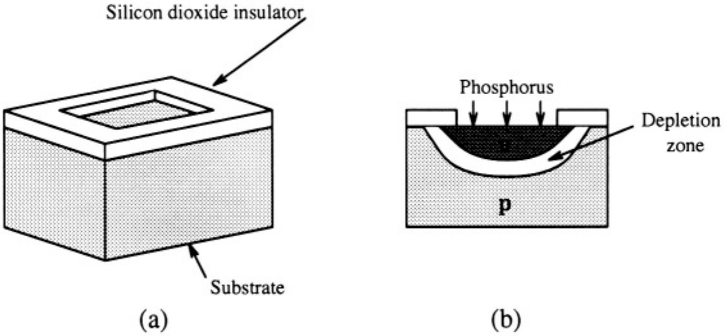


Figure 2.3: Formation of a diffused junction.

the diode, and it can be used to develop two types of devices: unipolar and bipolar. Unipolar devices are created by using the semiconductor junction, under suitable external conditions, to modulate the flow of charge between two regions of opposite polarity. On the other hand, bipolar devices are created, under suitable external conditions, by isolating one semiconductor region from another of opposite majority-carrier polarity, thus permitting a charge to flow within the one without escaping into the other.

Great numbers of both types of devices can be made rather easily by doping a silicon wafer with diffusion of either phosphorus or boron. The silicon wafer is pre-doped with boron and covered with silicon dioxide. Diffused regions are created near the surface by cutting windows into a covering layer of silicon dioxide to permit entry of the vapor, as shown in Figure 2.3. Phosphorus vapor, for example, will then form a bounded n-region in a boron doped substrate wafer if introduced in sufficient quantity to overwhelm the contribution of the boron ions in silicon. All types of regions, with differing polarities, can be formed by changing the diffusion times and diffusing vapor compositions, therefore creating more complex layered structures. The exact location of regions is determined by the mask that is used to cut windows in the oxide layer.

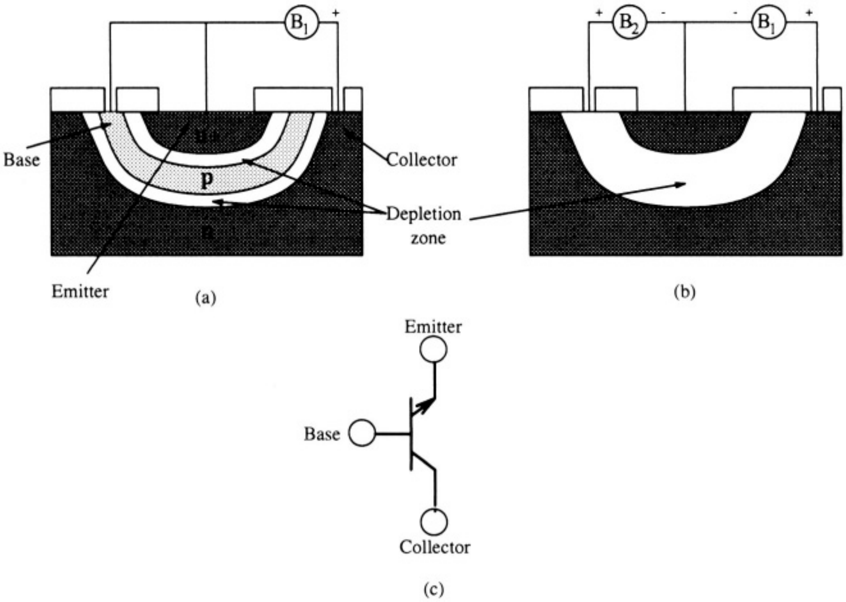


Figure 2.4: TTL transistor.

In the following, we will discuss how the semi-conductor junction is used in the formation of both TTL and MOS transistors. TTL is discussed rather briefly to enable more detailed discussion of the simpler, unipolar MOS technology.

2.2.2 TTL Transistors

A TTL transistor is an n-p-n device embedded in the surface of p-type semiconductor substrate (see Figure 2.4(a), the p-substrate is not shown). There are three regions in a TTL transistor, namely the emitter, the base, and the collector. The main idea is to control the flow of current between collector (n-region) and emitter (n^+-region) by using the base (p-region). The basic construction of a TTL transistor, both in its ‘on‘ state and ‘off’ state, along with its symbol is shown in Figure 2.4. In order to understand the operation of a TTL transistor, consider what happens if the regions of the transistor are connected to a battery B_1 as shown in Figure 2.4(a). A few charge carriers are removed from both base and collector; however, the depletion zones at the emitter-base and base-collector interfaces prevent the flow of currents of significant size along any pathway. Now if another battery with a small voltage B_2 is connected as shown in Figure 2.4(b), then two different currents begin to flow. Holes are introduced into the base by B_2, while electrons are sent into the emitter by both B_1 and B_2. The electrons in the emitter cross over into the base region. Some of these electrons are neutralized by some holes, and since the base region is rather thin, most of the electrons pass through the base and move into the collector. Thus a flow of current is established from emitter to collector. If B_2 is disconnected from the circuit, holes in the base cause the flow to stop. Thus the flow of a very small current in the B_2-loop modulates the flow of a current many times its size in the B_1-loop.

2.2.3 MOS Transistors

MOS transistors were invented before bipolar transistors. The basic principle of a MOS transistor was discovered by J. Lilienfeld in 1925, and O. Heil proposed a structure closely resembling the modern MOS transistor. However, material problems failed these early attempts. These attempts actually led to the development of the bipolar transistor. Since the bipolar transistor was quite successful, interest in MOS transistors declined. It was not until 1967 that the fabrication and material problems were solved, and MOS gained some commercial success. In 1971, nMOS technology was developed and MOS started getting wider attention.

MOS transistors are unipolar and simple. The field-induced junction provides the basic unipolar control mechanism of all MOS integrated circuits. Let us consider the n-channel MOS transistor shown in Figure 2.5(a). A p-type semiconductor substrate is covered with an insulating layer of silicon dioxide or simply oxide. Windows are cut into oxide to allow diffusion. Two separate n-regions, the source and the drain, are diffused into the surface of a p-substrate through windows in the oxide. Notice that source and drain are insulated from each other by a p-type region of the substrate. A conductive material (polysilicon or simply poly) is laid on top of the gate.

If a battery is connected to this transistor as shown in Figure 2.5(b), the poly acquires a net positive charge, as some of its free electrons are conducted away to the battery. Due to this positive charge, the holes in the substrate beneath the oxide are forced to move away from the oxide. As a result, electrons begin to accumulate beneath the oxide and form an n-type channel if the battery pressure, or more precisely the gate voltage V_g, is increased beyond a threshold value V_t. As shown in Figure 2.5(b), this channel provides a pathway for the flow of electrons from source to drain. The actual direction of flow depends on the source voltage (V_s) and the drain voltage (V_d). If the battery is now disconnected, the charge on the poly disappears. As a result, the channel disappears and the flow stops. Thus a small voltage on the gate can be used to control the flow of current from source to drain. The symbols of an n-channel MOS gate are shown in Figure 2.5(c). A p-channel MOS transistor is a device complementary to the n-channel transistor, and can be formed by using an n-type substrate and forming two p-type regions.

Integrated systems in metal-oxide semiconductor (MOS) actually contain three or more layers of conducting materials, separated by intervening layers of insulating material. As many as four (or more) additional layers of metal are used for interconnection and are called *metal1, metal2, metal3* and so on.

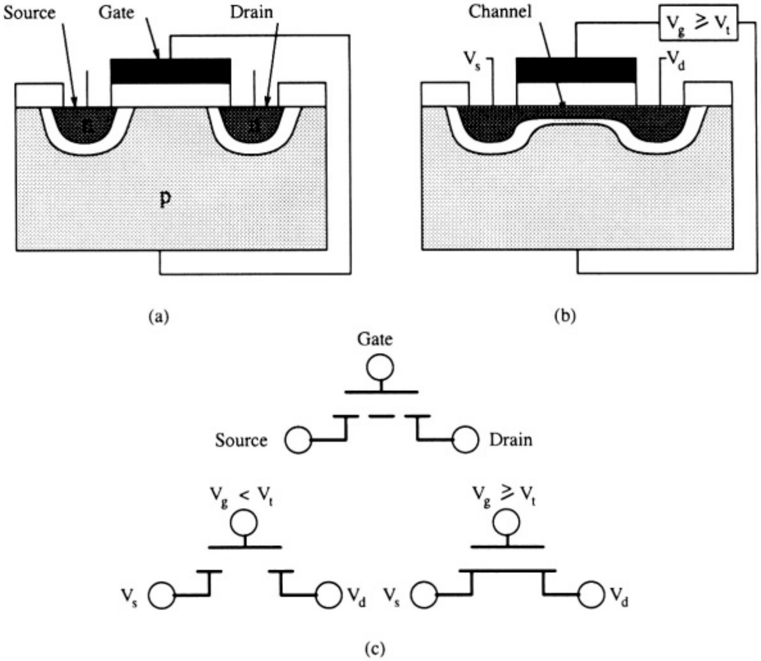


Figure 2.5: A nMOS transistor.

