Sajit Chandra Shakya
Physics Educator
Swarnim Basti, Chagal, Kathmandu

Hello: 4275803
Email: sajitcshakya@gmail.com
Web Address: http://sajitsir.tripod.com

Semiconductor

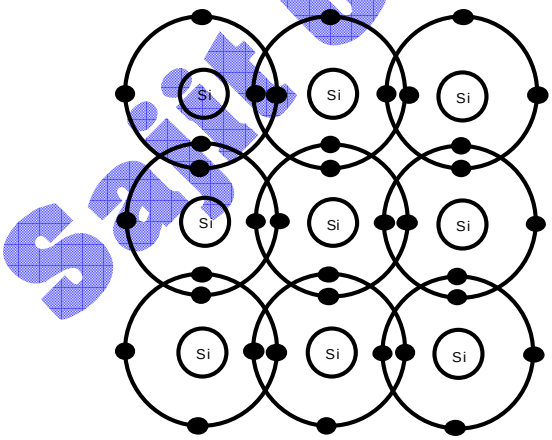
A semiconductor is a substance whose conductivity lies between that of conductors and insulators. They act as nonconductors at 0 K and neighboring temperatures but can be made conducting when some triggers act on them. The triggers may be some heat energy (even normal atmospheric temperatures are enough), light or other radiations and pressure. This is made possible by the absence of free electrons at 0 K but their energy levels are such that even small triggers as stated can create free electrons. These free electrons are then swept away by the electric field applied later.

A conductor is made of such atoms which possess free electrons even at 0 K, which is due to the presence of one or two electrons in the valence shell, which can easily be knocked off. An insulator is made of such atoms, which do not possess free electrons at 0 K, low temperature, normal temperature and even higher temperature. Very hardly that some of them give free electrons at very high temperatures, of the order of thousands of degrees, when they are close to melting.

So the substances which are eligible to be semiconductors are the Group IV elements. Their electrons are neither as loose as conductors nor as tight as insulators. For example Silicon and Germanium. Carbon does not qualify as a semiconductor because its valence electrons are very close to the nucleus (the atomic number of Carbon is 6, so it has only two orbits) and they are tightly held by the nucleus. Silicon is the most widely used because of its abundance on earth surface.

Pure Semiconductor (Intrinsic Semiconductor)

A pure semiconductor is made by melting a piece of element having four electrons in the outer shell, such as Silicon or Germanium and allowing to cool naturally. While cooling, each of the atoms shares one electron each with four other atoms, forming four covalent bonds. The solid piece formed thus is a lattice of silicon atoms attached to four others.

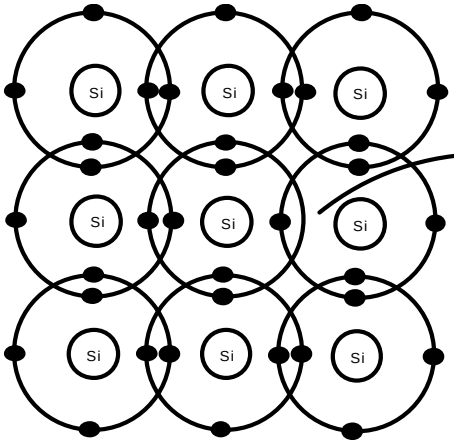


A pure semiconductor piece with no excitation of electrons

At 0 K and lower temperatures, each of the electrons of the covalent bonds is in its place. So no free electrons are present. However when the temperature is raised slightly, or some other triggers are given, the electrons of the bonds start vibrating and start to escape. The escaped electron is free from the attraction posed by the nuclei of atoms, so is termed as **free electron**. When the electron leaves its original position, it leaves an empty space. The space is called "**Hole**". Besides, the remaining structure, which has lost its electron develops a tendency to

Hello: 4275803
Email: sajitcshakya@gmail.com
Web Address: http://sajitsir.tripod.com

attract the just escaped or some other electron, so a **hole** is considered to be **positively charged** (remember that a hole is not matter, it is the absence of matter).



Free electron

Release of one free electron produces an empty space (or hole).

A pure semiconductor piece with excitation of an electron

The number of free electrons in a pure semiconductor can be increased by raising its temperature. However before any significant temperature is reached, the semiconductor piece starts to melt. So the number of free electrons has to be increased by some alternate method. The most widely used method is – **Doping**.

