Operation of Semiconductor Devices
a Simplified Explanation

by Kenneth A. Kuhn
Sept. 5, 2007, revised Sept. 2, 2012

Note: This document is best viewed or printed in color

Introduction

Explaining the operation of semiconductor devices is a challenge because the process is quite complicated. The approach taken here is to skip the complications and high math and develop an intuitive feel for operation. A good intuitive feel is the main thing needed to understand and design solid-state circuits.

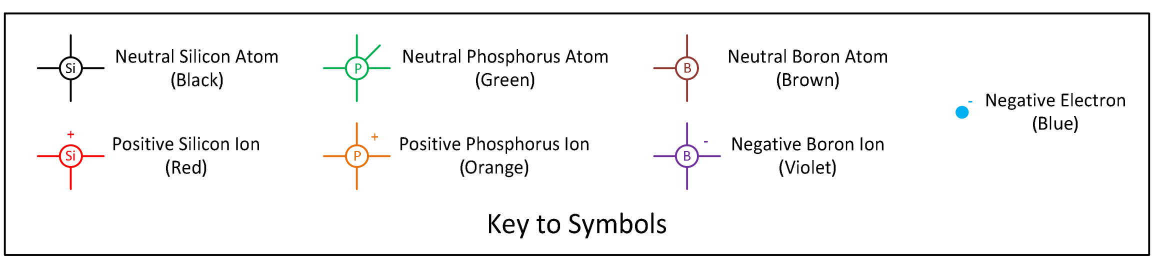
Conductors and Insulators

An electrical conductor is a material that has an abundance of free to roam charged particles (generally electrons). Typically, some electrons in the outer orbit of the material are very loosely held and little energy is required for them to escape orbit and wander freely. Normally, there is no net charge on the conductor as there is a balance between positive and negative charges – each electron that breaks orbit leaves a positive ion behind. During the course of wandering an electron may become close enough to a positive ion to re-enter orbit thus returning the ion to a neutral atom. In the absence of an applied voltage gradient the location of these charges randomly moves about the material at temperatures above absolute zero. These charges are free to move in an orderly fashion, thus forming a current, in response to an applied voltage gradient.

An electrical insulator is a material that has few if any charged particles free to roam. The electrons are very tightly bound to the atoms and it takes high energy to knock them loose. Thus, there is little if any current in response to an applied voltage gradient.

Semiconductors

The following chart is a key to the symbols in the figures.



Operation of Semiconductor Devices
a Simplified Explanation

A semiconductor is a special type of material that generally conducts poorly in its pure form but with the addition of specific impurities (a process called doping) the conductivity can be controlled by electric fields – that ability is what makes the material a semiconductor. Silicon is the most common semiconductor in use today. Previously, germanium was the material of choice.

Silicon atoms have four electrons in the outer orbit and these form covalent bonds with adjacent silicon atoms in a uniform crystalline structure. These electrons are fairly tightly held so the electrical resistance is high. However, a relatively small number of electrons do break free at any given time and are available for conduction. This is illustrated in Figure 1 which is a static picture. Imagine the locations of the positive silicon ions randomly changing and the locations of the free electrons are constantly moving. In all cases there is no net charge – the number of free electrons always exactly matches the number of positive silicon ions.

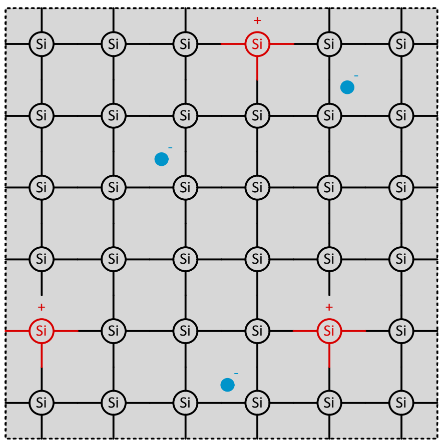


Figure 1: Pure Silicon

Note that only a few electrons are available at any given time for conduction

Type n semiconductor

When a material such as phosphorus which has five electrons in the outer orbit is added the fifth electron does not form any kind of bond and little energy is required for it to escape orbit. Thus, the phosphorus atom is referred to as a donor since it provides a free electron to the crystalline structure. This forms what is called a type n semiconductor because of the abundance of free electrons. In type n semiconductor material the free electrons are called majority carriers since they are the primary mechanism for conducting current. The positive ions (also known as holes) are referred to as minority carriers. The positive ions do not physically move but the location of positively charged entities does move thus constituting a current. This material is a much better conductor than pure silicon. This is illustrated in Figure 2 which is a static picture – imagine a random motion of the location of positive ions and free electrons.