Different patterns for paths on different levels, and the locations for contact cuts through the insulating material to connect certain points between levels, are transferred into the levels during the fabrication process from masks. Paths on the metal level can cross poly or diffusion levels in the absence of contact cuts with no functional effects other than a parasitic capacitance. However, when a path on the poly level crosses a path on the diffusion level, a transistor is formed.

The nMOS transistor is currently the preferred form of unipolar integration technology. The name MOS survives the earlier period in which gates were made of metal (instead of poly). Aluminum is the metal of choice for all conductivity pathways, although unlike the *aluminum/oxide/semiconductor* sandwich that provides only *two* topological levels on which to make interconnections, the basic *silicon-gate* structures provide three levels, and are therefore more compact and correspondingly faster. Recent advances in fabrication have allowed the use of up to four (or more) layers of metal. However, that process is expensive and is only used for special chips, such as microprocessors. Two or three metal technology is more commonly used for general purpose chips.

The transistors that are non-conducting with zero gate bias (gate to source voltage) are called *enhancement mode transistors.* Most MOS integrated circuits use transistors of the enhancement type. The transistors that conduct with zero gate bias are called *depletion mode transistors.* For a depletion mode transistor to turn off, its gate voltage V_g must be more negative than its threshold voltage (see Figure 2.6). The channel is enriched in electrons by an implant step; and thus an n-channel is created between the source and the drain. This channel allows the flow of electrons, hence the transistor is normally in its 'on' state. This type of transistor is used in nMOS as a resistor due to poor conductivity of the channel as shown in Figure 2.6(d).

The MOS circuits dissipate DC power i.e., they dissipate power even when the output is low. The heat generated is hard to remove and impedes the performance of these circuits. For nMOS transistors, as the voltage at the gate increases, the conductivity of the transistor increases. For pMOS transistors, the p-channel works in the reverse, i.e., as the voltage on the gate increases, the conductivity of the transistor decreases. The combination of pMOS and nMOS transistors can be used in building structures which dissipate power only while switching. This type of structure is called CMOS (Complementary Metal-Oxide Semiconductor). The actual design of CMOS devices is discussed in Section 2.5.

CMOS technology was invented in the mid 1960's. In 1962, P. K. Weimer discovered the basic elements of CMOS flip-flops and independently in 1963, F. Wanlass discovered the CMOS concept and presented three basic gate structures. CMOS technology is widely used in current VLSI systems. CMOS is an inherently low power circuit technology, with the capability of providing a lower power-delay product comparable in design rules to nMOS and pMOS technologies. For all inputs, there is always a path from '1' or '0' to the output and the full supply voltage appears at the output. This 'fully restored' condition simplifies circuit design considerably. Hence the transistors in the CMOS gate do not have to be 'ratioed', unlike the MOS gate where the lengths of load and driver transistors have to be adjusted to provide proper voltage at the output. Another advantage of CMOS is that there is no direct path between VDD and GND for any combination of inputs. This is the basis for the low static power dissipation in CMOS. Table 2.1 illustrates the main differences between nMOS and CMOS technology. As shown in the table, the major drawback of CMOS circuits is that they require more transistors than nMOS circuits. In addition, the CMOS process is more complicated and expensive. On the other hand, power consumption is critical in nMOS and bipolar circuits, while it is less of a concern in CMOS circuits. Driver sizes can be increased in order to reduce net delay in CMOS circuits without any major concern of power. This difference in power consumption makes CMOS technology superior to nMOS and bipolar technologies in VLSI design.

2.3 Fabrication of VLSI Circuits

Design and Layout of VLSI circuits is greatly influenced by the fabrication process; hence a good understanding of the fabrication cycle helps in designing efficient layouts. In this section, we review the details of fabrication.

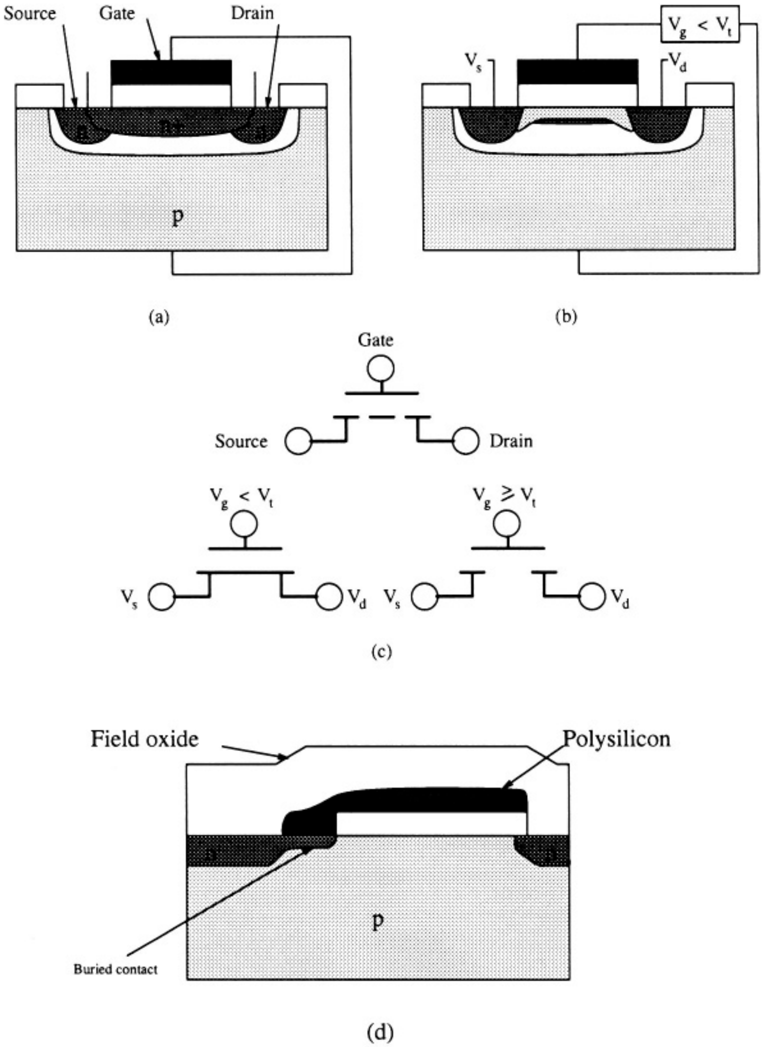


Figure 2.6: A depletion mode transistor.

CMOS	MOS
Zero static power dissipation	Power is dissipated in the circuit with output of gate at '0'
Power dissipated during logic transition	Power dissipated during logic transition
Requires 2N devices for N inputs for complementary static gates	Requires (N+1) devices for N inputs
CMOS encourages regular layout styles	Depletion, load and different driver transistors create irregularity in layout

Table 2.1: Comparison of CMOS and MOS characteristics.

Fabrication of a VLSI chip starts by growing a large silicon crystal ingot about 20 centimeters in diameter. The ingot is sliced into several wafers, each about a third of a millimeter thick. Under various atmospheric conditions, phosphorus is diffused, oxide is grown, and polysilicon and aluminum are each deposited in different steps of the process. A complex VLSI circuit is defined by 6 to 12 separate layer patterns. Each layer pattern is defined by a mask. The complete fabrication process, which is a repetition of the basic three-step process(shown in Figure 2.7), may involve up to 200 steps.

1. **Create:** This step creates material on or in the surface of the silicon wafer using a variety of methods. Deposition and thermal growth are used to create materials on the wafer, while ion implantation and diffusion are used to create material (actually they alter the characteristics of existing material) in the wafer.

2. **Define:** In this step, the entire surface is coated with a thin layer of light sensitive material called *photoresist.* Photoresist has a very useful property. The ultraviolet light causes molecular breakdown of the photoresist in the area where the photoresist is exposed. A chemical agent is used to remove the dis-integrated photoresist. This process leaves some regions of the wafer covered with photoresist. Since exposure of the photoresist occurred while using the mask, the pattern of exposed parts on the wafer is exactly the same as in the mask. This process of transferring a pattern from a mask onto a wafer is called *photolithography* and it is illustrated in Figure 2.8.

3. **Etch:** Wafers are immersed in acid or some other strong chemical agent to etch away either the exposed or the unexposed part of the pattern, depending on whether positive or negative photoresist has been used. The photoresist is then removed to complete the pattern transfer process.

This three step process is repeated for all the masks. The number of masks and actual details of each step depend on the manufacturer as well as the

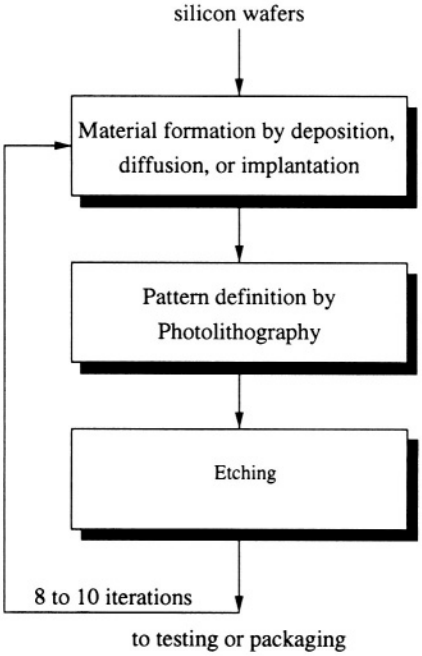

silicon wafers
Material formation by deposition, diffusion, or implantation
Pattern definition by Photolithography
Etching
8 to 10 iterations
to testing or packaging

Figure 2.7: Basic steps in MOS fabrication process.

technology. In the following analysis, we will briefly review the basic steps in nMOS and CMOS fabrication processes.

