

VLSI DESIGN AUTOMATION COURSE NOTES THE PRINCIPLES OF VLSI DESIGN

Peter M. Maurer
ENG 118
Department of Computer Science & Engineering
University of South Florida
Tampa, FL 33620

1. The Nature of Silicon Devices.

1.1 Conductors and Insulators.

The fundamental basis of all electronic equipment, including VLSI chips, is the controlled flow of electric current. Electric current is caused by the movement of electric charge, or more correctly, charged particles from one point to another. In most materials the charged particles, or *charge carriers*, are electrons. Materials that permit the flow of electric current are called *conductors*, and those that do not are called *insulators*. The primary difference between conductors and insulators is that in a conductor, some electrons are only weakly bound to their atoms, and may be moved with very little force, while in an insulator, the electrons are very tightly bound to their atoms and can be moved only with great difficulty. In either case, it is only the valence electrons of the substance which need be considered, because the other electrons of the substance are much too tightly bound to be used as charge carriers.

The amount of force required to create an electric current in a substance can be quantified in terms of *resistance*. The force used to create a current is termed *Electro-Motive Force* (EMF) and is measured in units of *voltage*. Electric current is determined by the amount of charge moving past a specific point in a fixed amount of time, and is measured in units of *amperage*. The resistance of a device is determined by the amount of voltage required to produce a fixed amount of current. For conductors and other materials, resistance varies directly with length, and inversely with width, or more correctly, cross-sectional area. Thus a two meter length of wire will have twice the resistance of a one-meter length of the same wire. Doubling the thickness of a wire decreases its resistance by a factor of four. Electronic materials can be classified by their *resistivity*, which is the amount of resistance per unit length for a section of material with a fixed width. Obviously, conductors have very low resistivity, while insulators have very high resistivity.

Semiconductors are materials that fall midway between conductors and insulators. Semiconductors do not conduct electricity readily enough to be useful as conductors, and at the same time, they do not have enough resistivity to be useful as insulators. Without some engineering of their properties, semiconductors would be practically useless as electronic materials.

1.2 The properties of silicon.

Silicon falls directly below carbon in the periodic table, and like carbon has four valence electrons. In its pure state, silicon is a semiconductor. The basic structure of the silicon atom is illustrated in Figure 1.

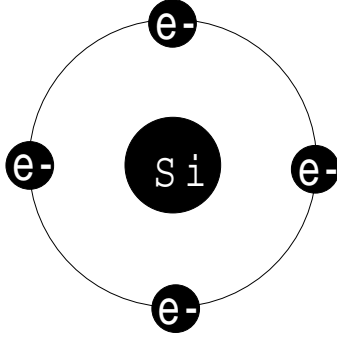


Figure 1. The Silicon Atom.

Although silicon has four valence electrons, these electrons form covalent bonds with neighboring silicon atoms, as illustrated in Figure 2, and are not available for use as charge carriers.

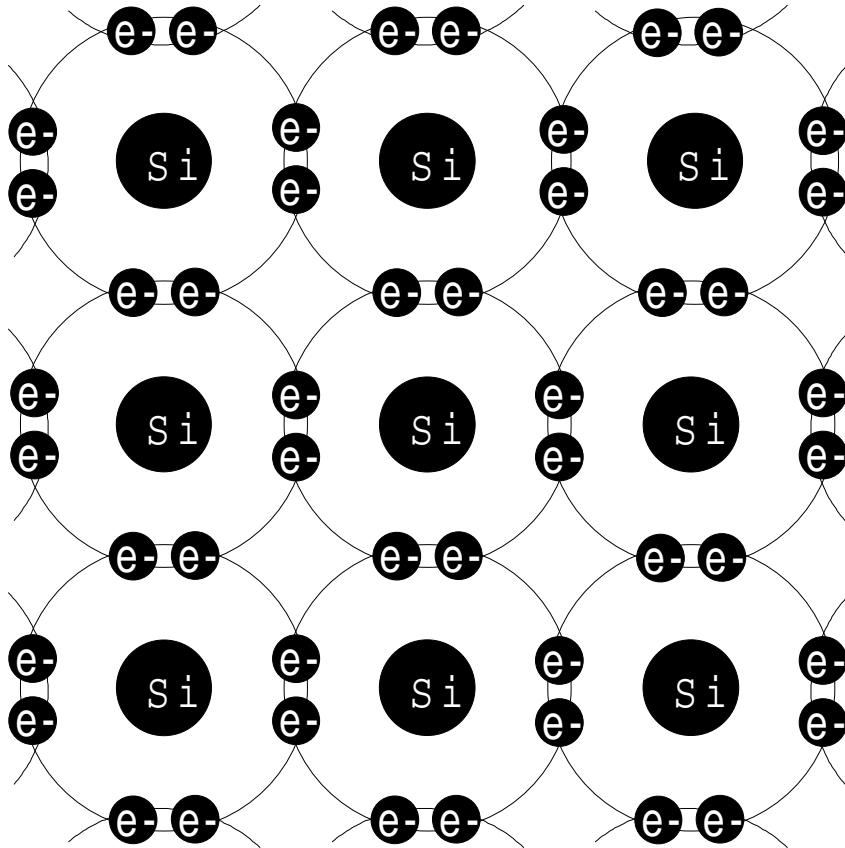


Figure 2. Silicon Crystals.

The true structure of silicon is 3-dimensional, not the 2-dimensional structure illustrated in Figure 2, but this figure illustrates that the valence electrons are bound in a structure of covalent bonds. Silicon is able to conduct a weak current, because the thermal vibration of the silicon atoms in the crystal will dislodge a few electrons, which then become charge carriers.

1.3 Doped Silicon.

Although pure silicon is a poor conductor, its electrical properties can be enhanced by introducing the right kind of impurities into the crystalline structure. The element Phosphorus has five valence electrons, one more than silicon. If a very small amount of phosphorous is introduced into the silicon crystal, a structure such as that pictured in Figure 3 will result.