Operation of Semiconductor Devices
a Simplified Explanation

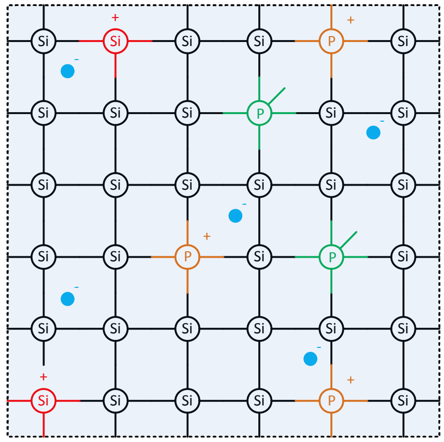


Figure 2: type n semiconductor formed by doping with donor atoms
Observe that the phosphorus atoms readily give up an electron.
The doping level is highly exaggerated.

Type p semiconductor

When a material such as boron which has only three electrons in the outer orbit is added to the silicon structure there is a dangling bond. Any nearby free electron tends to become captured and fairly tightly held by the boron ion. Thus, boron is called an acceptor since it accepts an electron. This leaves a negatively charged boron ion and a positively charged silicon ion which is known as a hole. Although the atoms do not move, the location of a positive silicon ion does move about randomly as different silicon atoms give up an electron or capture a nearby free electron. This forms what is called a type p semiconductor because of the abundance of free holes. In type p semiconductor material the free holes are called majority carriers since they are the primary mechanism for conducting current. The electrons are referred to as minority carriers. This material is also a much better conductor than pure silicon.

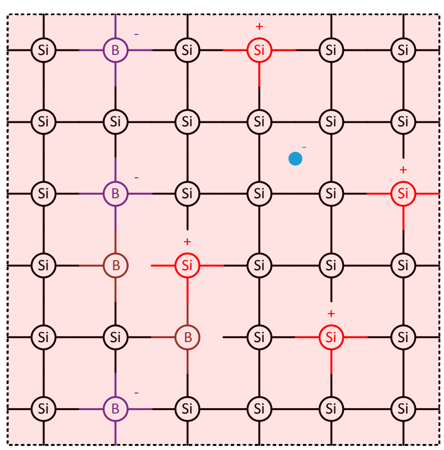


Figure 3:type p semiconductor formed by doping with acceptor atoms
Observe that the boron atoms acquire and hold on to free electrons.
The doping level is highly exaggerated.

Operation of Semiconductor Devices
a Simplified Explanation

A doped semiconductor is a fair conductor but in of itself is not particularly interesting. Both types conduct electrical current much better than pure silicon but that is the end of the story. Things get very interesting in the junction region when a type n and a type p are joined together. In actuality both semiconductor types must be built together rather than physically joined. But the discussion is much easier if we first think of them separately and then join them later.

The pn junction

When a type n semiconductor is joined with a type p semiconductor forming what is known as a pn junction, many free electrons in the type n semiconductor cross over the junction and become bound by the free holes in the type p semiconductor. This forms a net negative charge on the type p side of the junction which blocks further electrons from crossing. The holes on the type n side of the junction form a counter positive charge. Figure 4 illustrates this.