2.3.1 nMOS Fabrication Process

The first step in the n-channel process is to grow an oxide layer on lightly doped p-type substrate (wafer). The oxide is etched away (using the diffusion mask) to expose the active regions, such as the sources and drains of all transistors. The entire surface is covered with poly. The etching process using the poly mask removes all the poly except where it is necessary to define gates. Phosphorus is then diffused into all uncovered silicon, which leads to the formation of source and drain regions. Poly (and the oxide underneath it) stops diffusion into any other part of the substrate except in source and drain areas. The wafer is then heated, to cover the entire surface with a thin layer of oxide. This layer insulates the bare semiconductor areas from the pathways to be formed on top. Oxide is patterned to provide access to the gate, source, and drain regions as required. It should be noted that the task of aligning the poly and diffusion masks is rather easy, because it is only their intersections that define transistor boundaries. This self-alignment feature is largely responsible for the success of silicon-gate technology. The formation of a depletion mode

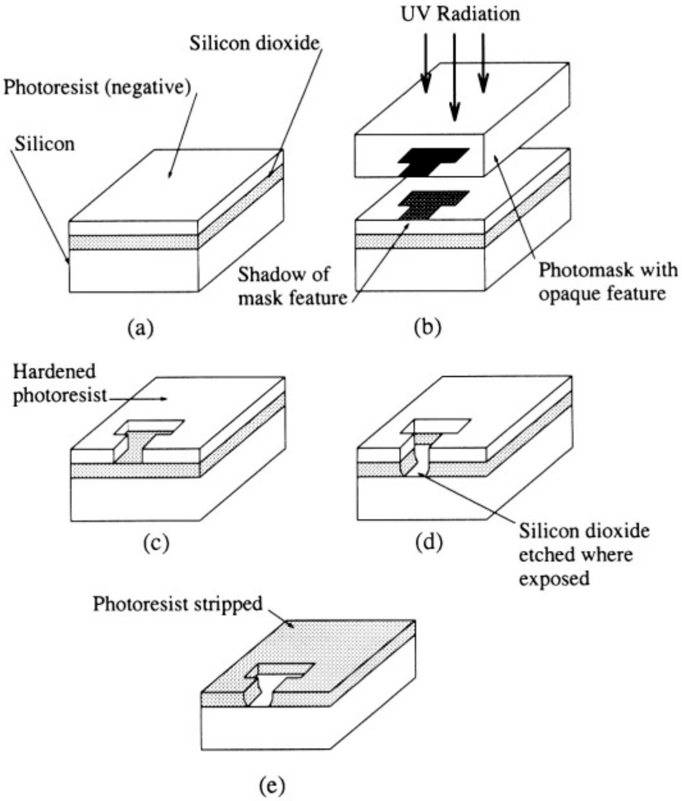


Figure 2.8: Photholithographic process.

transistor requires an additional step of ion implantation.

A thin covering of aluminum is deposited over the surface of the oxide, which now has 'hills' and 'valleys' in its structure. Etching then removes all but the requisite wires and contacts. Additional metal layers may be laid on top if necessary. It is quite common to use two layers of metal. At places where connections are to be made, areas are enlarged somewhat to assure good interlevel contact even when masks are not in perfect alignment. All pathways are otherwise made as small as possible, in order to conserve area. In addition to normal contacts, an additional contact is needed in nMOS devices. The gate of a depletion mode transistor needs to be connected to its source. This is accomplished by using a *buried contact,* which is a contact between diffusion and poly.

The final steps involve covering the surface with oxide to provide mechanical and chemical protection for the circuit. This oxide is patterned to form windows which allow access to the aluminum bonding pads, to which gold wires will be attached for connection to the chip carrier. The windows and pads are very

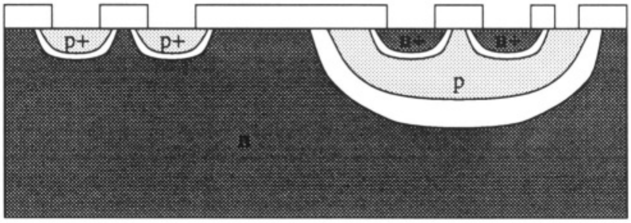


Figure 2.9: A p-well CMOS transistor.

large as compared to the devices.

2.3.2 CMOS Fabrication Process

CMOS transistors require both a p-channel and an n-channel device. However, these two types of devices require two different substrates. nMOS transistors require a p-type substrate, while pMOS transistors require a n-type substrate. CMOS transistors are created by diffusing or implanting an *n-type well* in the original p-substrate. The well is also called a *tub* or an *island.* The p-channel devices are placed in the n-well. This is called the *n-well CMOS process.* A complementary process of *p-well CMOS* starts with an n-type substrate for pMOS devices, and creates a p-well for nMOS devices. The structure of a CMOS transistor is shown in Figure 2.9 (p-substrate is not shown). A *twin-tub CMOS* process starts with a lightly doped substrate and creates both a n-well and a p-well.

As compared to the nMOS process, the CMOS process requires a additional mask for ion implanting to create the deep p-wells, n-wells, or both.

2.3.3 Details of Fabrication Processes

The complete fabrication cycle consists of the following steps: crystal growth and wafer preparation, epitaxy, dielectric and polysilicon film deposition, oxidation, diffusion, ion implantation, lithography and dry etching. These steps are used in a generic silicon process, however they are similar to those of other technologies. Below, we discuss details of specific fabrication processes.

1. **Crystal Growth and Wafer Preparation:** Growing crystals essentially involves a phase change from solid, liquid, or gas phases to a crystalline solid phase. The predominant method of crystal growth is Czochralski (CZ) growth, which consists of crystalline solidification of atoms from the liquid phase. In this method, single-crystal ingots are pulled from molten silicon contained in a fused silica crucible.

   Electronic-grade silicon, which is a polycrystalline material of high purity, is used as the raw material for the preparation of a single crystal. The conditions and the parameters during the crystal-pulling operation define many properties of the wafer, such as dopant uniformity and oxygen concentration. The ingots are ground to a cylindrical shape of precisely controlled diameter and one or more flats are ground along its length. The silicon slices are sawed from the ingot with an exact crystallographic orientation. Crystal defects adversely affect the performance of the device. These defects may be present originally in the substrate or may be caused by subsequent process steps. The harmful impurities and defects are removed by careful management of the thermal processes.

2. **Epitaxy:** This is the process of depositing a thin single-crystal layer on the surface of a single-crystal substrate. The word epitaxy is derived from two Greek words: *epi,* meaning 'upon', and *taxis,* meaning 'ordered'. Epitaxy is a Chemical Vapor Deposition (CVD) process in which a batch of wafers is placed in a heated chamber. At high temperatures (900° to 1250°C), deposition takes place when process gases react at the wafer surface. A typical film growth rate is about $1\ \mu\mathrm{m/min}$. The thickness and doping concentration of the epitaxial layer is accurately controlled and, unlike the underlying substrate, the layer can be made oxygen- and carbon-free. A limitation of epitaxy is that the degree of crystal perfection of the deposited layer cannot be any better than that of the substrate. Other process-related defects, such as slip or impurity precipitates from contamination can be minimized.

   In bipolar device technology, an *epi-layer* is commonly used to provide a high-resistivity region above a low-resistivity buried layer, which has been formed by a previous diffusion or 'implant and drive-in' process. The heavily doped buried layer serves as a low-resistance collector contact, but an additional complication arises when epitaxial layers are grown over patterned buried layer regions. To align the subsequent layers in relation to the pattern of the buried layer, a step is produced in the pre-epitaxial processing.

3. **Dielectric and Polysilicon film deposition:** The choice of a particular reaction is often determined by the deposition temperature(which must be compatible with the device materials), the properties, and certain engineering aspects of deposition (wafer throughput, safety, and reactor maintenance).

   The most common reactions for depositing silicon dioxide for VLSI circuits are:

   - Oxidizing silane (silicon hydrate) with oxygen at 400°-450°C.
   - Decomposing tetra-ethoxysilane at 650° to 750°C, and reacting dichlorosilane with nitrous oxide at 850° to 900°C.

   Doped oxides are prepared by adding a dopant to the deposition reaction. The hydrides arsine, phosphine, or diborane are often used because they are readily available gases. However, halides and organic compounds can also be used. Polysilicon is prepared by pyrolyzing silane at 600° to 650°C.

4. **Oxidation:** The process of oxidizing silicon is carried out during the entire process of fabricating integrated circuits. The production of high-quality IC's requires not only an understanding of the basic oxidization mechanism, but also the electrical properties of the oxide. Silicon dioxide has several properties:

   - Serves as a mask against implant or diffusion of dopant into silicon.
   - Provides surface passivation.
   - Isolates one device from another.
   - Acts as a component in MOS structures.
   - Provides electrical isolation of multilevel metalization systems.

