

Diodes and Transistors — How They Work

Now that we are familiar with electricity, and how dc electricity has current in one direction only, and that the current changes direction continuously in cycles in ac electricity, we should be able to understand how semiconductor diodes and transistors operate. The operation of all semiconductor devices is based on the same principles. To begin to understand these principles, we will start with a discussion of diodes and the P-N junction.

Diodes: How the Simplest Semiconductor Devices Work

The simplest semiconductor device is a diode. As shown in *Figure 4-1*, it is made up of a junction of N and P semiconductor material. We will be concerned only with silicon diodes. Diodes are also made from other semiconductor materials, such as germanium and gallium arsenide. Diodes made from these materials work essentially the same way.

Rectification Is a Form of Switching

It is easy to see that a semiconductor diode is basically an electrically-controlled switch. As an example, consider the silicon diode as shown in *Figure 4-1*. The working part of the diode is a specially processed piece of silicon that has two regions—an anode (explained later as a P-type semiconductor region), and a cathode (explained later as an N-type semiconductor region).

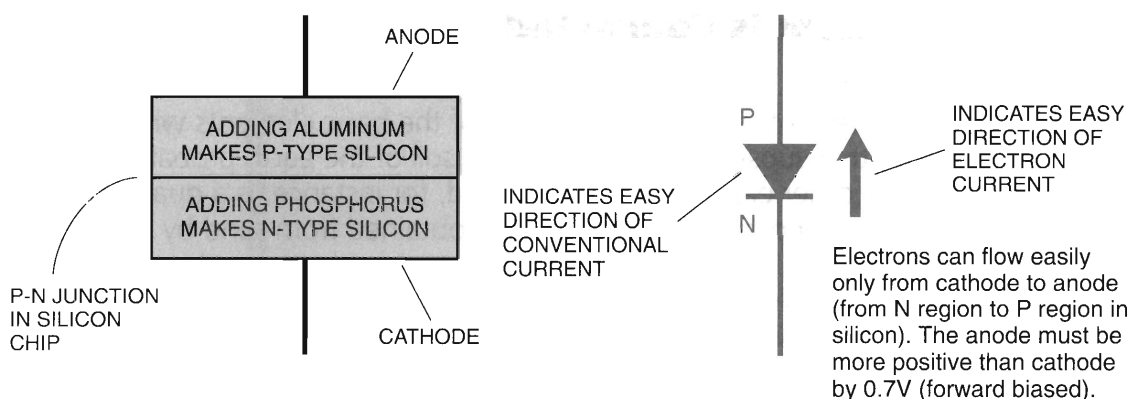


Figure 4-1. The P-N junction in a diode chip acts as a one-way valve for electrons.

The diode acts as a one-way valve for current. The diode allows no flow of electrons (turns off) when it is reverse biased, but it allows an easy flow of electrons (turns on) when it is forward biased. Forward bias is when the anode is more positive than the cathode and above the threshold voltage of about 0.7 volt. Reverse bias means the cathode is more positive than the anode, trying to cause current in the reverse direction. Thus, *a diode is basically a form of automatic switch*—when forward biased, current is allowed; when reverse biased, no current is allowed.

The switching occurs in response to an electrical signal (the voltage bias across the diode), and it can happen very rapidly. The process of allowing current in one direction and allowing no current in the other direction is called *rectification*. Some diodes can rectify ac at frequencies measured in gigahertz, which means billions of cycles per second.

The Junction Between P-Type and N-Type Silicon Rectifies Current

Let's talk more about the piece of silicon in *Figure 4-1* with its two regions called the anode and the cathode that form the diode. *Electrons* pass easily from the cathode to the anode. (Electron current, which is indicated by the separate arrow, is opposite from the conventional current direction, which is indicated by the arrowhead in the schematic symbol). Electron current from anode to cathode is blocked.

In studying semiconductor devices, it is easier to understand the electronic operation when we think about the flow of electrons instead of conventional current. Something about the anode and cathode regions of the silicon chip allows electrons to flow from cathode to anode, but not the other way.

The Different P and N Regions

What is different about the P anode and N cathode regions of silicon? To begin with, the basic material from which the silicon chip is made is a single crystal silicon. That means all the atoms inside it line up in the same rows and layers all through the chip without any interruptions. However, some modifications have been made to the single crystal silicon crystal for each region. The anode region has a few aluminum atoms mixed in with the silicon. As a result, for reasons we will see later, the anode material is called *P-type silicon*. In like fashion, the cathode region of the crystal has a few phosphorus atoms scattered here and there. This type of material is called *N-type silicon*.