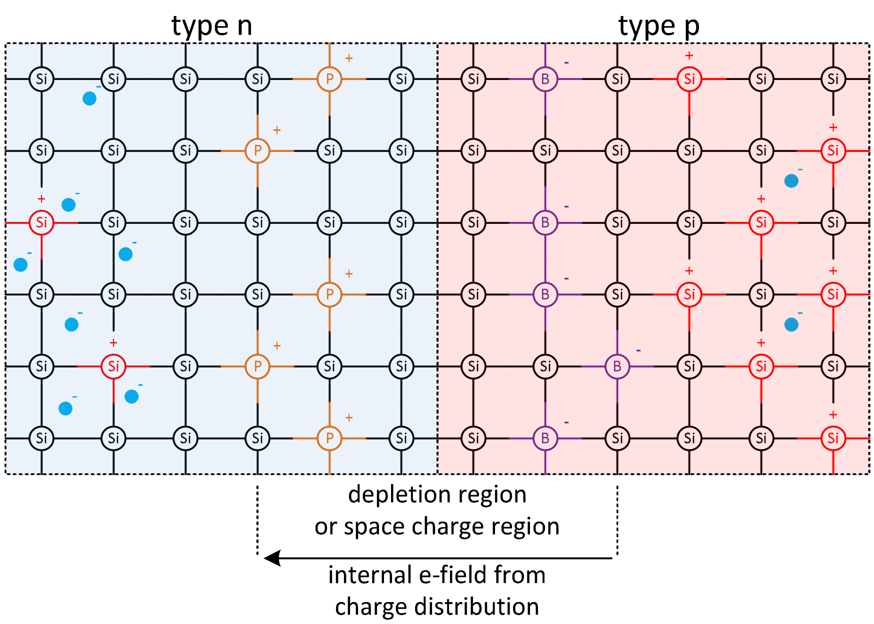


Figure 4: pn junction

The zone of the two charge regions on either side of the junction is known as the depletion region (also known as the space charge region) as there are few if any majority carriers available to conduct current. Thus the depletion region is an electrical insulator.

The pn semiconductor device is electrically neutral but has a very non-uniform distribution of charge as follows:

- On the type n side and away from the junction there is a net negative charge because there are few positive silicon ions available to capture free electrons.

- On the type n side near the junction there are an abundance of holes (i.e. positive charge) because the donor atoms have lost their electrons to the acceptor atoms on the type p side.
- On the type p side near the junction there are an abundance of acceptor atoms with bound electrons forming a net negative charge.
- On the type p side away from the junction there are an abundance of silicon atoms that have lost electrons thus forming a net positive charge.

There is no net voltage across the pn device because all of the internal voltage gradients sum to zero.

Forward bias condition

Forward bias by definition is when the applied voltage to a pn junction is such that the voltage at the type p material is more positive than the voltage at the type n material. With a small forward bias the positive ions on the type p side of the junction are repelled towards the junction and the free electrons on the type n side are also repelled towards the junction. This pressure forces the depletion region at the junction to become narrower thus increasing the probability that more energetic electrons on the type n side can cross the depletion region to the type p side. The depletion region is a barrier to conductivity and it takes a relatively high energy to penetrate this barrier. As the forward bias voltage is increased the depletion region becomes narrower still and the energy required for an electron to penetrate the region becomes smaller. Thus many more electrons cross the junction and form a higher current. Up until a point where ohmic losses in the material limits current the current increases exponentially with the applied forward voltage.