   Several techniques such as thermal oxidation, wet anodization, CVD etc. are used for forming the oxide layers.

   When a low charge density level is required between the oxide and the silicon, *Thermal oxidation* is preferred over other techniques. In the thermal oxidation process, the surface of the wafer is exposed to an oxidizing ambient of O_2 or H_2O at elevated temperatures, usually at an ambient pressure of one atmosphere.

5. **Diffusion:** The process in which impurity atoms move into the crystal lattice in the presence of a chemical gradient is called diffusion. Various techniques to introduce dopants into silicon by diffusion have been studied with the goals of controlling the dopant concentration, uniformity, and reproducibility, and of processing a large number of device wafers in a batch to reduce the manufacturing costs. Diffusion is used to form bases, emitters, and resistors in bipolar device technology, source and drain regions, and to dope polysilicon in MOS device technology. Dopant atoms which span a wide range of concentrations can be introduced into silicon wafers in the following ways:

   - Diffusion from a chemical source in vapor form at high temperatures.
   - Diffusion from doped oxide source.
   - Diffusion and annealing from an ion implanted layer.

6. **Ion Implantation:** Ion implantation is the introduction of ionized projectile atoms into targets with enough energy to penetrate beyond surface regions. The most common application is the doping of silicon during device fabrication. The use of 3-keV to 500-keV energy for doping of boron, phosphorus, or arsenic dopant ions is sufficient to implant the ions from about 100 to 10,000A° below the silicon surface. These depths

place the atoms beyond any surface layers of $30A^\circ$ native SiO_2, and therefore any barrier effect of the surface oxides during impurity introduction is avoided. The depth of implantation, which is nearly proportional to the ion energy, can be selected to meet a particular application.

With ion implantation technology it is possible to precisely control the number of implanted dopants. This method is a low-temperature process and is compatible with other processes, such as photoresist masking.

7. **Lithography:** As explained earlier, lithography is the process delineating the patterns on the wafers to fabricate the circuit elements and provide for component interconnections. Because the polymeric materials resist the etching process they are called *resists* and, since light is used to expose the IC pattern, they are called *photoresists.*

   The wafer is first spin-coated with a photoresist. The material properties of the resist include (1) mechanical and chemical properties such as flow characteristics, thickness, adhesion and thermal stability, (2) optical characteristics such as photosensitivity, contrast and resolution and (3) processing properties such as metal content and safety considerations. Different applications require more emphasis on some properties than on others. The mask is then placed very close to the wafer surface so that it faces the wafer. With the proper geometrical patterns, the silicon wafer is then exposed to ultraviolet (UV) light or radiation, through a photomask. The radiation breaks down the molecular structure of areas of exposed photoresist into smaller molecules. The photoresist from these areas is then removed using a solvent in which the molecules of the photoresist dissolve so that the pattern on the mask now exists on the wafer in the form of the photoresist. After exposure, the wafer is soaked in a solution that develops the images in the photosensitive material. Depending on the type of polymer used, either exposed or nonexposed areas of film are removed in the developing process. The wafer is then placed in an ambient that etches surface areas not protected by polymer patterns. Resists are made of materials that are sensitive to UV light, electron beams, X-rays, or ion beams. The type of resist used in VLSI lithography depends on the type of exposure tool used to expose the silicon wafer.

8. **Metallization:** Metal is deposited on the wafer with a mechanism similar to spray painting. Motel metal is sprayed via a nozzle. Like spray painting, the process aims for an even application of metal. Unlike spray painting, process aims to control the thickness within few nanometers. An uneven metal application may require more CMP. Higher metal layers which are thick may require several application of the process get the desired height. Copper, which has better interconnect properties is increasing becoming popular as the choice material for interconnect. Copper does require special handling since a liner material must be provide between copper and other layers, since copper atoms may migrate into other layers due to electr-migration and cause faults.

9. **Etching:** Etching is the process of transferring patterns by selectively removing unmasked portions of a layer. Dry etching techniques have become the method of choice because of their superior ability to control critical dimensions reproducibly. Wet etching techniques are generally not suitable since the etching is isotropic, i.e., the etching proceeds at the same rate in all directions, and the pattern in the photoresist is undercut. Dry etching is synonymous with plasma-assisted etching, which denotes several techniques that use plasmas in the form of low pressure gaseous discharges. The dominant systems used for plasma-assisted etching are constructed in either of two configurations: parallel electrode (planar) reactors or cylindrical batch (hexode) reactors. Components common to both of these include electrodes arranged inside a chamber maintained at low pressures, pumping systems for maintaining proper vacuum levels, power supplies to maintain a plasma, and systems for controlling and monitoring process parameters, such as operating pressure and gas flow.

10. **Planarization:** The Chemical Mechanical Planarization (CMP) of silicon wafers is an important development in IC manufacturing. Before the advent of CMP, each layer on the wafer was more un-even then the lower layer, as a result, it was not possible to icrease the number of metal layers. CMP provides a smooth surface after each metalization step. CMP has allowed essentially unlimited number of layers of interconnect. The CMP process is like "Wet Sanding" down the surface until it is even. Contact and via layers are filled with tungsten plugs and planarized by CMP. ILD layers are also planarized by CMP.

11. **Packaging:** VLSI fabrication is a very complicated and error prone process. As a result, finished wafers are never perfect and contain many 'bad' chips. Flawed chips are visually identified and marked on the wafers. Wafers are then diced or cut into chips and the marked chips are discarded. 'Good' chips are packaged by mounting each chip in a small plastic or ceramic case. Pads on the chip are connected to the legs on the case by tiny gold wires with the help of a microscope, and the case is sealed to protect it from the environment. The finished package is tested and the error prone packages are discarded. Chips which are to be used in an MCM are not packaged, since MCM uses unpackaged chips.

The VLSI fabrication process is an enormous scientific and engineering achievement. The manufacturing tolerances maintained throughout the process are phenomenal. Mask alignment is routinely held to 1 micron in 10 centimeters, an accuracy of one part in 10^5, which is without precedent in industrial practice. For comparison, note that a human hair is 75 microns in diameter.

2.4 Design Rules

The constraints imposed on the geometry of an integrated circuit layout, in order to guarantee that the circuit can be fabricated with an acceptable yield, are called *design rules.* The purpose of design rules is to prevent unreliable, or hard-to-fabricate (or unworkable) layouts. More specifically, layout rules are introduced to preserve the integrity of topological features on the chip and to prevent separate, isolated features from accidentally short circuiting with each other. Design rules must also ensure thin features from breaking, and contact cuts from slipping outside the area to be contacted. Usually, design rules need to be re-established when a new process is being created, or when a process is upgraded from one generation to the next. The establishment of new design rules is normally a compromise between circuit design engineers and process engineers. Circuit designers want smaller and tighter design rules to improve performance and decrease chip area, while process engineers want design rules that lead to controllable and reproducible fabrication. The result is a set of design rules that yields a competitive circuit designed and fabricated in a cost effective manner.

Design rules must be simple, constant in time, applicable in many processes and standardized among many fabrication facilities. Design rules are formulated by observing the interactions between features in different layers and limitations in the design process. For example, consider a contact window between a metal wire and a polysilicon wire. If the window misses the polysilicon wire, it might etch some lower level or the circuit substrate, creating a fatal fabrication defect. One should, undoubtedly, take care of basic width, spacing, enclosure, and extension rules. These basic rules are necessary parts of every set of design rules. Some conditional rules depend on electrical connectivity information. If, for instance, two metal wires are part of the same electrical node, then a short between them would not affect the operation of circuit. Therefore, the spacing requirement between electrically connected wires can be smaller than that between disconnected wires.

The design rules are specified in terms of microns. However, there is a key disadvantage of expressing design rules in microns. A chip is likely to remain in production for several years; however newer processes may be developed. It is important to produce the chip on the newer processes to improve yield and profits. This requires modifying or shrinking the layout to obey the newer design rules. This leads to smaller die sizes and the operation is called *process shifting.* If the layout is specified in microns, it may be necessary to rework the entire layout for process shifting. To overcome this scaling problem, Mead and Conway [MC79] suggested the use of a single parameter to design the entire layout. The basic idea is to characterize the process with a single scalable parameter called *lambda* (λ), defined as the maximum distance by which a geometrical feature on any one layer can stray from another feature, due to over-etching, misalignment, distortion, over or underexposure, etc, with a suitable safety factor included. λ is thus equal to the maximum misalignment of a feature from its intended position in the wafer. One can think of λ as either

Diffusion Region Width	2λ
Polysilicon Region Width	2λ
Diffusion-Diffusion Spacing	3λ
Poly-Poly Spacing	2λ
Polysilicon Gate Extension	2λ
Contact Extension	λ
Metal Width	3λ

Table 2.2: Basic nMOS design rules.

some multiple of the standard deviation of the process or as the resolution of the process. Currently, λ is approximately 0.25×10^{-6} m ($0.25\ \mu$m). In order to simplify our presentation, we will use lambda.

Design rules used by two different fabrication facilities may be different due to the availability of different equipment. Some facilities may not allow use of a fourth or a fifth metal layer, or they may not allow a 'stacked via'. Usually, design rules are very conservative (devices take larger areas) when a fabrication process is new and tend to become tighter when the process matures. Design rules are also conservative for topmost layers (metal4 and metal5 layers) since they run over the roughest terrain.