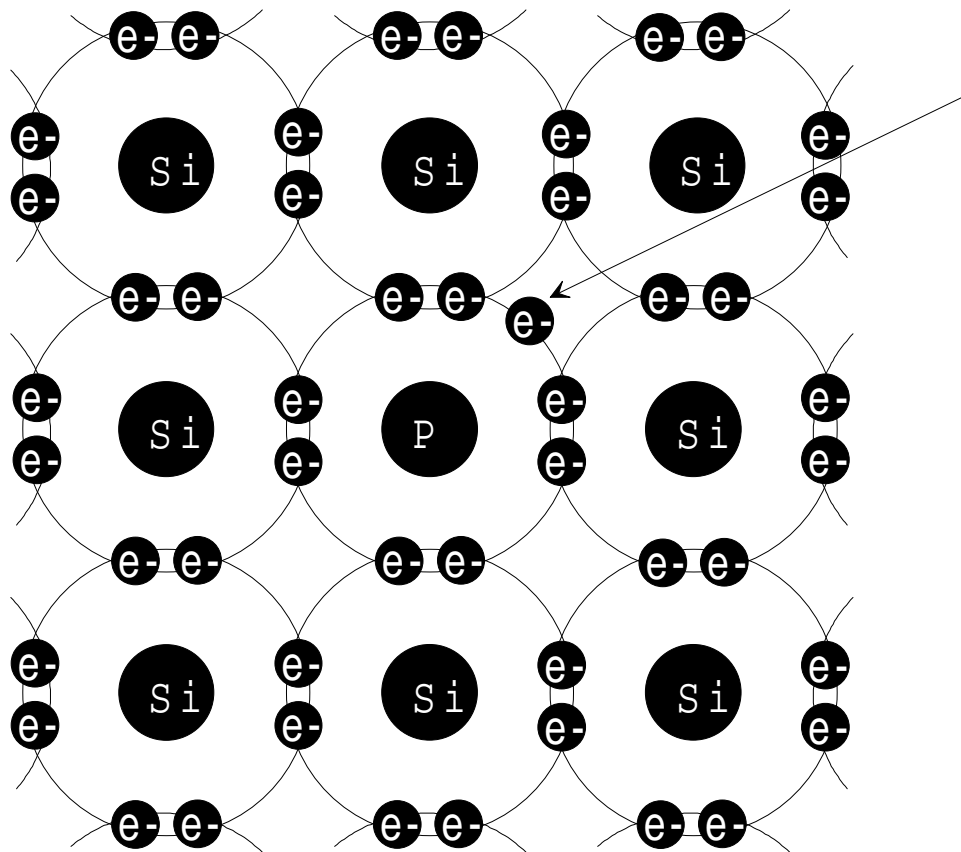


Figure 3. N-Type Silicon.

The amount of phosphorus must be so small that the crystalline structure of the silicon is not disturbed. The isolated phosphorus atoms will insert themselves into the structure as if they were silicon atoms. However, phosphorous has an extra valence electron which will not be bound into the structure. This electron can act as a charge carrier. Silicon doped with a small amount of phosphorous is called *N-Type* silicon, because the charge carriers are Negative electrons. N-Type silicon is a significantly better conductor than pure silicon.

It is also possible to introduce atoms with fewer than four valence electrons. For instance, boron has three valence electrons. When a small amount of boron is introduced into the silicon crystal, a structure such as that illustrated in Figure 4 is created.

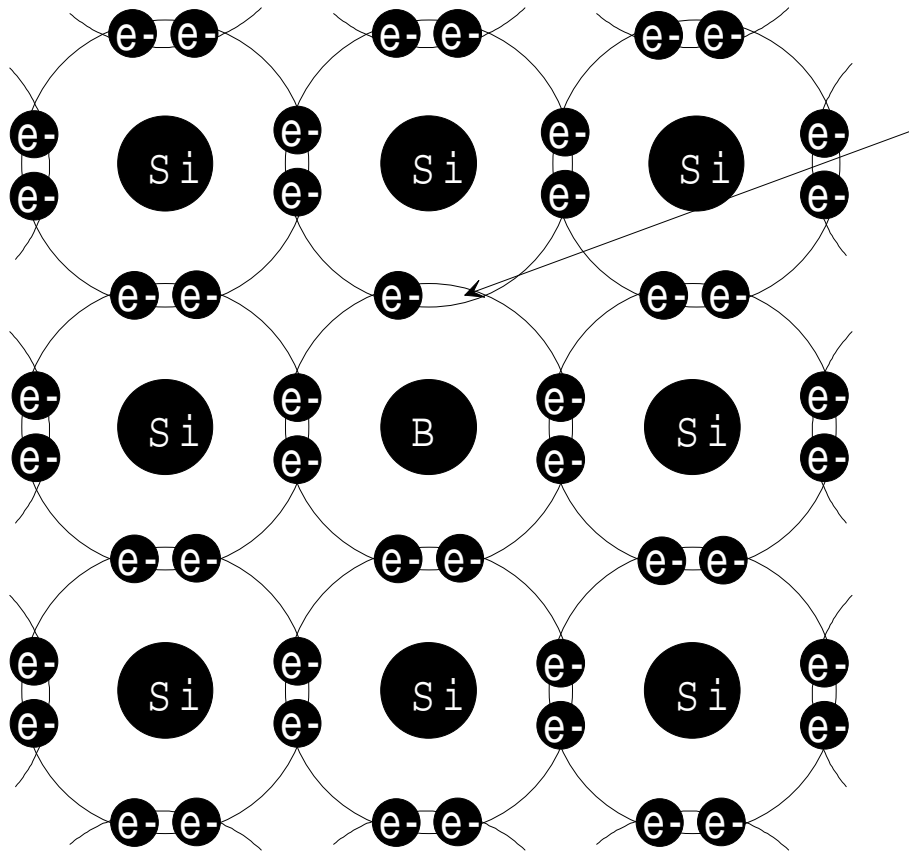


Figure 4. P-Type Silicon.

As with phosphorus, the isolated boron atoms bond into the crystalline structure as if they were silicon atoms. The missing electron creates a low-energy *hole* that can easily capture an electron from a neighboring atom. When the hole steals an electron from a silicon atom, another low energy hole appears at the silicon atom, and the atom becomes a positively charged ion. Applying a voltage will cause electrons to jump from one hole to another, causing a current to flow in the material. Although the current is due to the movement of electrons, it is convenient to think of the charge carriers being the positively charged ions moving in the opposite direction. Silicon doped with boron is known as *P-Type* silicon because the charge carriers are Positively charged holes. Holes are poorer charge-carriers than electrons, so P-Type silicon must be more heavily doped to provide the same resistivity as N-Type material.

1.4 N-P Junctions.

Neither N-Type silicon nor P-Type silicon are good conductors compared to metals. The useful properties of doped silicon only become apparent when a section of P-Type material is joined with a section of N-Type material to form a *N-P junction*, as illustrated in Figure 5.