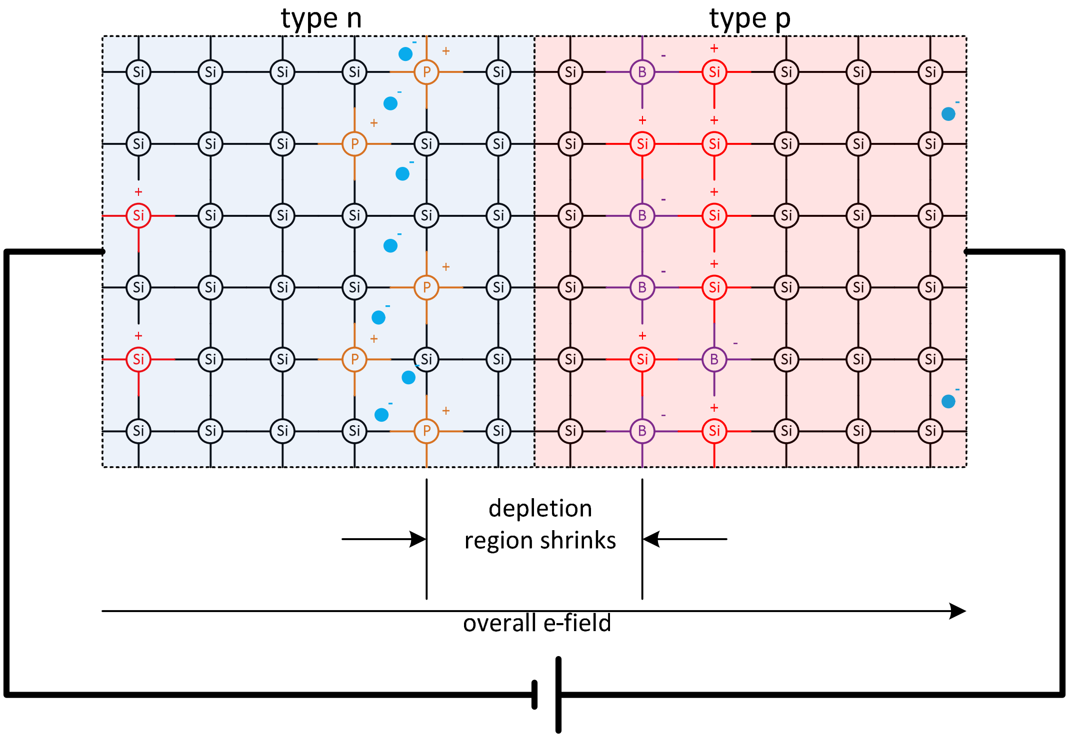


Figure 5: pn junction with forward bias

Reverse bias condition

Reverse bias is the opposite of forward bias – the voltage at the type p material is negative relative to the voltage at the type n material. The negative voltage applied to the type p side attracts the free holes away from the junction thus pulling the negative ions near the junction on the type p side. Correspondingly, the positive voltage applied to the type n side attracts the free electrons away from the junction thus pulling the positive ions near the junction on the type n side. This force increases the width of the depletion region and little if any current can exist. The higher the reverse voltage the wider the depletion region becomes. Thus, there can be no conduction when the pn device is reverse biased.