The actual list of design rules for any particular process may be very long. Our purpose is to present basic ideas behind design rules, therefore, we will analyze simplified nMOS design rules. Table 2.2 lists basic nMOS design rules. We have omitted several design rules dealing with buried contact, implant, and others to simply our discussion.

As stated earlier, design rules are specified in fractions of microns. For example, separation for five metal layers may be $0.35\ \mu$m, $0.65\ \mu$m, $0.65\ \mu$m, $1.25\ \mu$m, and $1.85\ \mu$m respectively. Similar numbers are specified for each rule. Such rules do make presentation rather difficult, explaining our motivation to use the simpler lambda system. Although the lambda system is simple, sometimes it can become over-simplifying or even misleading. At such places we will indicate the problems caused by our simplified design rules.

In order to analyze design rules it is helpful to envision the design rules as a set of constraints imposed on the geometry of the circuit layout. We classify the rules in three types.

1. **Size Rules:** The minimum feature size of a device or an interconnect line is determined by the line patterning capability of lithographic equipment used in the IC fabrication. In 1998, the minimum feature size is $0.25\ \mu$m. Interconnect lines usually run over a rough surface, unlike the smooth surface over which active devices are patterned. Consequently, the minimum feature size used for interconnects is somewhat larger than the one used for active devices, based on patternability considerations. However, due to advances in planarization techniques, roughness problem

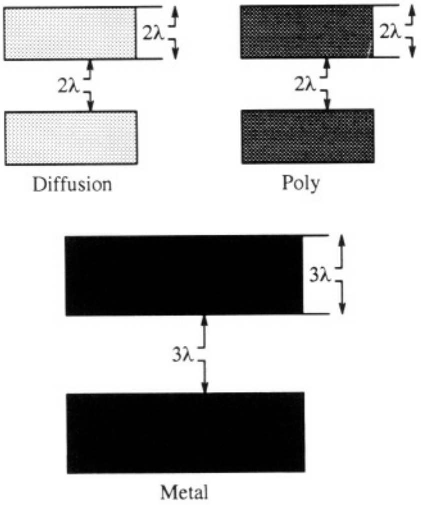


Figure 2.10: Size and separation rules.

of higher layers is essentially a solved problem.

The design rule must specify the minimum feature sizes on different layers to ensure a valid design of a circuit. Figure 2.10 shows different size rules for feature sizes in different layers.

2. **Separation Rules:** Different features on the same layer or in different layers must have some separation from each other. In ICs, the interconnect line separation is similar to the size rule. The primary motivation is to maintain good interconnect density. Most IC processes have a spacing rule for each layer and overlap rules for vias and contacts. Figure 2.10 also shows the different separation rules in terms of λ.

3. **Overlap Rules:** Design rules must protect against fatal errors such as a short-circuited channel caused by the mismigration of poly and diffusion, or the formation of an enhancement-mode FET in parallel with a depletion-mode device, due to the misregistration of the ion-implant area and the source/drain diffusion as shown in Figure 2.11. The overlap rules are very important for the formation of transistors and contact cuts or vias.

   Figure 2.12 shows the overlap design rules involved in the formation of a contact cut.

In addition to the rules discussed above, there are other rules which do not scale. Therefore they cannot be reported in terms of lambda and are reported in terms of microns. Such rules include:

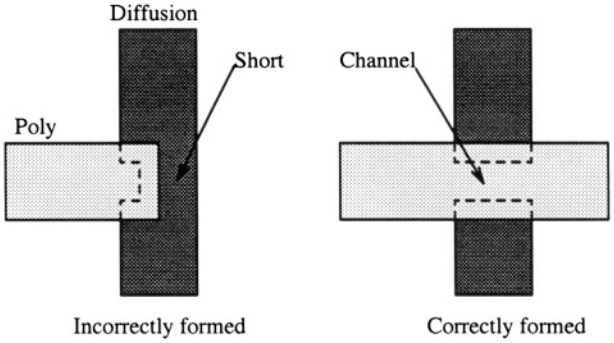


Figure 2.11: (a) Incorrectly formed channel; (b) Correctly formed channel.

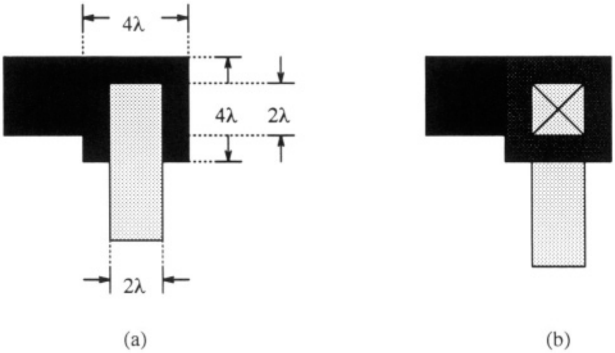


Figure 2.12: overlap rules for contact cuts.

1. The size of bonding pads, determined by the diameter of bonding wire and accuracy of the bonding machine.
2. The size of cut in overglass (final oxide covering) for contacts with pads.
3. The scribe line width (The line between two chips, which is cut by a diamond knife).
4. The feature distance from the scribe line to avoid damage during scribing.
5. The feature distance from the bonding pad, to avoid damage to the devices during bonding.
6. The bonding pitch, determined by the accuracy of bonding machine.

We have presented a simple overview of design rules. One must study actual design rules provided by the fabrication facility rather carefully before one starts the layout. CMOS designs rules are more complicated than nMOS design rules, since additional rules are needed for tubs and pMOS devices.

The entire layout of a chip is rarely created by minimum design rules as discussed above. For performance and/or reliability reasons devices are designed with wider poly, diffusion or metal lines. For example, long metal lines are sometimes drawn using using two or even three times the minimum design rule widths. Some metal lines are even tapered for performance reasons. The purpose of these examples is to illustrate the fact that layout in reality is much more complex. Although we will maintain the simple rules for clarity of presentation, we will indicate the implications of complexity of layout as and when appropriate.

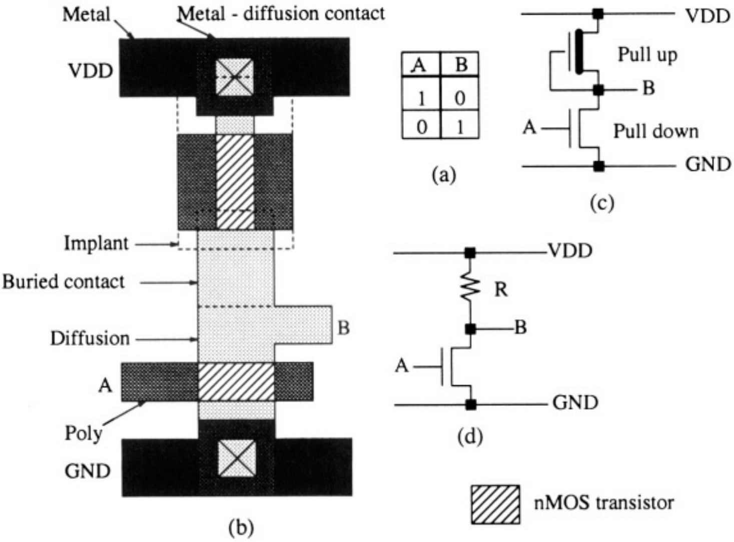
2.5 Layout of Basic Devices

Layout is a process of translating schematic symbols into their physical representations. The first step is to create a plan for the chip by understanding the relationships between the large blocks of the architecture. The layout designer partitions the chip into relatively smaller subcircuits (blocks) based on some criteria. The relative sizes of blocks and wiring between the blocks are both estimated and blocks are arranged to minimize area and maximize performance. The layout designer estimates the size of the blocks by computing the number of transistors times the area per transistor. After the top level ‘floorplan’ has been decided, each block is individually designed. In very simple terms, layout of a circuit is a matter of picking the layout of each subcircuit and arranging it on a plane. In order to design a large circuit, it is necessary to understand the layout of simple gates, which are the basic building blocks of any circuit. In this section, we will discuss the structure of various VLSI devices such as the Inverter, NAND and NOR gates in both MOS and CMOS technologies.

2.5.1 Inverters

The basic function of an *inverter* is to produce an output that is complement of its input. The logic table and logic symbol of a basic inverter are shown in Figure 2.13(a) and (d) respectively. If the inverter input voltage A is less than the transistor threshold voltage V_t then the transistor is switched off and the output is pulled up to the positive supply voltage VDD. In this case the output is the complement of the input. If A is greater than V_t, the transistor is switched on and current flows from the supply voltage through the resistor R to GND. If R is large, V_{out} could be pulled down well below V_t, thus again complementing the input.