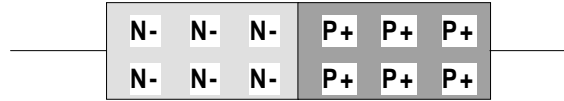


Figure 5. An N-P Junction.

The behavior of electrons near the N-P junction is different from those in pure N-type or P-type material. The holes in the P-type material can easily capture the loosely bound electrons in the N-type material. When an electron moves from the N-type material into the P-type material, it leaves a positively charged ion behind, and creates a negatively charged ion in the P-type material. Furthermore, this movement of electrons decreases the number of charge carriers in the region of the N-P junction. There is a very narrow zone around the N-P junction called the *depletion zone*, because it has been depleted of charge carriers. Furthermore, the N side of the depletion zone will be positively charged, while the P side will be negatively charged. If one attempts to cause current to flow from the P side to the N side, electrons will move through the P-type material by jumping from hole to hole. Once they reach the depletion zone, there will be no place for them to go. Furthermore, the negative charge on the P side of the junction will repel the electrons, making it extremely difficult for a current to flow through the junction.

However, if the current flow is reversed, the positive polarity of the P-side electrode will cause electrons to move out of the depletion zone into the P-type material. This will create charge carriers in the depletion zone, and will cause more electrons to be captured from the N side of the junction. These electrons will be replaced due to the negative polarity of the N-side electrode. The result will be a steady current flowing through the junction. Thus an N-P junction can be used to restrict the direction of current flow.

Things become even more interesting in devices that have more than one N-P junction. A simple transistor, such as that illustrated in Figure 6, has two N-P junctions in series.

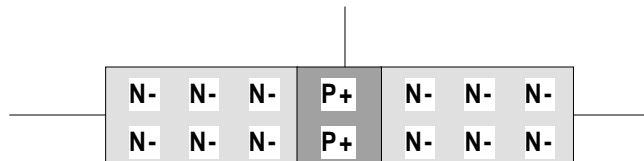


Figure 6. An N-P-N Transistor.

The transistor illustrated in Figure 6 has three electrodes, two primary electrodes attached to the N-type material, and a secondary electrode attached to the P-type material. Without using the secondary electrode, it is impossible, under normal conditions, to make a current flow between the primary electrodes. However, the third electrode can be used to alter the characteristics of the central band of P-type material permitting current flow between the main electrodes.

1.5 MOS Transistors.

There are many different types of transistors, but the primary type that is of interest to VLSI designers is the MOSFET, or Metal-Oxide-Semiconductor Field Effect Transistor. To construct such a device, one begins with a wafer of pure silicon, which has been

doped to create P-type material. The silicon wafer is a highly polished single crystal of silicon, with the atoms of the crystal aligned to the polished face. The wafer is often referred to as the *substrate*. Ordinary pure silicon is polycrystalline, which means that it consists of many regular crystalline structures joined together and aligned randomly with respect to one another. Polycrystalline silicon is unsuitable for use as substrate material, but has other uses in the fabrication of VLSI chips.

To create a MOSFET, one begins by creating two N-type regions in the substrate, as illustrated in Figure 7. These regions are created by exposing portions of the substrate to a gas such as phosphene, which bears the proper type of impurities. These impurities will diffuse into the surface of the silicon, creating the N-type material. Although the material is already doped as P-material, the N-type doping is much heavier than the P-type doping.

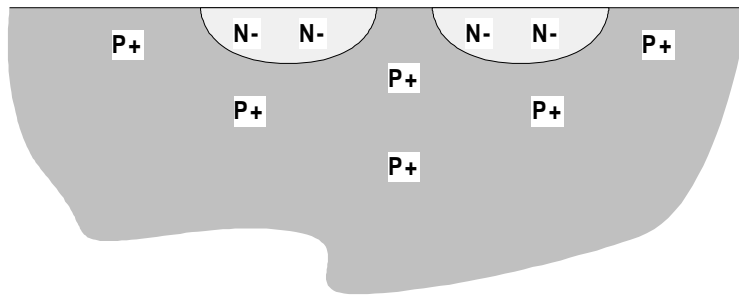


Figure 7. Two Diffused Regions.

The first step in creating a MOS transistor is to attach electrodes to the two N-type diffused regions illustrated in Figure 7. To complete the transistor the additional elements illustrated in Figure 8 are added.

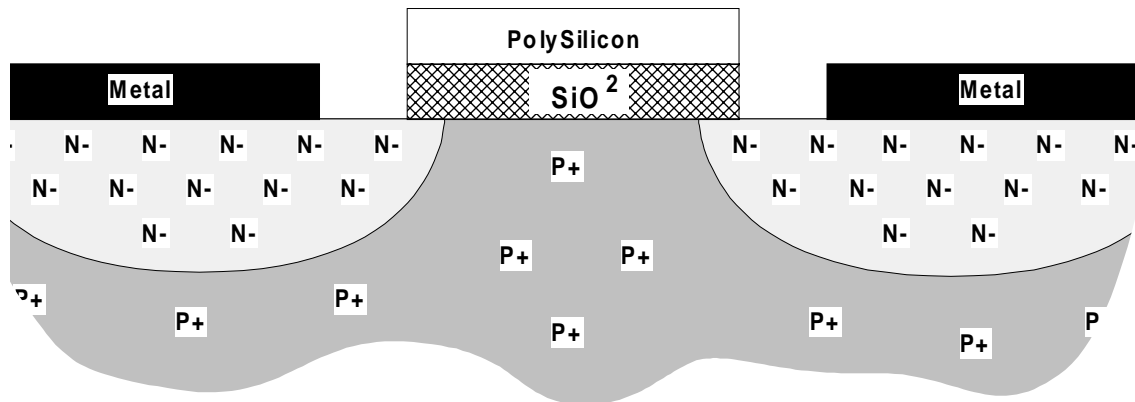


Figure 8. A MOS Transistor.

The metal contacts illustrated in Figure 8 provide electrical connections to the N-type diffused regions. The area between the two N-Type regions is coated with Silicon Dioxide (SiO_2) which is a good insulator. A layer of highly doped polycrystalline silicon (polysilicon) is placed on top of the SiO_2 . The two metal contacts are known as the *source* and the *drain* of the transistor, while the polysilicon contact is called the *gate* of the transistor. (The source and drain are identical, and can be distinguished only when the

transistor is wired into a circuit.) When a MOSFET is in its normal state, no current will flow between source and drain, however the transistor can be placed in a conducting state by placing a positive charge on the gate, as illustrated in Figure 9.