Doping is the process in which calculated amount of impurities are added to pure semiconductors so as to increase the number of free electrons or holes or at least make the electrons come out of the shells so that they can be made to be swept off by the electric field, if applied. Two types of impurities are generally used for the purpose.

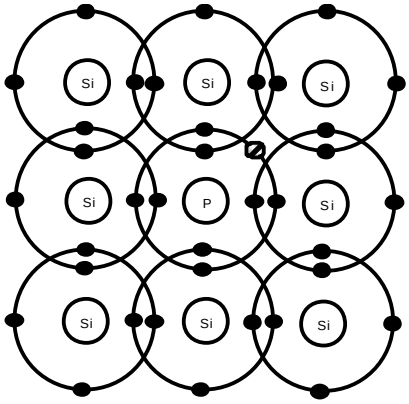
One of them is adding pentavalent impurities and the other is trivalent ones, the addition of each of them giving different names to the impure semiconductors, the former being called n-type semiconductors and the latter p-type. Such impure semiconductors are then called as **Extrinsic Semiconductors**.

N-type semiconductor

The semiconductors which possess more free electrons as a result of doping by pentavalent impurities is called as **N-type semiconductors**. This is constructed by first mixing pentavalent substance such as phosphorus or arsenic into pure silicon or germanium in the ratio of one is to million (roughly), then fusing and then allowing to cool slowly. When the mixture solidifies, the pentavalent atom gets entangled in between silicon atoms. Among the five valence electrons of the pentavalent atom, four get involved with one electron each of the four neighboring silicon atoms, where one remains idle. This electron is not attracted by anyone, in fact it is repelled by all. So this electron goes hither and thither, being "hated" by all and unattached to any particular atom. So this electron is called the **free electron** and there is a continuous and random flow of these electrons. So naturally there is the creation of free electrons, at the rate of one electron per impurity atom which help increase the conductivity of

Hello: 4275803
Email: sajitcshakya@gmail.com
Web Address: http://sajitsir.tripod.com

the material.



Extra unwanted electron and hence free

One free (extra) electron does not mean one empty space

A n-type semiconductor piece in natural state

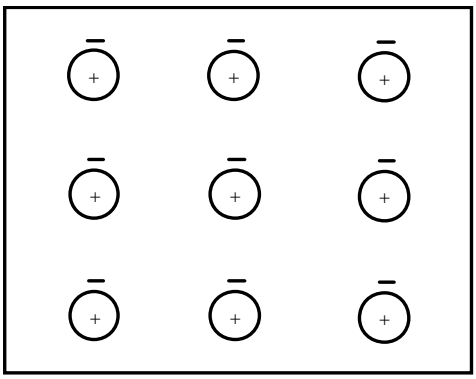
The material is called n-type because of this natural existence of free electrons, which are officially considered negatively charged. Since the pentavalent impurity gives such rich supply of negative particles, they are called as "**Donors**".

The free electrons are not only produced in this manner, some extra electrons might become free as a result of heat energy, but they are very minor compared to those produced due to bonding. The free electrons that existed due to impurity do not give any empty spaces, while those due to heat give few. So, the number of free electrons far exceeds the number of holes. So the free electrons are considered the "**Majority Charge Carriers**" and holes the "**Minority Charge Carriers**".

The n-type semiconductors are depicted as in the figure. For simplicity, the silicon atoms are not shown, only the donors are. Under normal conditions, each donor is represented by a plus (+) sign bounded by a circle with a minus (–) sign above it. The minus sign outside the circle denotes the free electron which has the potential to escape any moment (that's why it is not within the circle). After the free electron has escaped, what remains is the donor atom which has just lost an electron, so it is considered positively charged, and hence the plus sign. Unlike the free electron, it is not free, but stuck due to the silicon atoms around it. So it is shown bounded by the circle.

Sajit Chandra Shakya
Physics Educator
Swarnim Basti, Chagal, Kathmandu

Hello: 4275803
Email: sajitcshakya@gmail.com
Web Address: http://sajitsir.tripod.com



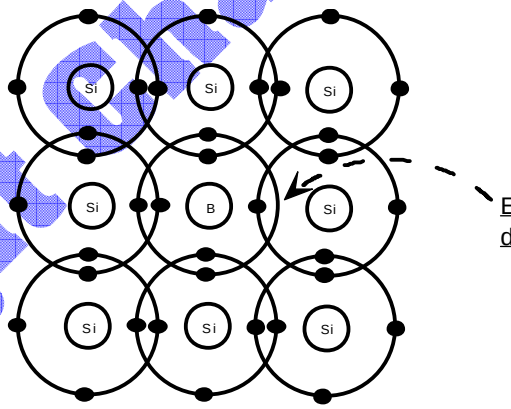
Minus sign (-) outside the circle denotes electrons which are free and can escape any moment.