The place where the two types of silicon meet inside the crystal is called the *P-N junction*. What we will find out in this chapter is how a P-N junction acts as a one-way valve for electrons. This will help us understand how transistors work.

Each Silicon Atom Is Connected to Four Others by Covalent Bonds

Silicon is a chemical element; that is, silicon is one of the basic elements which are combined to make other substances. The rocks and soil of the earth probably contain more silicon than any other element. A grain of sand, for instance, is a quartz crystal which is made of silicon and oxygen. Pure silicon is obtained from sand by separating the silicon from the oxygen. Silicon is used to make semiconductor devices because of the special ways in which electrons flow among the atoms of a silicon crystal. These ways depend on how the atoms are connected together.

If we could look inside a piece of silicon crystal with a microscope with super magnification, we would see silicon atoms arranged in very even rows and layers. As shown in *Figure 4-2a*, each atom would look like a fuzzy, cloudy ball, with four fuzzy extensions that connect it to four other atoms. The ball part of an atom is called the

atom's *core*. The extensions stretching between atoms are called *covalent bonds*. This particular arrangement of atoms in a crystal is called the *lattice*. This same model could represent any semiconductor crystal; for example, germanium, silicon or carbon. Yes, carbon, in the form of diamond, can be used as a semiconductor material that operates at extremely high temperatures.

Figure 4-2b shows the usual way of drawing a diagram of silicon atoms and covalent bonds in a crystal. The circles represent atoms or atomic cores, and the lines represent covalent bonds. The atoms are placed in a square pattern, with each one connected to four neighbors. In explaining how electrons flow through the crystal, we can pretend that the crystal is a single flat layer of silicon atoms arranged in squares instead of a three-dimensional model with the bonds in the X, Y and Z directions.

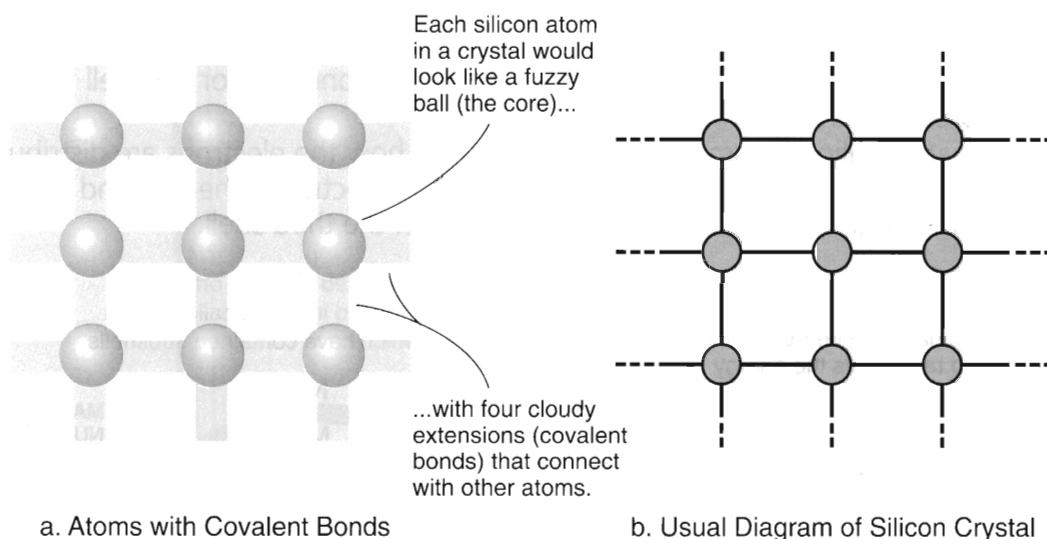


Figure 4-2. Each atom of a silicon crystal is connected to four other atoms by covalent bonds in an orderly arrangement called the diamond lattice. For simplicity, we can use a simple, flat diagram with atoms in a square pattern.

An Atom's Electrons Are Arranged In Shells

To understand how electrons flow in a semiconductor crystal, we have to see how covalent bonds work. Those bonds are a result of the way the electrons of each atom are arranged.

As stated in Chapter 1, an atom consists of a tiny, positively charged nucleus surrounded by a swarm of negatively charged electrons. The speeding electrons are held in orbits around the nucleus by electrostatic attraction. The nucleus receives its positive charge from positively charged protons. Each chemical element (hydrogen, oxygen, and silicon, for example) has a different number of positive protons in its nucleus. In a normal atom, there are just as many negatively-charged electrons as there are positively-charged protons. So the entire atom is neutral in charge.