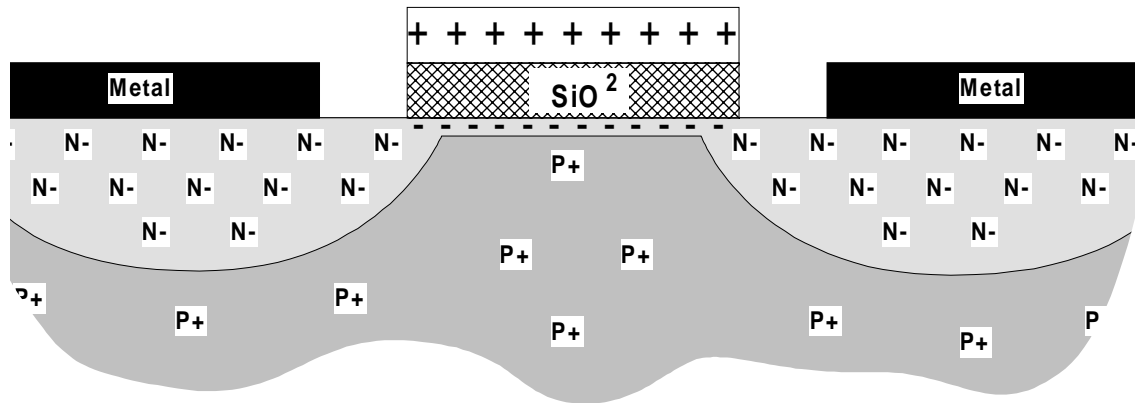


Figure 9. MOSFET in conducting state.

The positive charge on the gate of a MOSFET will attract electrons from the substrate material, converting the P-type material to N-type material. If the positive charge is strong enough, sufficient electrons will accumulate under the gate to form a continuous N-Type channel between the source and the drain. When this point is reached, current will flow readily between the source and the drain. In VLSI the MOSFETs are used as electronically controlled switches.

The MOSFET illustrated in Figure 9 is more formally known as an *enhancement mode N-Channel* MOSFET. Other types of MOSFETS are *depletion mode* transistors, and *P-Channel* transistors. Depletion mode transistors are conducting in their normal state, but can be made non-conducting by placing a charge on the gate. P-Channel transistors are created by diffusing P-type regions into an N-type substrate. The N-type substrate is usually created by diffusing a large N-type region into a P-type substrate, which allows both N-channel and P-channel transistors to be used in the same circuit. A P-Channel transistor functions much the same as an N-Channel transistor, except a negative charge on the gate is required to make the transistor conducting.

2. MOS Circuits.

2.1 Series/Parallel Connections.

One important principle in the design of digital circuits, is the equivalence of switching networks and boolean functions. xxx illustrates two switching networks which can be modeled using the and and or boolean functions.

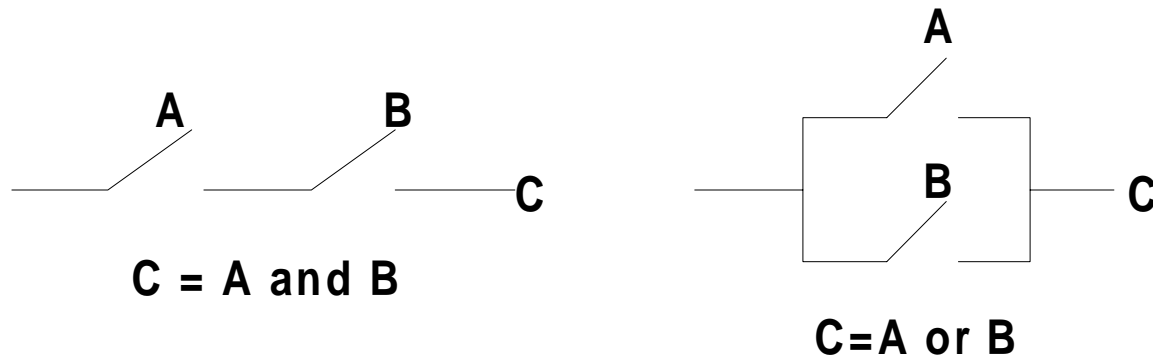


Figure 10. Series and Parallel Switches.

In the first circuit of Figure 10 current will flow in **C** only if both switches **A** and **B** are closed. In the second circuit, current will flow at **C** only if either **A** or **B** is closed.

2.2 Amplification.

Constructing logic gates is more complicated than stringing together series and parallel switches, but only slightly so. Before discussing how to construct logic gates, it is necessary to cover a bit more basic electronics. The primary reason why series and parallel connections cannot be used by themselves to construct logic gates, is that each successive connection adds resistance. A weak signal coming into a gate would be even weaker going out.. Such circuits would quickly accumulate enough resistance to become non-functional.

To avoid the problem of accumulating resistance, gates must be constructed so that they amplify their input signals. This means that a weak signal coming into a gate results in a strong signal coming out. In any digital circuit, the two strongest signals are power and ground, which represent one and zero respectively. Rather than simply passing along its input signal, a gate is constructed so that its output signal is connected to either power or ground, and weak input signal is sufficient to switch the output of the gate from one to another. To understand how this is accomplished, it is necessary to cover a bit more basic electronics. Consider the circuit pictured in Figure 11.

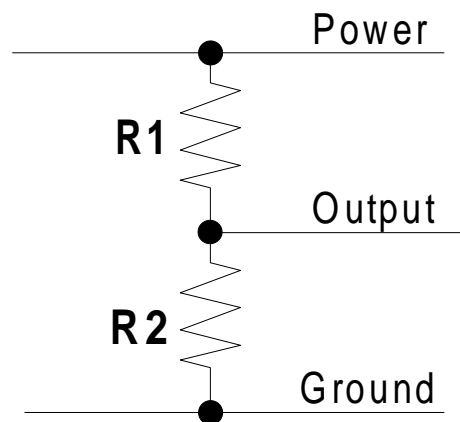


Figure 11. A Two-Resistor Circuit.