Plus sign (+) inside the circle denotes the donor atom which hosted the free electron and become positively charged after giving the electron. The circle outside the '+' sign denotes that the atom is not free to move, it is bounded (stuck).

Symbol of n-type semiconductor

P-type semiconductor

The impure semiconductors in which empty spaces (holes) are created and electrons are made to venture out of shells by doping with trivalent impurities are called **P-type Semiconductors**. They are constructed by first mixing trivalent elements such as Aluminium or Boron with Silicon or Germanium in the ratio of one in a million (roughly), fusing and then allowing to cool naturally. When the mixture solidifies, the trivalent atom gets entangled in between silicon atoms. All the three valence electrons of the trivalent atom get involved with one electron each of the three neighboring silicon atoms. But one electron of the remaining silicon atom is still unaccounted for. This creates a space at the location which otherwise would have had a bond. This space is the hole. So when such doping is done, holes are created naturally, at the rate of one hole per impurity atom.

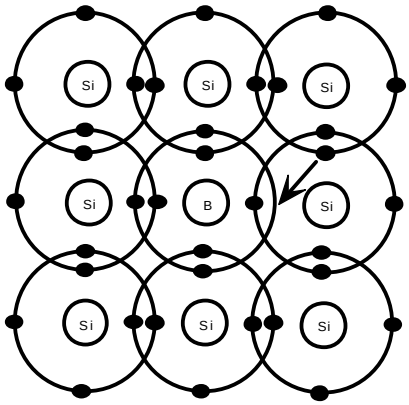


A p-type semiconductor piece just at the time of formation

But the impurity atom wants to reach its octet, so it snatches an electron from nearby bond. This transfers the hole in the opposite direction. The atom, in which vacant space is now created because of the loss of electron, again attracts an electron from nearby to reach the octet state giving the hole to that atom. So a continuous snatching of electrons and giving of holes occur throughout the piece, resulting in the flow of holes opposite to the flow of

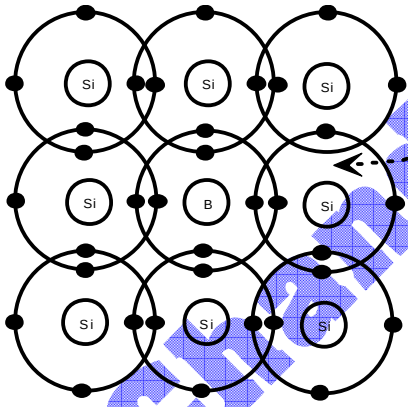
Hello: 4275803
Email: sajitcshakya@gmail.com
Web Address: http://sajitsir.tripod.com

electrons. So there are two types of flows, electron flow and hole flow, which are opposite to each other.



Boron atom attracting nearby electron to reach octet state

A p-type semiconductor piece after some time of formation



Position of the new space or hole

Electron filling up the existing space, which then creates a new space in its original location.

A p-type semiconductor piece after some time of formation

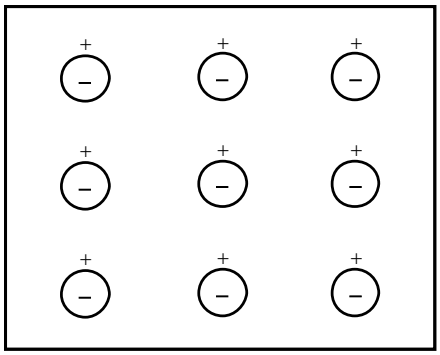
The material is called p-type because of this natural existence of empty spaces or holes, which are officially considered positively charged. Since the trivalent impurity gives such rich supply of positive entities, which are eager to accept something (electrons) from others, they are called as "**Acceptors**".

The holes are not only produced in this manner, some other holes might be created as a result of heat energy, which kick off electrons from the bonds, making them free and leaving behind vacant spaces. So, all in all the number of free electrons is very less than the number of holes. So the holes are considered the "**Majority Charge Carriers**" and free electrons the "**Minority Charge Carriers**" in these substances.