Now, as shown in Chapter 1, an atom's electrons do not orbit just anywhere at random around the nucleus. Instead, the orbits in all atoms follow a certain plan. That plan is determined by a set of rules from physics called quantum mechanics.

Electron Orbits

Let's review the general plan for electron orbits in all atoms started in Chapter 1. We repeat the customary model shown in *Figure 4-3*. The actual orbits of electrons may be oval-shaped, and they do not all lie in the same plane. But in this model, *each electron's orbit is a circle* that represents the electron's average distance from the nucleus.

In *Figures 4-3a* and *4-3b*, the radius or height of a circular orbit also represents the *energy* that the electron has in its actual orbit. *The greater the height of the orbit above the nucleus, the more energy the electron has.* This is because it takes energy to raise an electron to a higher orbit against the electrostatic pull of the nucleus.

The plan for all atoms is that *the electrons are permitted to have only certain amounts of energy.* In this model, that means orbits with certain heights. *The permitted orbits are grouped in separate layers called shells.* In *Figure 4-3b*, the shells are shown as thick, shaded circles. The shells (K, L, M, N, O) and their contained subshells were shown in *Figure 1-2b*.

There are empty gaps between the shells. That means it takes a certain amount of energy to raise an electron from a permitted orbit in one shell or subshell to a permitted orbit in the next higher shell or subshell.

Look at the silicon atom in *Figure 4-4b*. Notice how the electrons are distributed in the shells. The silicon atom has 14 protons and 14 electrons. The first and second shells are full, and there are four valence electrons in the third shell.

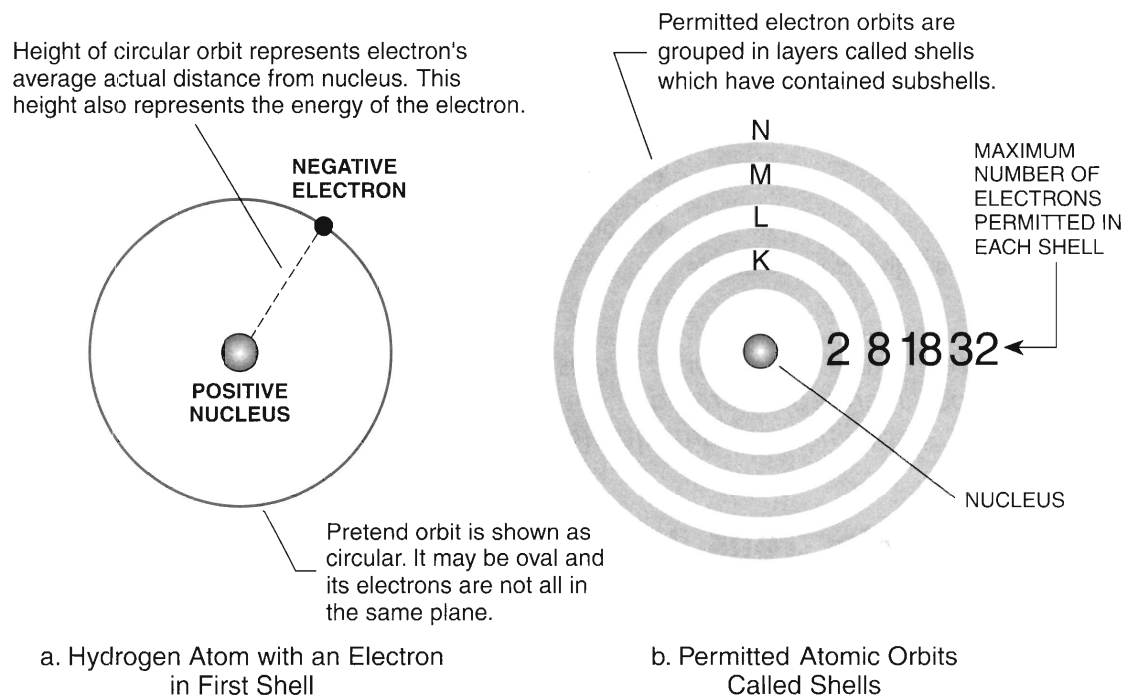


Figure 4-3. The electrons of an atom are permitted to have only certain energies (orbit heights), grouped in layers called shells. All atoms follow the same plan of shells and subshells dictated by physical laws.

Atoms Prefer to Have Full Valence Shells

What does all this have to do with covalent bonds and the flow of electrons in a semiconductor crystal? Well, *an atom uses the valence electrons of its outermost shell or subshell to form bonds with other atoms.* These bonds determine whether the material conducts current, and if so, how. The bonds also determine the kind of chemical reactions in which the element engages.