The main problem in the design of an inverter layout is the creation of the resistor. Using a sufficiently large resistor R would require a very large area compared to the area occupied by the transistor. This problem of large resistor can be solved by using a *depletion mode* transistor. The depletion mode transistor has a threshold voltage which is less than zero. Negative voltage is required to turn off a depletion mode transistor. Otherwise the gate is always



A	B
1	0
0	1

Figure 2.13: An nMOS inverter.

turned on. The circuit diagram of an inverter is shown in Figure 2.13(c). The basic inverter layout on the silicon surface in MOS is given in Figure 2.13(b). It consists of two polysilicon (*poly*) regions overhanging a path in the diffusion level that runs between VDD and GND. This forms the two MOS transistors of the inverter. The transistors are shown by hatched regions in Figure 2.13(b). The upper transistor is called pull-up transistor, as it pulls up the output to 1. Similarly, the lower transistor is called the pull-down transistor as it is used to pull-down the output to zero. The inverter input A is connected to the poly that forms the gate of the lower of the two transistors. The pull-up is formed by connecting the gate of the upper transistor to its drain using a buried contact. The output of the inverter is on the diffusion level, between the drain of the pull-down and the source of the pull-up. The pull-up is the depletion mode transistor, and it is usually several times longer than the pull-down in order to achieve the proper inverter logic threshold. VDD and GND are laid out in metal1 and contact cuts or vias are used to connect metal1 and diffusion.

The CMOS inverter is conceptually very simple. It can be created by connecting a p-channel and a n-channel transistor. The n-channel transistor acts as a pull-down transistor and the p-channel acts as a pull-up transistor. Figure 2.14 shows the layout of a CMOS inverter. Depending on the input, only one of two transistors conduct. When the input is low, the p-channel transistor between VDD and output is in its "on" state and output is pulled-up to VDD, thus inverting the input. During this state, the n-channel transistor does not conduct. When input is high, the output goes low. This happens due to the "on" state of the n-channel transistor between GND and output. This pulls the

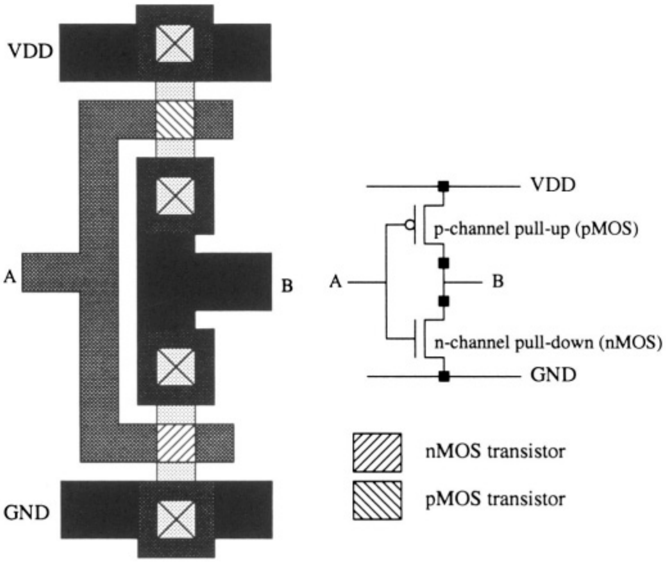


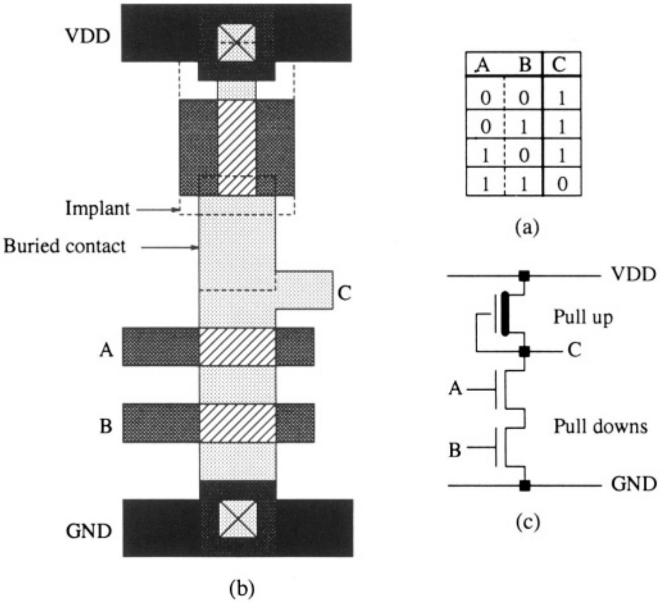
Figure 2.14: A CMOS inverter.

output down to GND, thus inverting the input. During this state, p-channel transistor does not conduct. The design rules for CMOS are essentially same as far as poly, diffusion, and metal layers are concerned. Additional CMOS rules deal with tub formation.

2.5.2 NAND and NOR Gates

NAND and NOR logic circuits may be constructed in MOS systems as a simple extension of the basic inverter circuit. The circuit layout in nMOS, truth tables, and logic symbols of a two-input NAND gate are shown in Figure 2.15 and NOR gate is shown in Figure 2.16.

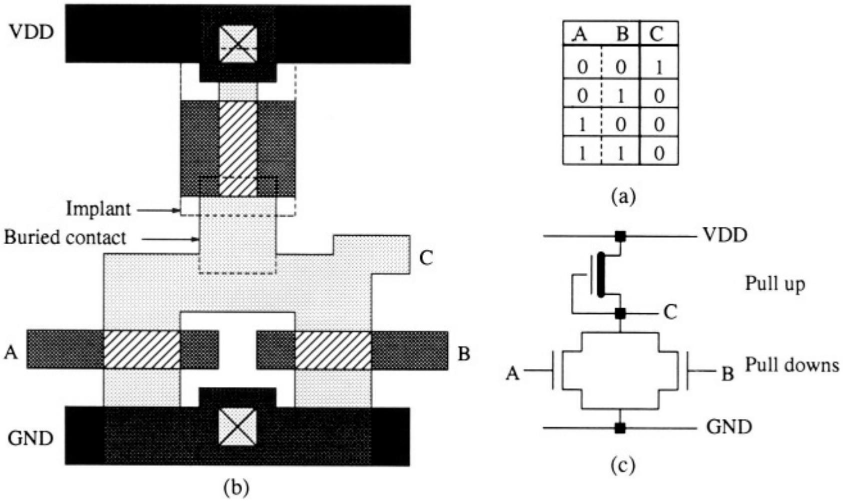
In the NAND circuit, the output will be high only when both of the inputs *A* and *B* are high. The NAND gate simply consists of a basic inverter with an additional enhancement mode transistor in series with the pull-down transistor (see Figure 2.15). NAND gates with more inputs may be constructed by adding more transistors in series with the pull-down path. In the NOR circuit, the output is low if either of the inputs, *A* and *B* is high or both are high. The layout (Figure 2.16) of a *two-input* NOR gate shows a basic inverter with an additional enhancement mode transistor in parallel with the *pull-down* transistor. To construct additional inputs, more transistors can be placed in parallel on the pull-down path. The logic threshold voltage of an n-input NOR circuit decreases as a function of the number of active inputs (inputs moving together from *logic-0* to *logic-1*). The delay time of the NOR gate with one input active



A	B	C
0	0	1
0	1	1
1	0	1
1	1	0

(a)

Figure 2.15: A nMOS NAND gate.



A	B	C
0	0	1
0	1	0
1	0	0
1	1	0

(a)

Figure 2.16: A nMOS NOR gate.

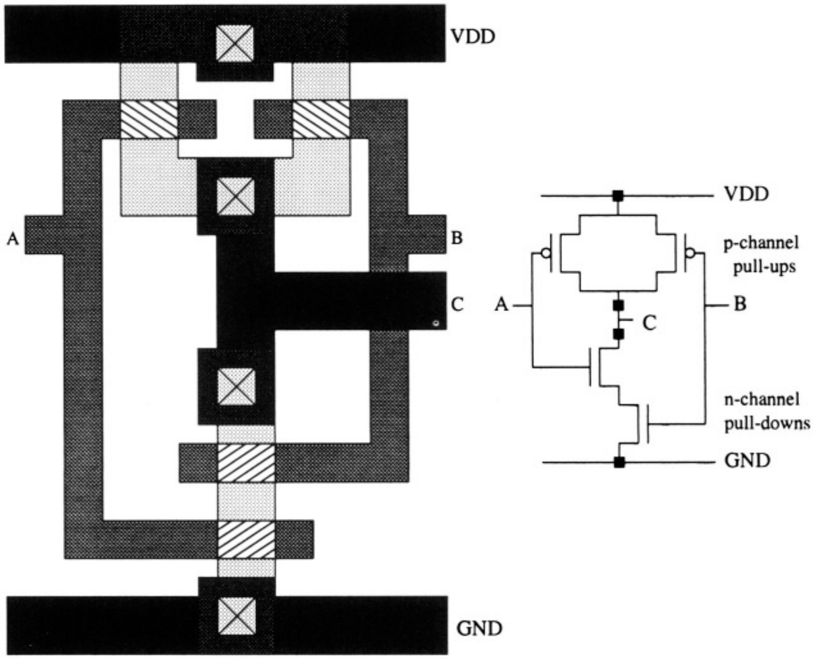


Figure 2.17: A CMOS NAND gate.

is the same as that of an inverter of equal transistor geometries, except for added stray capacitance. In designing such simple combined circuits, a single pull-up resistor must be fixed above the point of output.