Operation of Semiconductor Devices
a Simplified Explanation

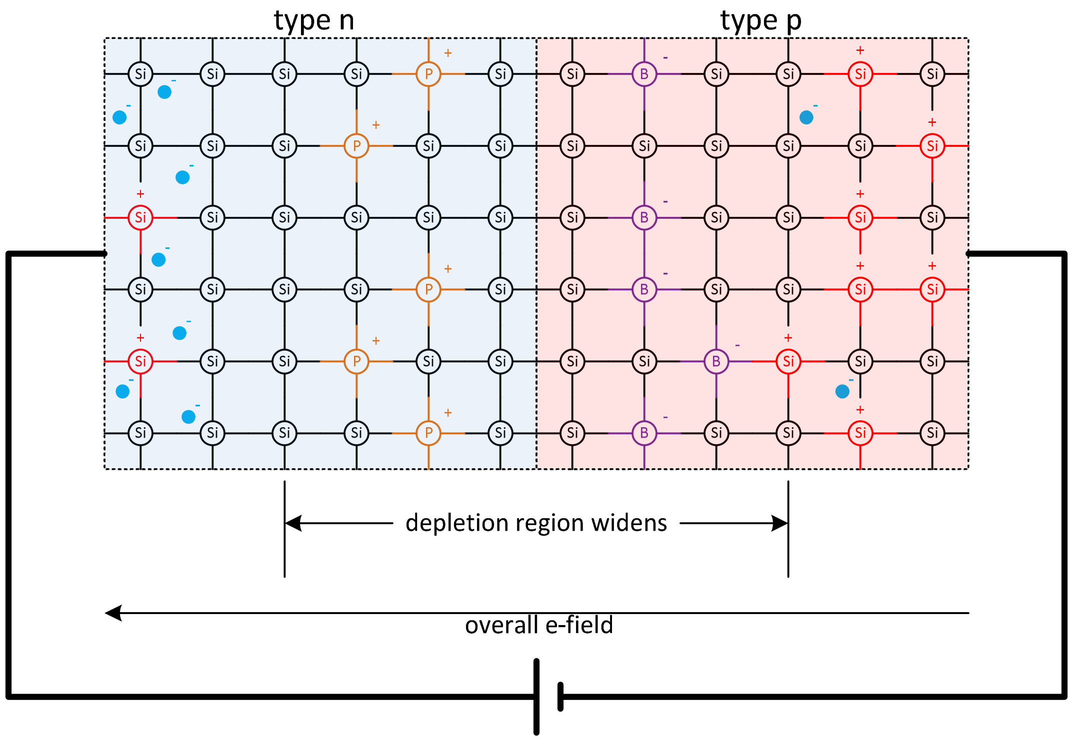


Figure 6: pn junction with reverse bias
Free electrons on the type n side are attracted to the positive supply.
Free holes on the type p side are attracted to the negative supply.

Diode operation

Thus, the diode can conduct current in only one direction. This unidirectional feature has a variety of useful applications ranging from rectifying an AC current to various signal processing functions to performing various digital logic functions. The two terminals of the diode have names – anode for the type p semiconductor and cathode for the type n semiconductor. These names come from the original vacuum tube diode. Under the forward bias condition the anode is more positive than the cathode and current can exist. Under the reverse bias condition the anode is more negative than the cathode and the device appears as an open circuit.

Bipolar junction transistor ***(note: this material and figures will be updated soon)***

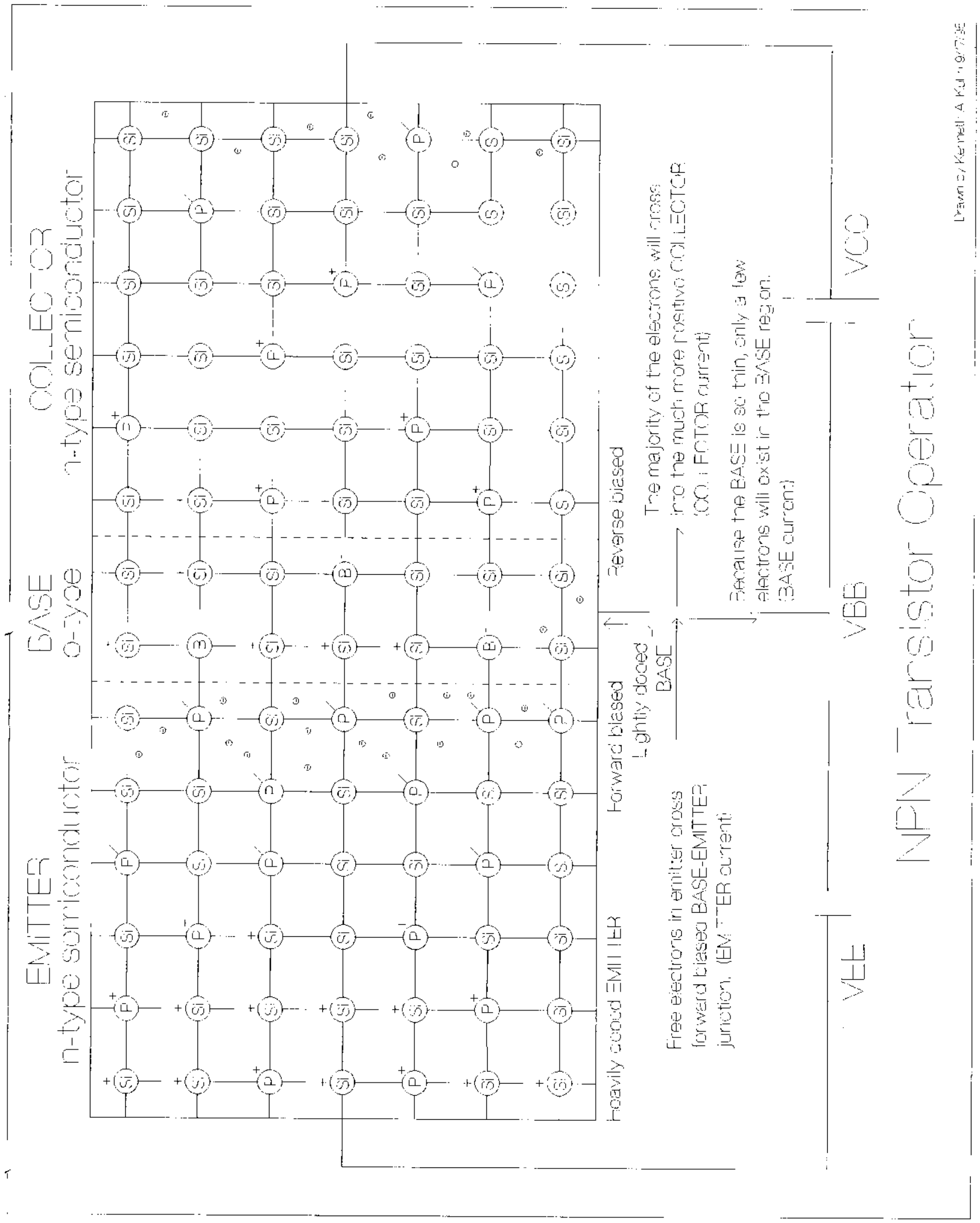
The bipolar junction transistor is a three-layer semiconductor device constructed either of two ways – NPN or PNP. This discussion will be about the NPN. Operation of the PNP is analogous by swapping donor and acceptor atoms.