In the circuit of Figure 11, suppose that a power source has been applied to the **Power** and **Ground** terminals, so that **Power** is at +5 volts and **Ground** is at 0 volts. Unlike charge or current, voltage is not an absolute measure. Voltage is the propensity for an electric current to flow from one point to another, and can be measured only with respect to *two* terminals. To simplify matters, a reference point, called *Ground*, is chosen in the circuit. Ground is arbitrarily assigned a voltage of zero. All voltages in the circuit are measured with respect to Ground. Voltages may be either positive or negative, although in most circuits all non-zero voltages are positive. Electrons flow from Ground toward the positive voltages.

When a resistance is placed between a positive voltage and ground, there is said to be a *voltage drop* across the the resistor. In Figure 11, the two resistors in series act as a single resistor. If **Power** is at +5 volts, the total voltage drop across both resistors is 5 volts. Since the voltage drop across the total resistance, $R1+R2$, is 5 volts, the voltage drop across the resistance **R1** must be less than 5 volts. To determine the voltage drop across **R1**, one must compare the two resistances **R1** and **R2**. The total voltage drop must be prorated based on the amount of resistance provided by each resistor. If **R1** and **R2** are equal then each gets half of the voltage drop. If **R2** is twice as large as **R1** then **R2** gets two thirds of the voltage drop, and **R1** gets one third, and so forth. In calculating voltage drop, the absolute values of R1 and R2 are unimportant.. The voltage at the terminal **Output** is equal to the supply voltage (5 volts) minus the voltage drop across **R1**. If **R1** and **R2** are equal, then the voltage at Output is 2.5 volts. If **R2** is twice as large as **R1**, then the voltage at **Output** is 3.33 volts.

2.3 NMOS Gates.

Figure 11 illustrates the principle around which many logic gates are designed. The fixed resistances **R1** and **R2** are replaced with variable resistances. During operation the resistances are varied to switch the terminal **Output** between (almost) 5 volts and (almost) zero volts. One of the simplest logic gates is the NMOS inverter illustrated in Figure 12.

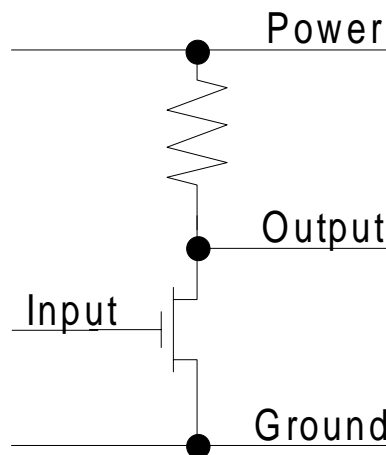


Figure 12. An NMOS inverter.

NMOS circuits are so named because they use only N-Channel transistors. The resistor illustrated in Figure 12 is chosen so that it has a very high resistance compared to the conducting state of the transistor. When the transistor is in its non-conducting state, its resistance is essentially infinity, so the voltage drop across the resistor is essentially zero. This causes the output voltage to equal the supply voltage of +5 volts. When the transistor is in its conducting state, its resistance is negligible compared to that of the resistor, causing the voltage drop across the transistor to be essentially zero. This causes the output voltage to equal the ground voltage of zero. The resistor in Figure 12 is generally formed from a depletion mode transistor which is wired to remain constantly in the conducting state, as illustrated in Figure 13, because this is the most effective way to create a resistor.

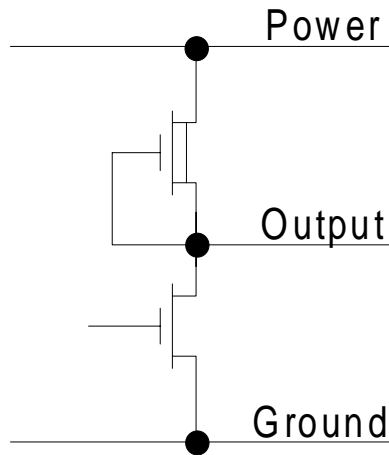


Figure 13. Inverter with Enhancement Mode Transistor.

The output of any logic gate must be connected to both power and ground through various types of electrical devices. The power connection is known as the *pull-up circuit* or simply the *pull-up*, while the ground connection is known as the *pull-down circuit* or the *pull-down*. All NMOS gates use a fixed pull-up and an active pull-down. The term *active* implies that the resistance of the connection can be dynamically altered. Using the principles of series/parallel connections described in section 2.1, it is possible to create NAND and NOR gates as illustrated in Figure 14.

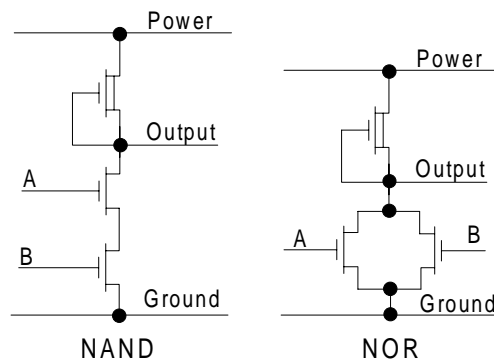


Figure 14. NMOS Nand and Nor Gates.

2.4 Cmos Gates.

The main disadvantage of NMOS gates is their high power consumption caused by the fixed pullup. When the pull-down circuit is in its conducting state, a steady current will flow between power and ground. On the other hand, when the pull-down is non-conducting, current may flow between the power and output, but not between power and ground. In most cases, the output of the gate will be the input of another gate. This implies that the output will be connected to the gate electrode of one or more transistors, and to nothing else. When the gate of a transistor is connected to power, current will flow briefly, until the gate becomes charged, then the current flow in the circuit becomes negligible.

Most early MOS chips were designed using NMOS gates because they operate faster than other alternatives. However as chips became larger and larger, the power consumption of the NMOS gate became a severe liability. The main problem is the heat generated by the circuit, which is proportional to the square of the current flow. CMOS gates solve this problem by using active pull-ups at the expense of somewhat slower gates. Improvements in processing technology have all but eliminated the disadvantages of using CMOS technology. Virtually all VLSI chips today use CMOS technology.

The active pull-up of CMOS gates is constructed from P-channel transistors. One of the simplest CMOS gates is the CMOS inverter illustrated in Figure 15.

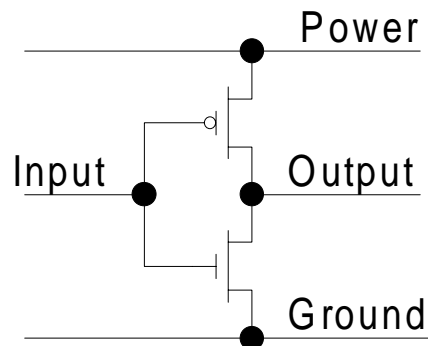


Figure 15. A CMOS Not Gate.