The layouts of CMOS NAND and NOR gates are shown in Figure 2.17 and Figure 2.18 respectively. It is clear from Figure 2.17, that both inputs must be high in order for the output to be pulled down. In all other cases, the output will be high and therefore the gate will function as a NAND. The CMOS NOR gate can be pulled up only if both of the inputs are low. In all other cases, the output is pulled down by the two n-channel transistors, thus enabling this device to work as a NOR gate.

2.5.3 Memory Cells

Electronic memory in digital systems ranges from fewer than 100 bits from a simple four-function pocket calculator, to $10^5 - 10^7$ bits for a personal computer. Circuit designers usually speak of memory capacities in terms of bits since a unit circuit (for example a filp-flop) is used to store each bit. System designers in the other hand state memory capacities in the terms of *bytes* (typically 8-9 bits) or *words* representing alpha-numeric characters, or scientific numbers. A key characteristic of memory systems is that only a single byte or word is stored

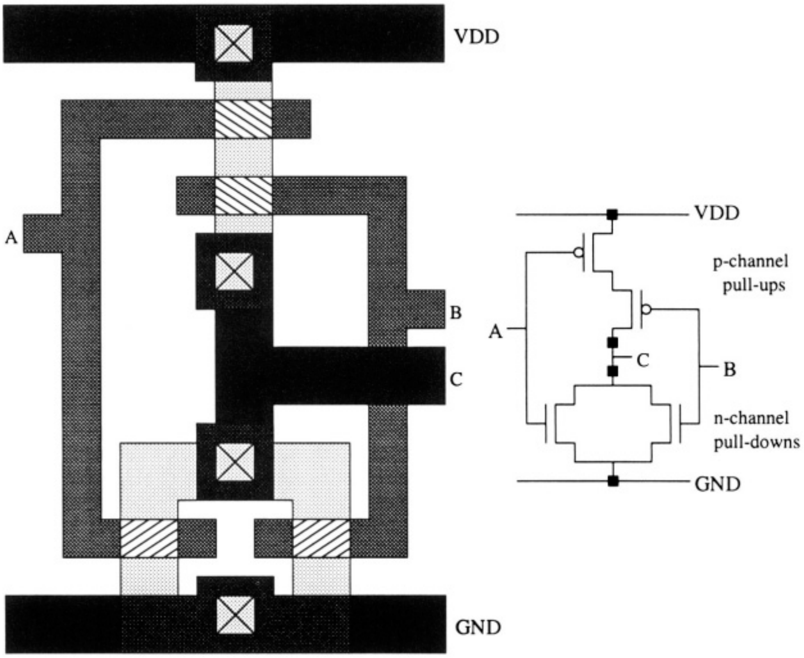


Figure 2.18: A CMOS NOR gate.

or retrieved during each cycle of memory operation. Memories which can be accessed to store or retrieve data at a fixed rate, independent of the physical location of the memory address are called Random Access Memories or RAMs. A typical RAM cell is shown in Figure 2.19.

2.5.3.1 Static Random Access Memory (SRAM)

SRAMs use static CMOS circuits to store data. A common CMOS SRAM is built using cross-coupled inverters to form a bi-stable latch as shown in Figure 2.20. The memory cell can be accessed using the pass transistors P1 and P2 which are connected to the BIT and the BIT′ lines respectively. Consider the read operation. The n-MOS transistors are poor at passing a *one* and the p-transistors are generally quite small (large load resistors). To overcome this problem, the BIT and the BIT′ lines are precharged to a n-threshold below VDD before the pass transistors are switched on. When the SELECT lines (word line) are asserted, the RAM cell will try to pull down either the BIT or the BIT′ depending on the data stored. In the write operation, data and it's complement are fed to the BIT and the BIT′ lines respectively. The word line is then asserted and the RAM cell is charged to store the data. A low on the SELECT lines, decouples the cell from the data lines and corresponds

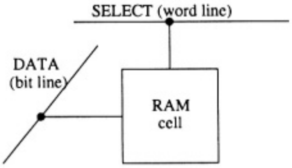


Figure 2.19: Block diagram of a generic RAM cell.

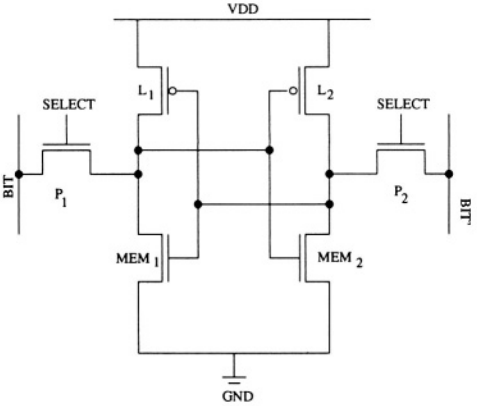


Figure 2.20: A CMOS Static RAM cell.

to a hold state. The key aspect of the precharged RAM read cycle is the timing relationship between the RAM addresses, the precharge pulse and the row decoder (SELECT line). A word line assertion preceding the precharge cycle may cause the RAM cell to flip state. On the other hand, if the address changes after the precharge cycle has finished, more than one RAM cell will be accessed at the same time, leading to erroneous *read* data.

The electrical considerations in such a RAM are simple to understand as they directly follow the CMOS Inverter characteristics. The switching time of the cell is determined by the output capacitance and the feedback network. The time constants which control the charging and discharging are

$$\tau_{ch} = \frac{C_L}{\beta_p(VDD - |V_{Tp}|)}$$

$$\tau_{dis} = \frac{C_L}{\beta_n(VDD - V_{Tn})}$$

where C_L is the total load capacitance on the output nodes and β_n and β_p are the transconductance parameters for the n and p transistors respectively.

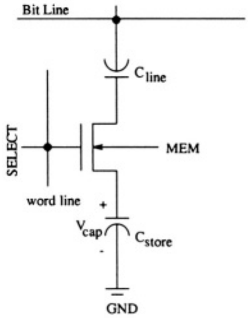


Figure 2.21: A CMOS Dynamic RAM cell.

Minimizing the time constants within the constraints of the static noise margin requirements gives a reasonable criterion for the initial design of such cells.

2.5.3.2 Dynamic Random Access Memory (DRAM)

A DRAM cell uses a dynamic charge storage node to hold data. Figure 2.21 shows a basic 1-Transistor cell consisting of an access nMOS MEM, a storage capacitor C_{store} and the input capacitance at the *bit line* C_{line}.

When the *Select* is set to high, C_{store} gets charged up to the bit line voltage according to the formula,

$$V_{cap}(t) = V_{max}[\frac{t/\tau_{ch}}{1 + t/\tau_{ch}}]$$

where $\tau_{ch} = 2C_{store}/\beta_n V_{max}$ is the charging time constant. The 90% voltage point $(0.9V_{max})$ is reached in a low-high time of $t_{LH} = 9\tau_{ch}$ and is the minimum logic 1 loading interval. Thus a logic one is stored into the capacitor. When a logic zero needs to be stored, the *Select* is again set to high and the charge on C_{store} decays according to the formula,

$$V_{cap}(t) = V_{max}[\frac{2e^{-t/\tau_{dis}}}{1 + e^{-t/\tau_{dis}}}]$$

where $\tau_{dis} = C_{store}/\beta_n V_{max}$ is discharge time constant. The 10% voltage point $(0.1V_{max})$ requires a high-low time of $t_{HL} = 2.94\tau_{dis}$. Thus it takes longer to load a logic 1 $(t_{LH} = 6.11t_{HL})$ than to load a logic 0. This is due to the fact that the gate-source potential difference decreases during a logic 1 transfer.

The read operation for a dynamic RAM cell corresponds to a charge sharing event. The charge on C_{store} is partly transferred onto C_{line}. Suppose C_{store} has an initial voltage of V_C. The bit line capacitance C_{line} is initially charged to a voltage V_{pre} (typically 3V). The total system charge is thus given by $Q_T = V_C C_{store} + V_{pre} C_{line}$. When the *SELECT* is set to a high voltage, *M*

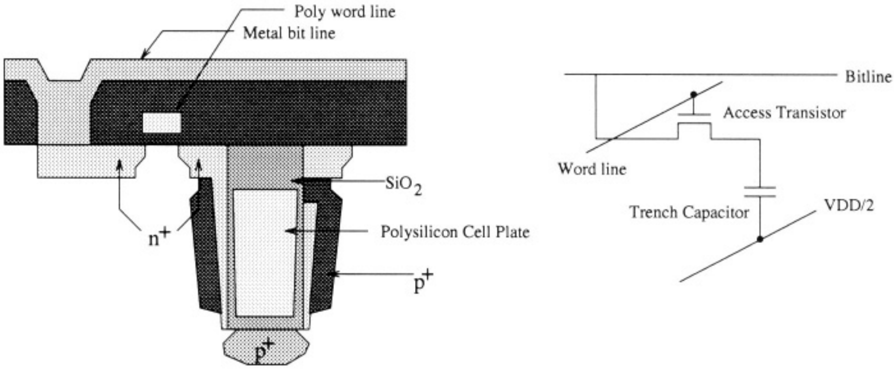


Figure 2.22: A deep trench CMOS Dynamic RAM cell.

becomes active and conducts current. After the transients have decayed, the capacitors are in parallel and equilibrate to the same final voltage V_f such that

$$V_f = \frac{V_C C_{store} + V_{pre} C_{line}}{C_{store} + C_{line}}.$$

Defining the capacitance ratio $r = C_{line}/C_{store}$ yields the final voltage V_f as

$$V_f = \frac{V_C + rV_{pre}}{1+r}.$$

If a logic 1, is initially stored in the cell, then $V_C = V_{max}$ and

$$V_1 = \frac{V_{max} + rV_{pre}}{1+r}.$$

Similarly for $V_C = 0$ volts,

$$V_0 = \frac{rV_{pre}}{1+r}.$$

Thus the difference between a logic 1 and a logic 0 is

$$\Delta V = \frac{V_{max}}{1+r}.$$

The above equation clearly shows that a small r is desirable. In typical 16 Mb designs, $C_{store} \approx 30fF$ and $C_{line} \approx 250fF$giving $r \approx 8$.