Example 1. Determining Distribution of Orbiting Electrons

If an oxygen atom has an atomic number of 8 and an arsenic atom has an atomic number of 33, how are the orbiting electrons distributed around the nucleus?

The atomic number identifies the number of positive protons in the nucleus, and, as a result, the number of orbiting electrons permitted in the following shells:

| Shell | | K | L | M | N | O | |
|--------------------------------------|---|---|---|----|----|----|--|
| Subshells | s | 2 | 2 | 2 | 2 | 2 | |
| | p | | 6 | 6 | 6 | 6 | |
| | d | | | 10 | 10 | 10 | |
| | f | | | | 14 | 14 | |
| Maximum Orbiting Electrons Permitted | | 2 | 8 | 18 | 32 | 32 | |
| Oxygen has 8 Orbiting Electrons | | 2 | 6 | | | | Oxygen has 2 electrons in first (K) shell and 6 in second (L) shell. |
| Arsenic has 33 Orbiting Electrons | | 2 | 8 | 18 | 5 | | Arsenic has 2 electrons in first (K) shell, 8 in second (L) shell, 18 in third (M) shell, and 5 in fourth (N) shell. |

The reason that the outer electrons form bonds is that *atoms prefer to have the outer subshell or shell full of electrons*. The other shells down below, if there are any, are already full. An atom forms bonds only if the outer shell is not filled with its maximum permitted number of electrons. This has to do with the fact that *when a shell or subshell is filled, its electrons are held especially tightly* by the nucleus. That is why elements like helium (atomic number 2), neon (atomic number 10) and argon (atomic number 18, shown in *Figure 4-4c*) are very stable elements. Their outer shells or subshells are full and they don't particularly want to react with other atoms.

Valence Electrons and the Atom's Core

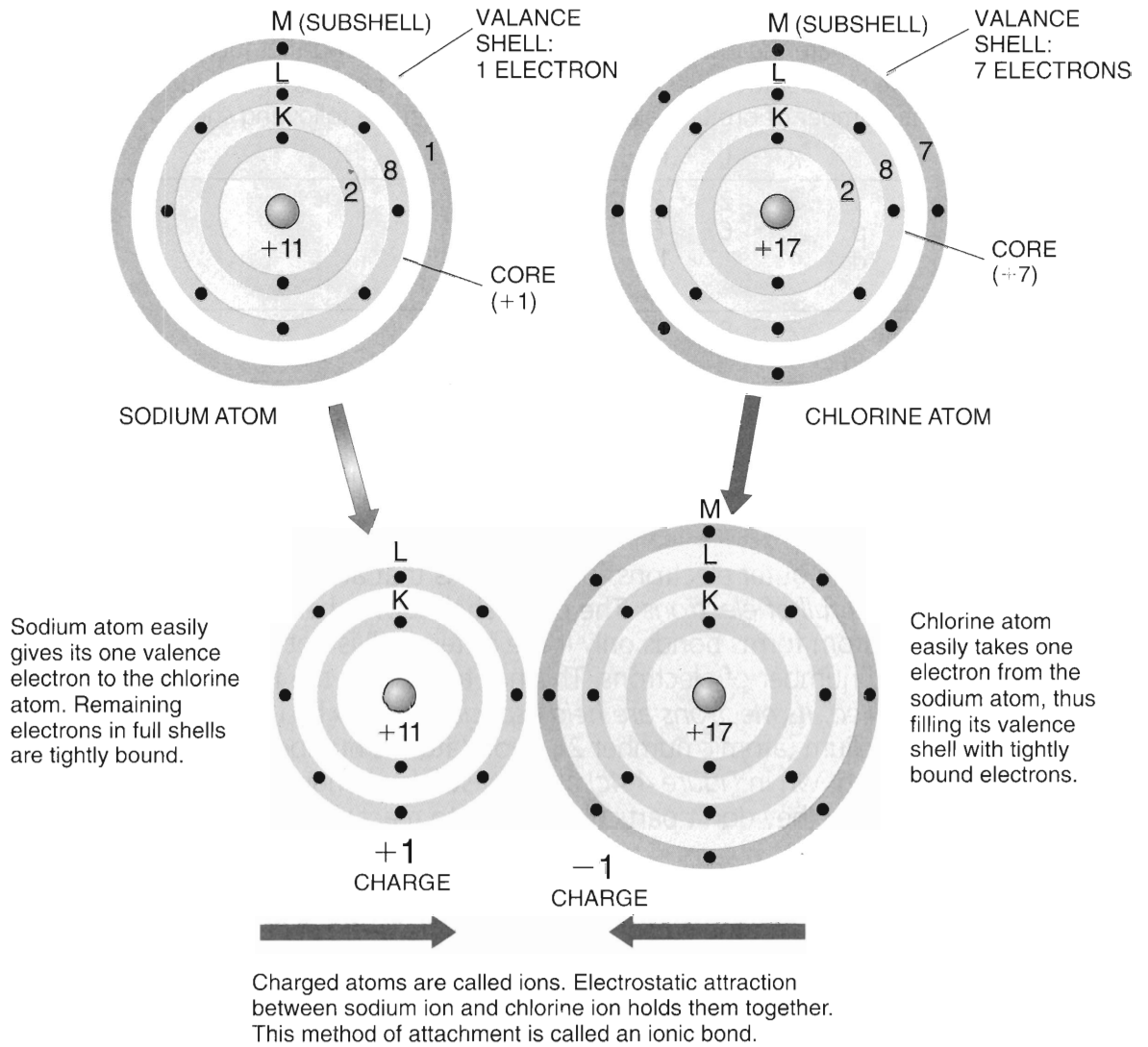
Look at the sodium atom and the chlorine atom in *Figure 4-4*. There is a special name for an atom's outer subshell or shell if it is not full. It is called the atom's *valence shell*. "Valence" means the number of bonds the atom forms. For instance, the valence of silicon atoms in a crystal is four, because every atom forms four bonds. As mentioned previously, the electrons in the valence shell are called the atom's *valence electrons*.