The p-type semiconductors are depicted as in the figure. For simplicity, the silicon atoms are not shown, only the acceptors are. Under normal conditions, each acceptor is represented by a minus (-) sign bounded by a circle with a plus (+) sign above it. The plus sign outside the circle denotes the hole which has the potential to disappear any moment (that's why it is not

Hello: 4275803
Email: sajitcshakya@gmail.com
Web Address: http://sajitsir.tripod.com

within the circle), due to the absorption of electron. After the hole is gone, an electron has been absorbed by the acceptor, so it is considered negatively charged, and hence the minus sign. Unlike the hole, it cannot escape or disappear, but stuck due to the silicon atoms around it. So it is shown bounded by the circle.



Plus sign (+) outside the circle denotes space (hole) which are free and can be lost any moment.

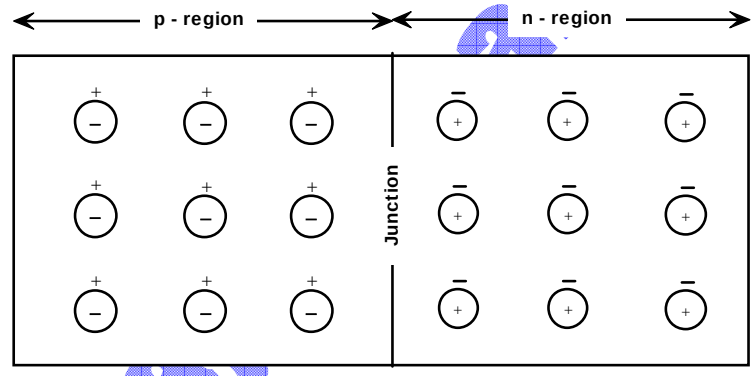
Minus sign (-) inside the circle denotes the acceptor atom which hosted the space and become negatively charged after receiving the electron. The circle outside the '-' sign denotes that the atom is not free to move, it is bounded (stuck).

Symbol of a p-type semiconductor

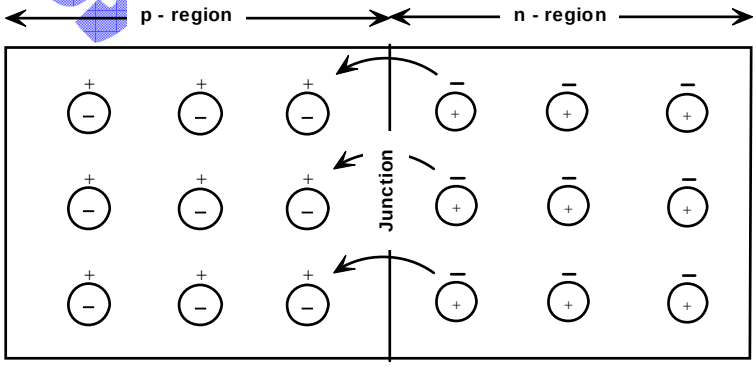
Que:

1. ***Which semiconductor would have higher conductivity: p-type or n-type?***
2. ***What is the overall charge in a n-type semiconductor?***
3. ***What is the overall charge is a p-type semiconductor?***

A semiconductor junction



Initial arrangement of atoms after the composite piece is formed.



Free electrons jumping into the holes in the nearby regions of the p - layer

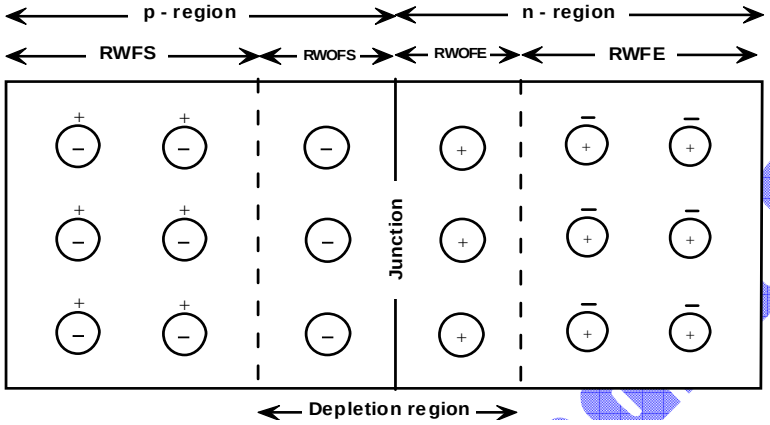
The doping process that was carried out on semiconductors need not be of only single type. Two doped regions can be created in the single piece; by diffusing the two types of substance in the liquid state and then allowing the whole thing to cool and solidify. The rate of cooling and solidification is balanced in such a way that when the pentavalent and trivalent atoms touch each other inside the semiconductor, the piece solidifies and there can be no further penetration as well as diffusion. This creates the composite piece whose middle has the boundary between negative region and the positive region. The so

Hello: 4275803
Email: sajitcshakya@gmail.com
Web Address: http://sajitsir.tripod.com

formed boundary is called as the "**Junction**".