The three terminals of the transistor are referred to as the emitter which is a heavily doped type N semiconductor and is a source of majority carriers (electrons for type N). The middle type P layer is extremely thin physically and is referred to as the base. The remaining type N material is called the collector because it literally collects majority carriers from the emitter.

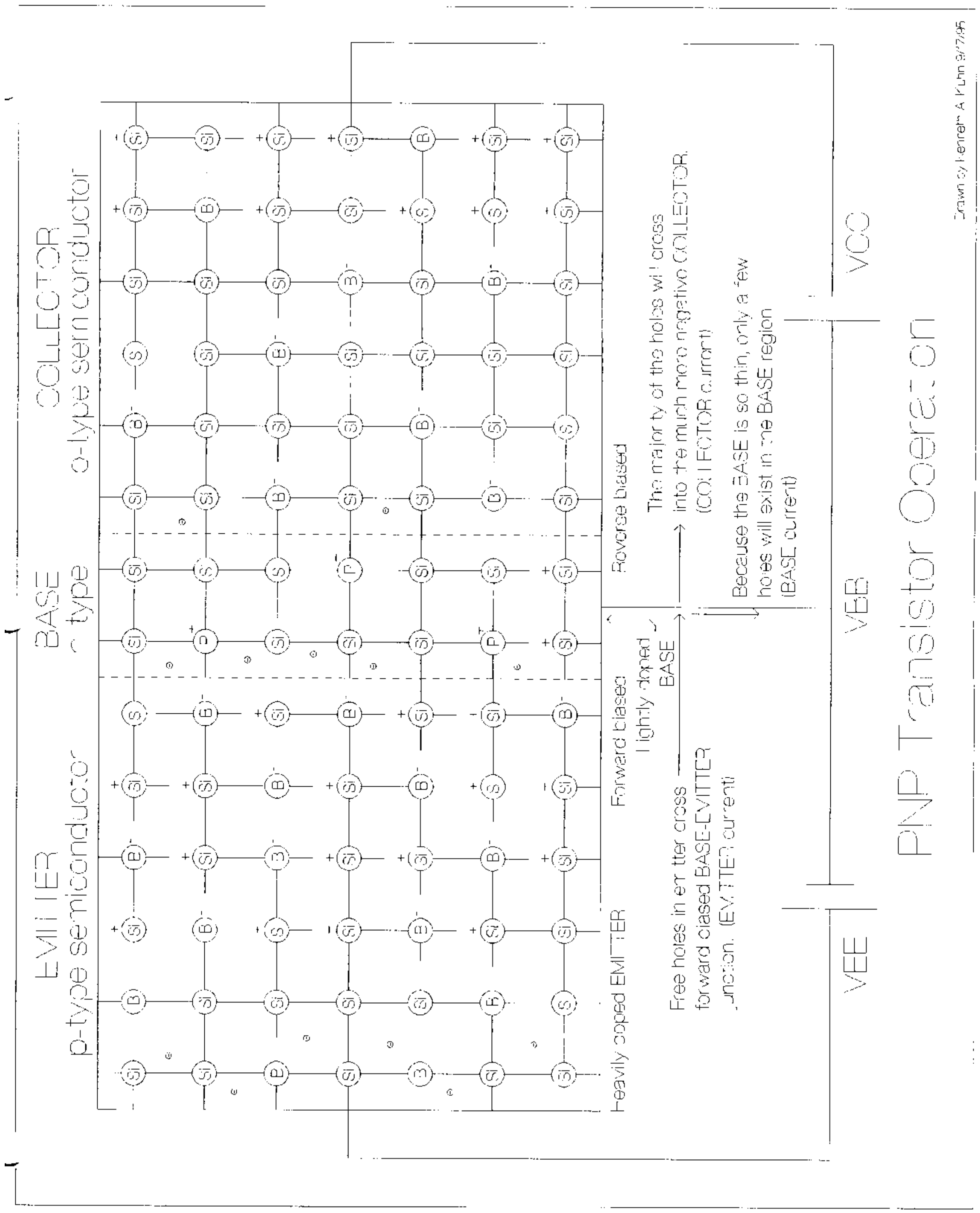
In operation the base-emitter junction is forward biased and the base-collector junction is reverse biased. This means that the base is slightly positive relative to the emitter and the collector is very positive relative to the base.