In Figure 15, the P-Channel transistor is attached to the power source, while the N-Channel transistor is attached to ground. A high (positive) voltage on the input will cause the N-Channel transistor to become conducting, and the P-Channel transistor to become non-conducting. A low voltage on the input will cause the P-Channel transistor to become conducting and the N-Channel transistor to become non-conducting. A high voltage on the input will cause a low voltage to appear on the output, and vice versa. When the output of the NOT gate is low, the P-Channel transistor will be non-conducting, preventing a current from flowing between Power and Ground. Similarly, when the output is high, the N-Channel transistor will be non-conducting. This means that the gate will not consume power in either state. There will be some power consumption when the gate switches from one state to another.

CMOS circuits were originally used only for very low power circuits, such as those used in watches and telephones. However, as circuits became larger and denser, the high power consumption of NMOS circuits began to cause serious heat problems. Today's

densest VLSI chips could not be constructed using NMOS gates, due to the enormous amount of heat generated by the circuit. Switching from NMOS to CMOS has not eliminated heat problems. CMOS gates consume power when switching states, and the amount of switching activity is related to the clock rate of the circuit. Today's clock-rates of 100MHz or more are causing new heat problems in the densest VLSI circuits. To combat this problem, many manufacturers are redesigning their chips to use an operating voltage of 3.5 volts, rather than the old standard of 5 volts.

As mentioned above, the source and drain of a MOSFET cannot be distinguished until the transistor has been wired into a circuit. When this has been done, the connection wired to the highest voltage is called the *source*, while the other is called the *drain*. The relative voltage appearing on the gate and the drain of a transistor is what causes charge to accumulate on the gate. When the gate voltage is positive with respect to the drain voltage, a positive charge accumulates, when the gate voltage is negative with respect to the drain, a negative charge accumulates. This explains why a zero voltage (or a very low positive voltage) can be used to switch the P-Channel transistor. Because of the positioning of the P-Channel transistor in the circuit, the voltage at the drain will be significantly higher than the zero voltage at the gate. This will cause a negative charge to accumulate on the gate, which will cause the transistor to become conducting. This also (partly) explains why N-Channel transistors are not used in the pull-up, nor P-Channel transistors in the pull-down.

When constructing more complex CMOS gates, it is important to remember that a parallel connection in the pull-down circuit must be matched by a series connection in the pull-up circuit, and vice versa. (This gives rise to the name CMOS: *Complementary MOS*.) CMOS NAND and NOR gates are illustrated in Figure 16.

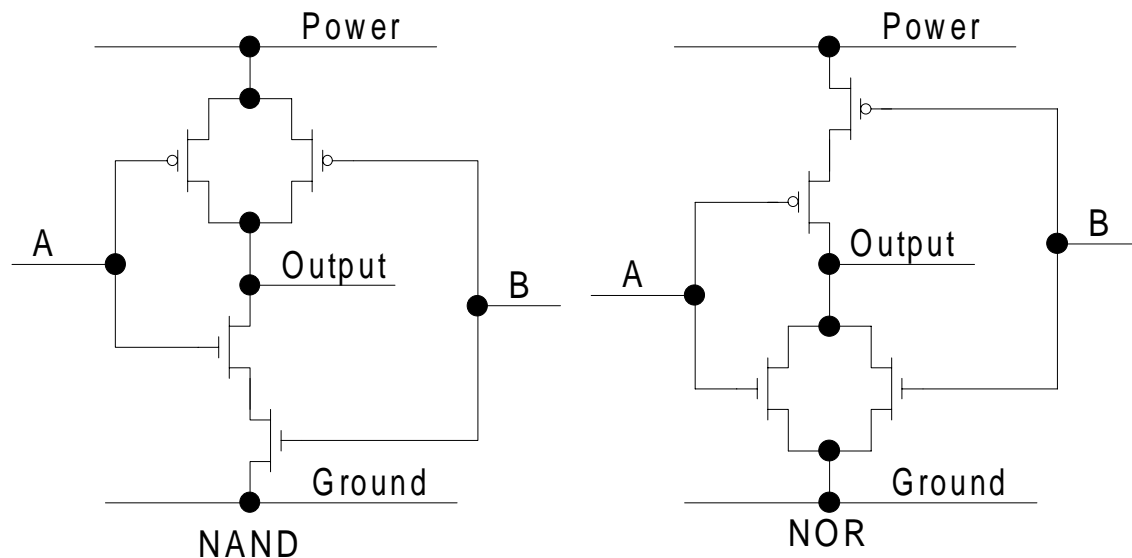


Figure 16. CMOS Nand and Nor Gates.

3. Layout and Fabrication.

3.1 Layout Design Techniques.

VLSI circuits are constructed by arranging rectangles of various types to form wires and transistors. The rectangles are placed in individual layers representing P and N-type diffusion, polysilicon, and one or more layers of metal. The layers are isolated from one another with a layer of insulator. Special rectangles are used to represent *contact cuts*, which are electrical connections between layers. Figure 17 illustrates the layout of a CMOS NAND gate.

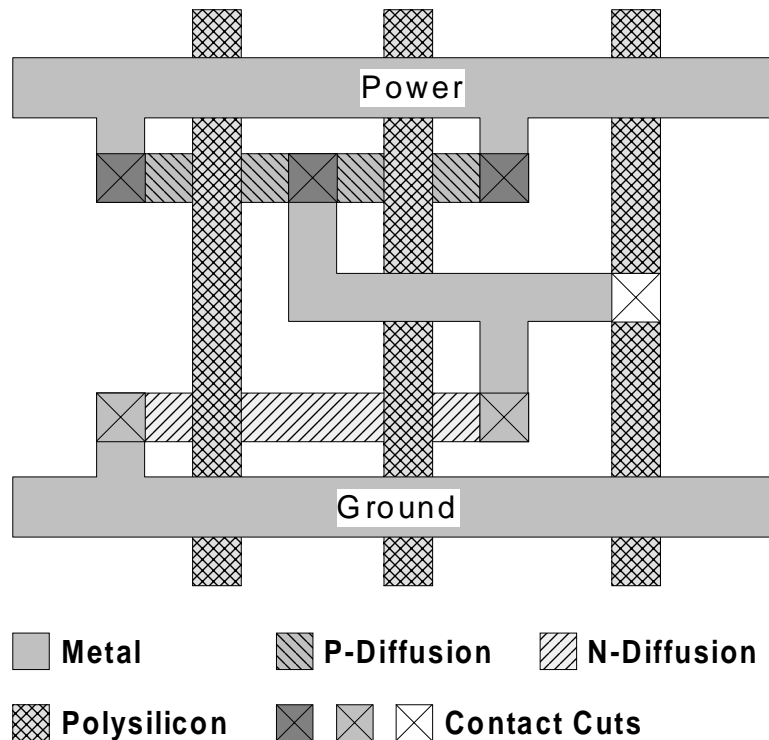


Figure 17. The Layout of a NAND Gate.