The junction has two different types of environment at its two sides. At the right are the free electrons which simply want to escape or go away, while at the left, there are holes, which are eager to absorb any electrons coming nearby. So the first change observed is that some free electrons near the junction cross over the boundary to fill up the holes at the other side. So there form a cluster of positively charged atoms at the right of the boundary and negatively charged atoms at the left.

The n-region now has two different types of territories – one with free electrons and the other without. Similarly the p-region also has one territory with free holes and the other without. The regions which have lost their free electrons and holes touch each other and combinedly form a larger region which "does not" have anything free. So the region is called as "**Depletion Region**" (being depleted means emptied). It spreads up to few micrometers at both sides of the junction.



RWFS = Region with free spaces (holes).
RWOFS = Region without free spaces (holes).
RWFE = Region with free electrons.
RWOFE = Region without free electrons.

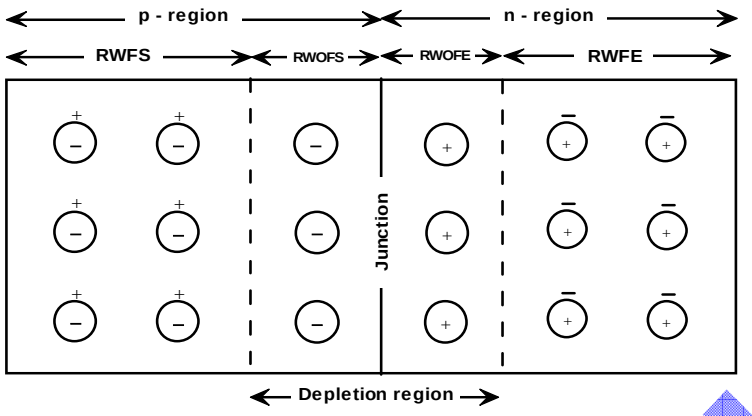
At the two sides of the depletion region, there is the grouping of positively charged atoms at one side and negatively charged at the other. So there exists a potential difference across the junction, named as "**Barrier Potential Difference**" or simply "**Barrier Potential**" in short. For silicon semiconductors, its value is 0.7 V and for germanium, it is 0.3.

The movement of the free electrons from the n-region to the p-region does not continue for long. The first reason is that as the electrons jump over, the width of the depletion region increases and the distance between the other free electrons and their target atoms at the other side increases. In addition, if the electron wants to jump, they are repelled by the lining of the newly formed negatively charged atoms right after the junction. So for a particular composite piece and at a particular temperature, the width of the depletion region is almost constant.

Forward Bias

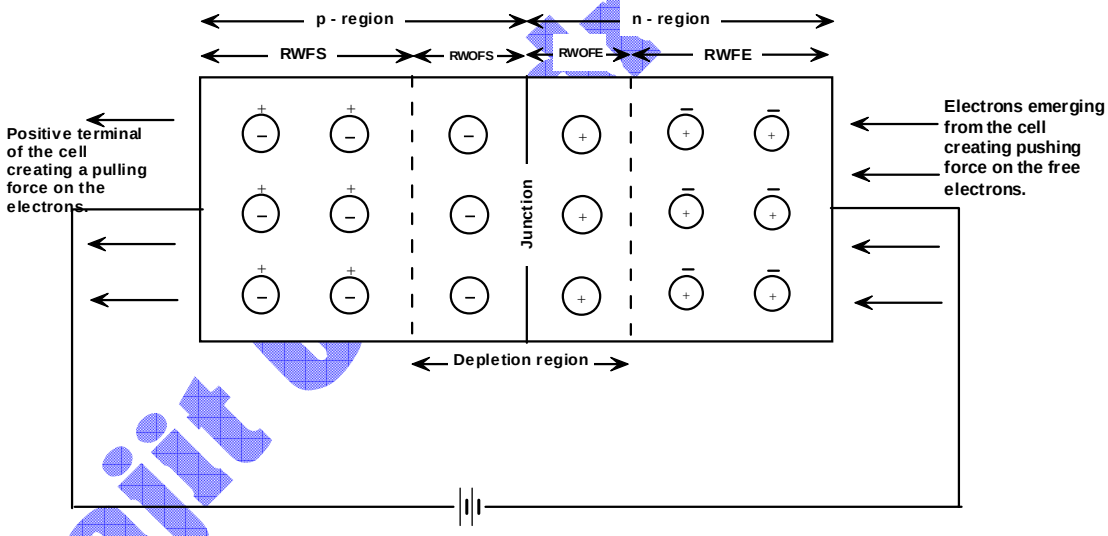
In a normal semiconductor composite piece, there exists a depletion layer at the two sides of the junction and free electrons from the n-side can not reach the holes of the other side due to the lining of newly formed negatively charged atoms at the start of the p-region.