Dynamic RAM cells are subject to charge leakage, and to ensure data integrity, the capacitor must be *refreshed* periodically. Typically a dynamic refresh operation takes place at the interval of a few milliseconds where the peripheral logic circuit reads the cell and re-writes the bit to ensure integrity of the stored data.

High-value/area capacitors are required for dynamic memory cells. Recent processes use three dimensions to increase the capacitance/area. One such

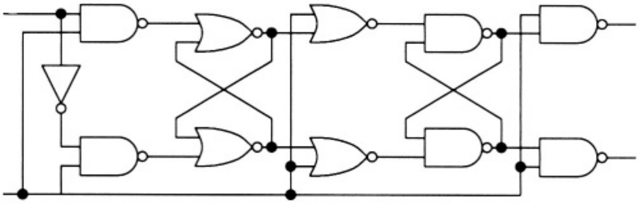


Figure 2.23: Circuit 1.

structure is the trench capacitor shown in Figure 2.22. The sides of the trench are doped n^+ and coated with a thin 10 nm oxide. Sometimes a thin oxynitride is used because its high dielectric constant increases the capacitance. The cell is filled with a polysilicon plug which forms the bottom plate of the cell storage capacitor. This is held at *VDD/2* via a metal connection at the edge of the array. The sidewall n^+ forms the other side of the capacitor and one side of the pass transistor that is used to enable data onto the bit lines. The bottom of the trench has a p^+ plug that forms a channel-stop region to isolate adjacent capacitors.

2.6 Summary

The fabrication cycle of VLSI chips consists of a sequential set of basic steps which are crystal growth and wafer preparation, epitaxy, dielectric and polysilicon film deposition, oxidation, lithography, and dry etching. During the fabrication process, the devices are created on the chip. When some fixed size material crosses another material, devices are formed. While designing the devices, a set of design rules has to be followed to ensure proper function of the circuit.

2.7 Exercises

1. Draw the layout using nMOS gates with minimum area for the circuit shown in Figure 2.23. For two input nMOS NAND gates, assume the length of pull-up transistor to be eight times the length of either pull-down. (Unless stated otherwise, use two metal layers for routing and ignore delays in vias).

2. Compute the change in area if CMOS gates are used in a minimum area layout of the circuit in Figure 2.23.

3. For the circuit shown in Figure 2.24, generate a layout in which the longest wire does not have a delay more than 0.5 psec. Assume that the

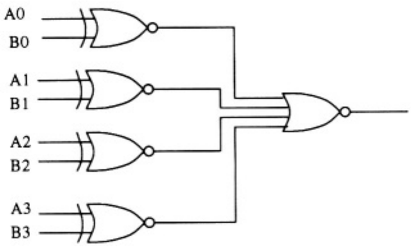


Figure 2.24: A 4-bit comparator.

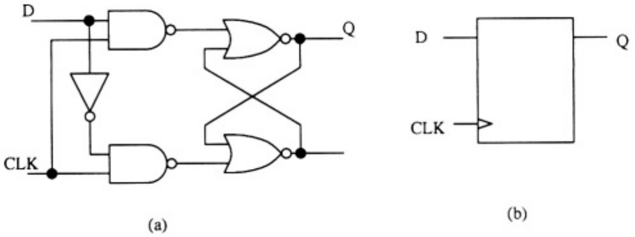


Figure 2.25: A D flip-flop.

width of the wire is 2 μm, the height is 0.5 μm, the thickness of the oxide below the wire is 1.0 μm. $\epsilon_o = 3.9$ and $\epsilon_s = 8.845 \times 10^{-14}$ F/cm.

4. Layout the circuit given in Figure 2.23 so that the delay in the longest path from input to output is minimized. Assume 2 μm CMOS process and assume each gate delay to be 2 nsec.

†5. In order to implement a memory, one needs a circuit element, which can store a bit value. This can be done using flip-flops. A D flip-flop is shown in Figure 2.25. Memories can be implemented using D flip-flops as shown in Figure 2.26. A 2 x 2 bit memory is shown in the figure. The top two flip-flops store word 0, while the bottom two flip flops store the word 1. *A* indicates the address line, if *A* = 0 the top two bits consisting of top word 0 are selected, otherwise the bottom word is selected. *CS*, *RD* and *OE* indicate chip select, read and output enable signals, respectively, I_0 and I_1 are two input lines, while O_0 and O_1 indicate output lines.

(a) Layout the 4 bit memory shown in Figure 2.26.

(b) Calculate the read and write times for this memory in 2 μm CMOS process. Assume gate delay to be 2 nsec.

(c) Estimate the total area (in microns) for a 256 KByte memory using 2 μm CMOS process.

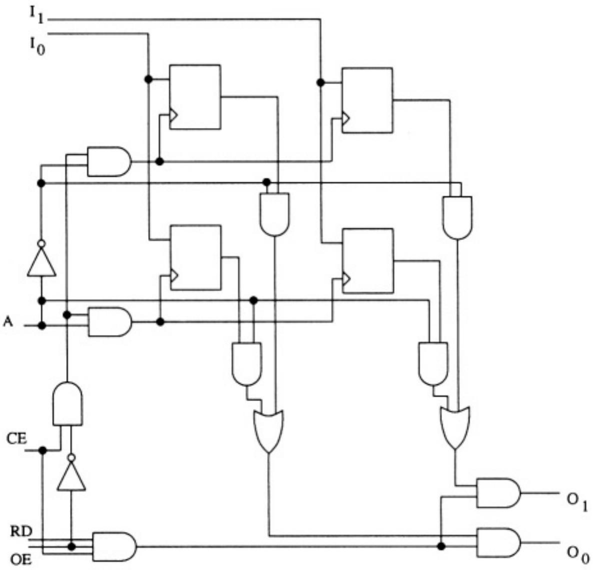


Figure 2.26: A 2×2 bit memory.

6. (a) Draw the circuit diagram of a half-adder.

   (b) Draw the layout of the of the half-adder with minimum area.

   (c) How many such chips may be fabricated on a wafer of diameter 10 cm ? (Assume scrap edge distance, $\alpha = 5$ mm.)

7. For the function $F = \bar{A}BC + A\bar{B}C + AB\bar{C} + ABC$

   (a) Draw the logic circuit diagram.

   (b) Generate a minimum area layout for the circuit.

8. Estimate the number of transistors needed to layout a k-bit full adder. Compute the area required to layout such a chip in nMOS and CMOS.

9. *Skew* is defined as the difference in the signal arrival times at two difference devices. Skew arises in the interconnection of devices and in routing of a clock signal. Skew must be minimized if the system performance is to be maximized. Figure 2.27 shows a partial layout of a chip. Complete the layout of chip by connecting signal source to all the terminals as shown in the Figure 2.27. All paths from the source to the terminals must be of equal length so as to have zero skew.

10. Suppose a new metal layer is added using present design rules. How many design rules would be needed ?

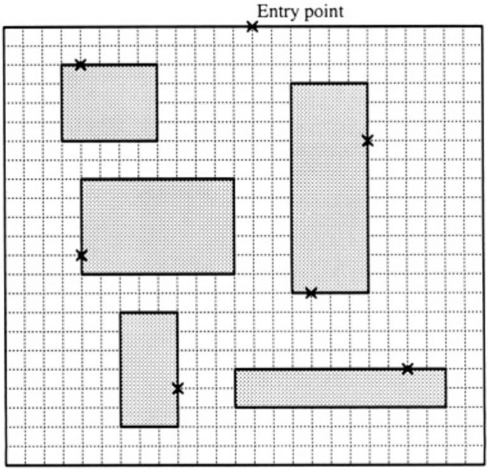


Figure 2.27: Routing with Minimum skew.

11. Compute the number of masks needed to produce a full custom chip using k-metal layer CMOS technology.

Bibliographic Notes

Weste and Eshraghian[WE92] cover CMOS design from a digital system level to the circuit level. In addition to the traditional VLSI design technologies it covers the emerging BiCMOS technology. Mead and Conway [MC79] discuss the physical design of VLSI devices. The details about the design rules can also be found in [MC79]. Advanced discussion on VLSI circuit and devices can be found in [Gia89]. The book by Bakoglu [Bak90] covers interconnects and the effects of scaling on device dimensions in detail. The basic functions of nMOS and CMOS VLSI devices are discussed in [Muk86].