The forward bias condition between the base and emitter makes it possible for electrons from the type N emitter to enter the type P base region just as in the diode. Now things get very interesting. Because the base region is very thin, the depletion region formed by the reverse biased base-collector junction, electrons that enter the base region from the emitter are free to move towards the more positively biased collector. Thus, most of the electrons take that route and only a few exit the base terminal as a base current. The collector current is then much higher than the base current – typically a factor between around 50 and 300. Stately differently, a small base current controls a much larger collector current. This is the basis for gain or amplification. The action of the bipolar junction transistor is a current controlled current source.

Operation of Semiconductor Devices
a Simplified Explanation



Operation of Semiconductor Devices
a Simplified Explanation



Junction field-effect transistor

Junction field-effect transistors (JFET) control a current by using an electric field to modulate the width of a depletion region. The junction field-effect transistor is a two layer semiconductor device consisting of a channel of one semiconductor type surrounded by a control gate made of the opposite semiconductor type. The device is known by the type of semiconductor comprising the channel. Thus we have N-channel JFETs and P-channel JFETs. This discussion will be about the operation of the N type. Operation of the P type is analogous.

The N-channel JFET is comprised of a type N semiconductor channel surrounded by a type P control gate. Each end of the channel is connected to a terminal and the two terminals are called the source and drain. The connection to the type P material is called the gate.

In operation the drain is made to be more positive than the source and the gate is always at some reverse bias potential relative to the source terminal. This reverse bias causes a depletion region to exist in the channel. Remember that in the depletion region that there are no free charges available for conduction and so the region acts as an insulator.

Thus, there can only be a current from source to drain near the center of the channel away from the depletion region. The width of the depletion region is controlled by the degree of reverse bias between the gate and the source terminal.

Thus, when the reverse bias is at or near zero volts, most of the channel area is available for conduction and the current between the source and drain is relatively high. When the reverse bias is high the channel area is narrow and the current between the source and drain is relatively small.

At a sufficient reverse bias the depletion region occupies the entire channel area and there is no current between source and drain. This condition is called pinch-off.

The preceding discussion implies that the FET is a voltage controlled resistor. That is true only when the voltage across the source to drain terminals is small (generally no more than a fraction of a volt). This condition is known as the ohmic region of operation and is useful for making voltage controlled resistors.

At higher voltage drops the depletion region becomes wider near the drain terminal (because that part of the channel is more positive than the gate thereby increasing the reverse bias) thus limiting the current. The result is that the current reaches a saturation level for a given gate to source reverse bias. Normal operation of the junction FET as an amplifier is in this saturation region. The junction FET thus acts as a voltage controlled current source.