The NAND layout pictured in Figure 17 is somewhat unrealistic, because too much polysilicon has been used for conductors. In a more realistic layout only short strips of polysilicon would be used, joined by first or second layer metal. The pictured layout *might* be realistic in a single-metal-layer fabrication process. The polysilicon strips have been extended above and below the power and ground connections to form connections for the gate inputs and output. Power and ground connections have been extended to the right and left for easy attachment to neighboring gates. Another item missing from the pictured layout is the N-Well surrounding the P-Diffusion. Without this diffused N-Well, P-Diffusion would not form a transistor. In most cases, layout design tools will automatically add an N-Well surrounding any P-Diffusion. If *you* are the tool designer, you must be aware of this and provide for it.

3.2 Fabrication Processes.

The lowest layer in a layout is the N-Well layer, followed by P or N diffusion. The next highest layer is the polysilicon layer, followed by one or more layers of metal. When more than one layer of metal is used, the layers are called *first-level metal*, *second-level metal*, and so forth, starting with the bottommost layer. For reasons that need not concern us here, each successive layer of metal is harder to create than the previous, which limits the number of metal layers used in most fabrication processes. Each layer is isolated from the one below it by a layer of insulating material, usually Silicon Dioxide, which is an excellent insulator.

The fabrication process begins with a circular wafer of pure mono-crystalline silicon. The wafer is sliced from a large single crystal of silicon, with the saw aligned as closely as possible to the crystal structure. The surface of the wafer is then polished. Ideally, the surface of the wafer should be a single layer of atoms, without dislocations. The silicon is given a light P-Type doping during the crystallization process.

The various layers of the circuit are created using a photographic process that uses light sensitive substances known as *Photoresists*. A mask is made for each layer of the circuit indicating which areas of the circuit contain elements from that layer. For example, the N-Diffusion mask indicates where N-Type diffusion should be created. The surface of the wafer is coated with photoresist, and exposed to light through the mask. (The light is usually in the ultraviolet range or above.) Exposure to light makes the photoresist harder to wash off, so after exposure and washing, only certain areas of the wafer will be covered with photoresist. When the wafer is exposed to a gas, to create the N-Diffusion layer for example, only those areas not covered by photoresist are affected. After exposure, the remainder of the photoresist is removed, and the process begins again with the next layer.

Obviously, it is necessary for each successive layer to be aligned with the lower layers. Nowhere is this more critical than in the alignment of polysilicon and diffusion to create transistors. To guarantee correct alignment of diffusion with the polysilicon layer, the diffusion layer is actually created *after* the polysilicon layer. The first step is to heat the wafer and expose it to oxygen. This creates a layer of silicon dioxide covering the wafer. Then the polysilicon layer is added on top of the silicon dioxide. Once the polysilicon layer has been created, cuts are made in the silicon dioxide to expose the pure silicon beneath. These areas are then exposed to the proper mixture of gasses to create the diffused areas. This processing method guarantees that the diffusion will be precisely aligned with the polysilicon gates. This also implies that any time polysilicon crosses diffusion, a transistor is created.

4. Networks of gates.

The gate layout illustrated in Figure 17 is typical of most gate layouts. A generic gate layout is illustrated in Figure 18.

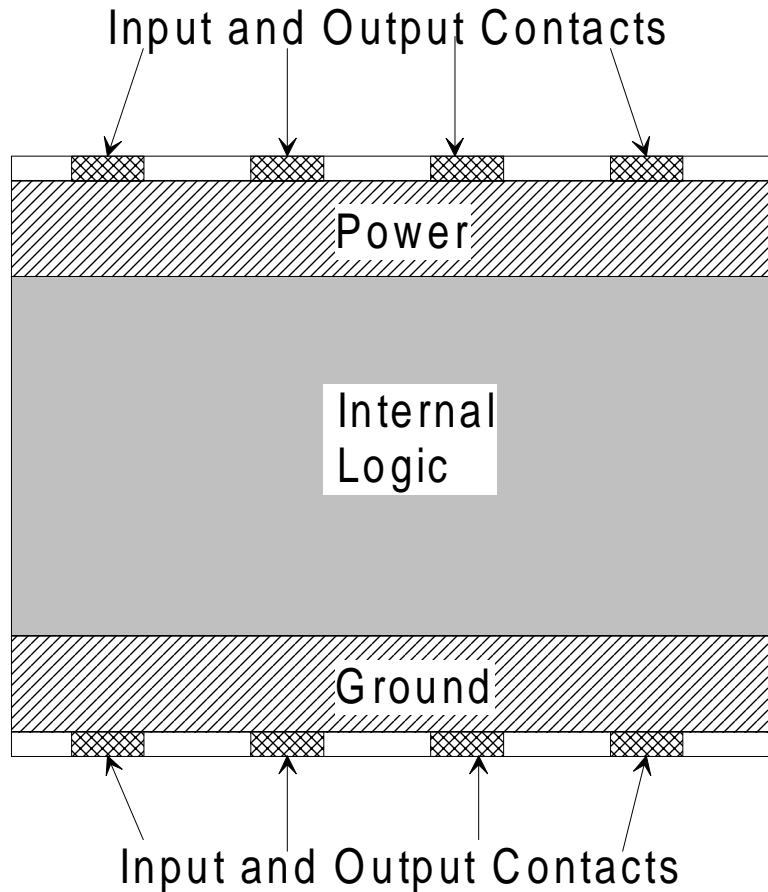


Figure 18. A Typical Layout of a Standard Cell.