Hello: 4275803
Email: sajitcshakya@gmail.com
Web Address: http://sajitsir.tripod.com



RWFS = Region with free spaces (holes).
RWOFS = Region without free spaces (holes).
RWFE = Region with free electrons.
RWOFE = Region without free electrons.

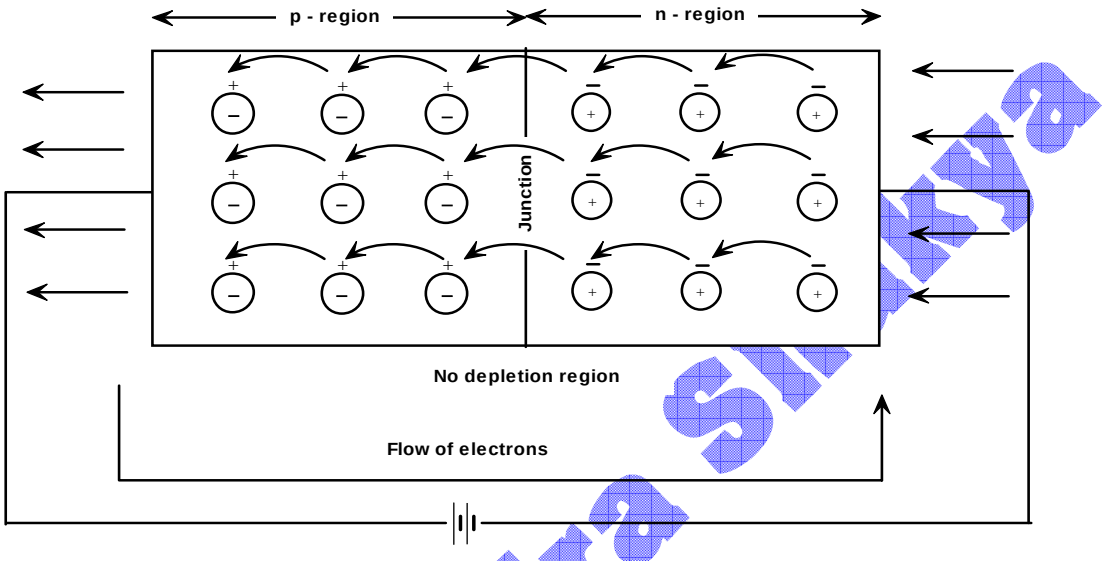
If the electrons are to be forced into the p-region, they have to be provided with greater energy from the n-side. The easiest method of doing so is connecting the piece to electric source, n-region connected to the negative terminal and p-region to the positive terminal of the source.



When such connection is done, the first effect is that the electrons arriving from the source start pushing the free electrons to the left (in the figure) and they are pushed instantly to the lining of the positively charged atoms, making them no more “depleted”. So they possess free electrons once again. Similarly, at the p-side, the positive terminal of the cell creates more lack and thirst for electrons, so the electron of the lining of the negatively charged atoms near the junction get knocked off and enter the cell. In this case electrons are not knocked off from the RWFS because they already lack electrons. So holes again emerge on the atoms which have just lost their electrons. It means both the regions which previously had no free electrons and holes, now have them and are no more “depleted”. So the effect is that the depletion region collapses or disappears.