The rest of the atom, consisting of filled shells and the nucleus, is what is called the *core*. Remember, shells filled with electrons don't have anything to do with bonds, chemical reactions, or current, so they can be considered as separate from the valence electrons. The core has a positive charge equal to the number of electrons in the valence shell.

For example, look again at the sodium atom in *Figure 4-4*. Its outermost shell contains only one electron. Since this shell is not filled to its capacity, it is the atom's valence shell. The nucleus and the filled first and second shells are the core of the sodium atom. The core has a positive charge of plus one, which is balanced by the negative charge of the single valence electron. The chlorine atom, on the other hand, has a core with a positive charge of plus seven because it has seven electrons in its valence shell.

When nearly all the orbits of an atom's valence shell are empty, the atom easily gives up the few electrons in that shell. So the single electron in a sodium atom's valence shell is not bound (tied) very tightly to the core.

The chlorine atom's valence shell has seven of the eight electrons that it desires, so a chlorine atom can easily grab and hold one extra electron in its valence shell. If a chlorine atom bumps against a sodium atom, *the chlorine atom steals the sodium atom's single valence electron*. In this way, the chlorine atom achieves a full outer shell, and the sodium atom ends up with an outer shell completely void of electrons.



a. Ionic Bond

Argon atom (inert gas) has no valence shell because its outer subshell is full. No bonds are possible.

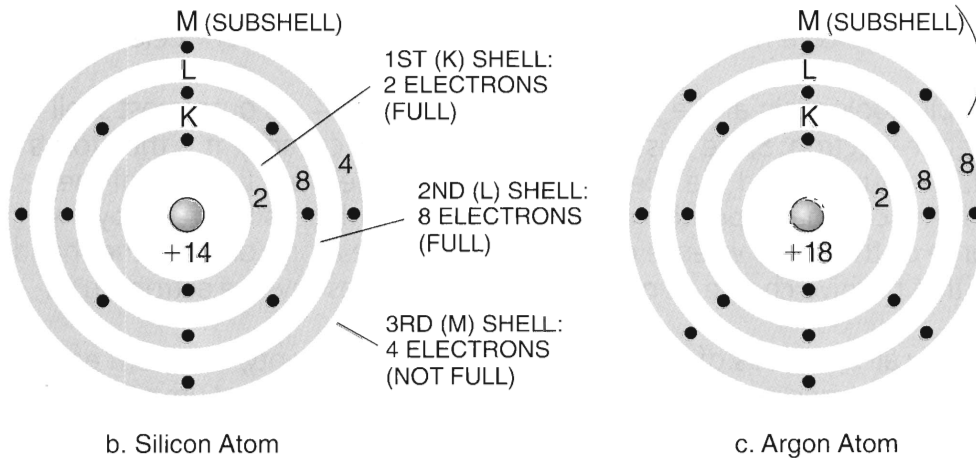


Figure 4-4. Atoms form bonds because they prefer to have full outer subshells or shells. An unfilled outer subshell or shell is called the valence shell.