The generic layout illustrated in Figure 18 is known as a *Standard Cell*. A standard cell can be rotated 90 or 180 degrees in either direction without changing its characteristics. It can also be reflected about the X or Y axis. Because the standard cell is simply a collection of rectangles, these operations are quite trivial. Standard cells are grouped into libraries of compatible cells. Two cells are compatible when they are of the same height, so that the power and ground connections are in the same position. More sophisticated layout tools will allow cells to be stretched into position if they are not initially compatible.

The positioning of the power and ground rails in the standard cell allows several cells to be placed in a row creating a single pair of power and ground connections at the ends of the row. This form of layout is illustrated in Figure 19.

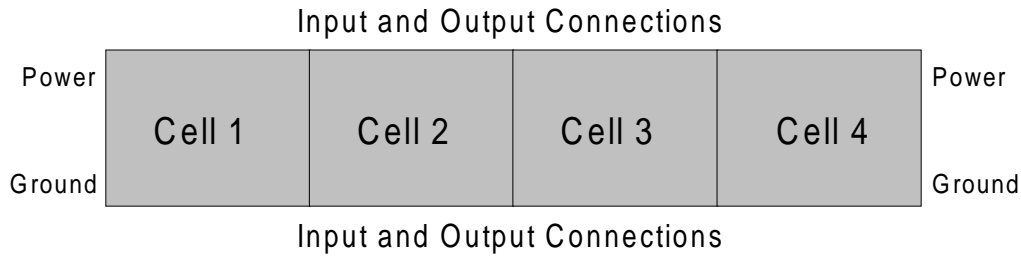


Figure 19. A Row of Standard Cells.

A row of cells can become quite long, but because it extends in only one dimension, most circuits consist of several rows, as illustrated in Figure 20.

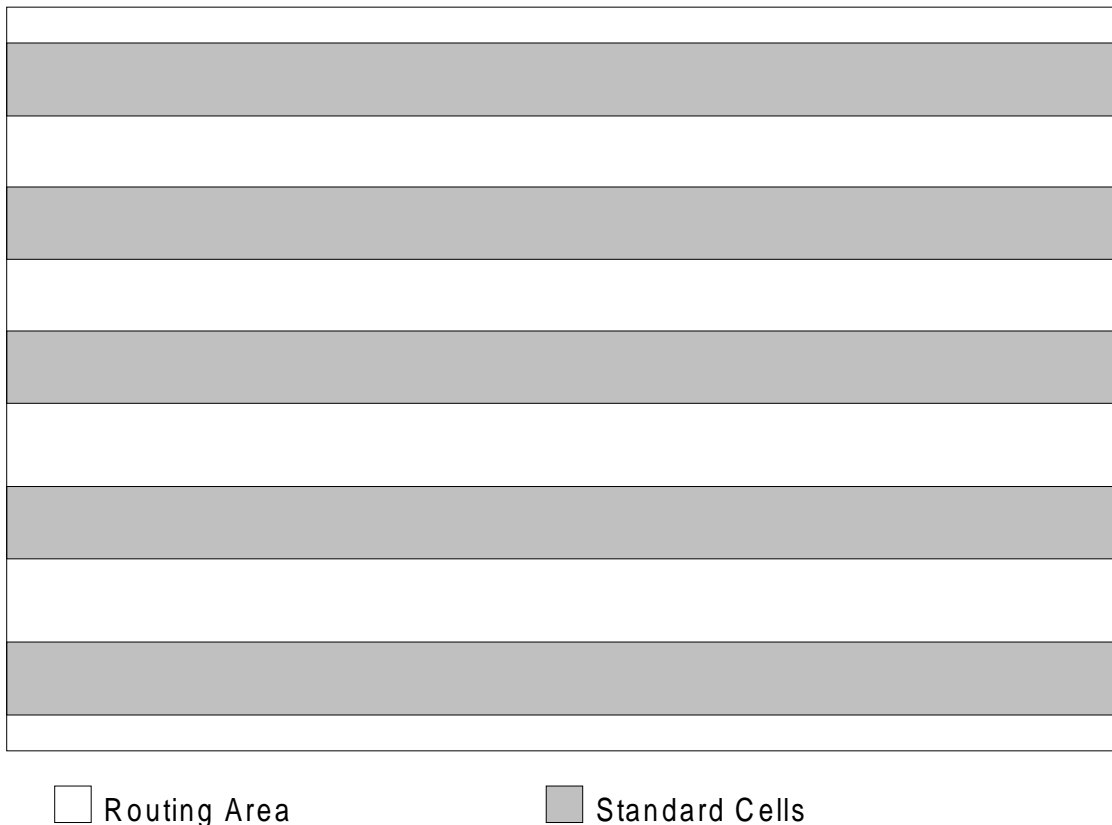


Figure 20. A Circuit Created from Standard Cells.

Creating circuits such as that illustrated in Figure 20 poses several problems for the Design Automation Specialist. Some standard cells should be placed as close as possible to one another to minimize the length of the connections between them. In fact, the size of the routing areas illustrated in Figure 20 is highly dependent on the method used to arrange the cells. It is usually extremely important to minimize the total area occupied by the circuit, so it is important to arrange the cells in such a way as to minimize the routing area between the rows. This procedure is known as *placement*.

Once the cells have been arranged in rows as optimally as possible, then it is necessary to create the connections between the cells, as well as the connections to the

“outside world.” Various *routing* algorithms are used to create these connections. For the circuit pictured in Figure 20, the routing problem is not particularly difficult because there are fixed connections only at the top and bottom of each routing area. It may be necessary to create an external connection at the end of a row, but this connection can be allowed to “float” to the most convenient vertical position.

For reasons that are too complicated to discuss here, it is generally not possible to create an entire circuit as a collection of rows of standard cells. The circuit is usually *partitioned* either automatically or along functional lines. For example, in a microprocessor, the ALU might be treated as a single circuit and the register array as another. If a sub-circuit is large enough it may be necessary to partition it into two or more smaller circuits for layout. When this is done, the partitioning is usually at least partially automated. There are several partitioning algorithms that are used for this purpose.

Once all subcircuits have been laid out, the circuit consists of a collection of irregularly shaped rectangles that must be positioned and connected to form the final circuit. Arranging irregularly shaped rectangles is known as *floorplanning*. Floorplanning is one of the least well optimized steps in the VLSI design process, although progress is continually being made.

After the floorplan of the circuit has been completed, it is necessary to connect all the rectangles and attach them to the I/O pads at the edge of the circuit. Various types of routing algorithms are used for this purpose, but some of these algorithms may be quite different from those used standard cell routing.

These techniques describe the core problems of Physical Design Automation.