Hello: 4275803
Email: sajitcshakya@gmail.com
Web Address: http://sajitsir.tripod.com

After that the free electrons again fill up the holes at the other side, which are immediately absorbed by the cell. So there goes a continuous chain of electrons emerging from the cell, going to the n-region in the form of free electrons, being absorbed into the holes and again into the cells. This results in a unidirectional flow of electrons (anticlockwise in the figure). So this arrangement of the semiconductor junction piece, in which electrons can flow forward, is called as "**Forward Bias**".

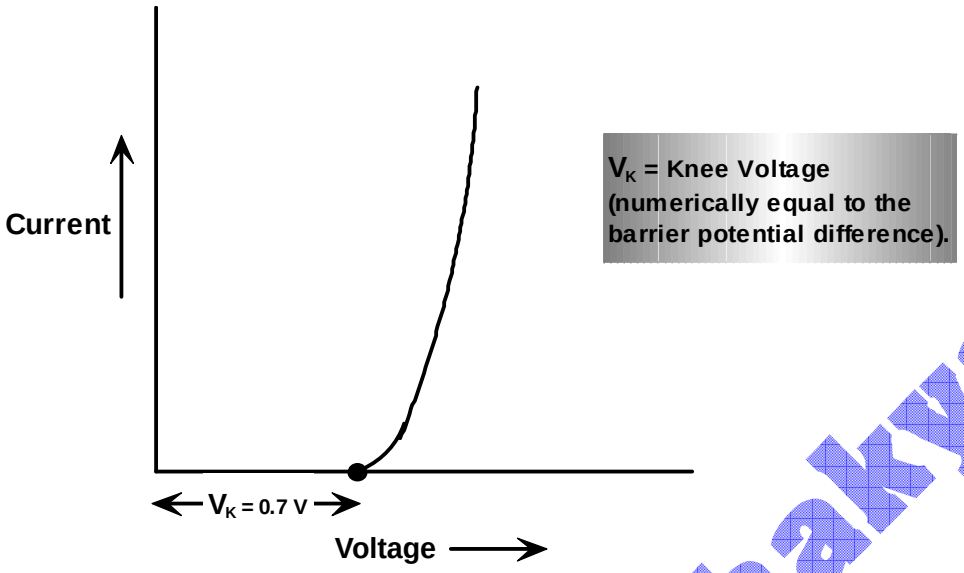


Electrons flipping from atom to atom through the junction

However, before the electrons can flow easily, the depletion region has to collapse. So the potential difference of the source that is used should be the one able to overcome the potential difference due to the depletion region. So it should be larger than 0.7 V for Silicon structures and 0.3 V for germanium structures. If the voltage is not enough, the depletion region reduces only slightly, but it still remains there. So electrons can not jump over and cause flow, that's why no current is observed for such low voltages.

The pattern of current flow for varying voltages would be like:

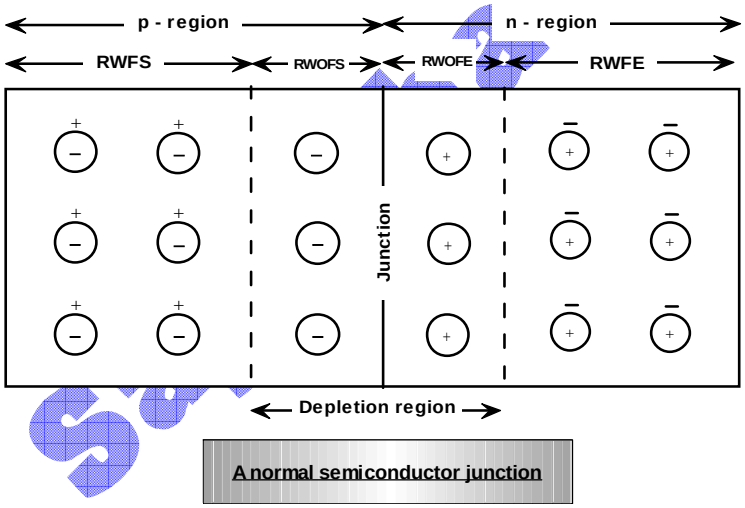
Hello: 4275803
Email: sajitcshakya@gmail.com
Web Address: http://sajitsir.tripod.com



The study of current variation according to the input voltage when it is forward biased is called as **Forward Characteristics**. The small resistance the piece offers to current when it starts to allow electrons (after **Knee Voltage**) is called **Forward Resistance**.

Reverse bias

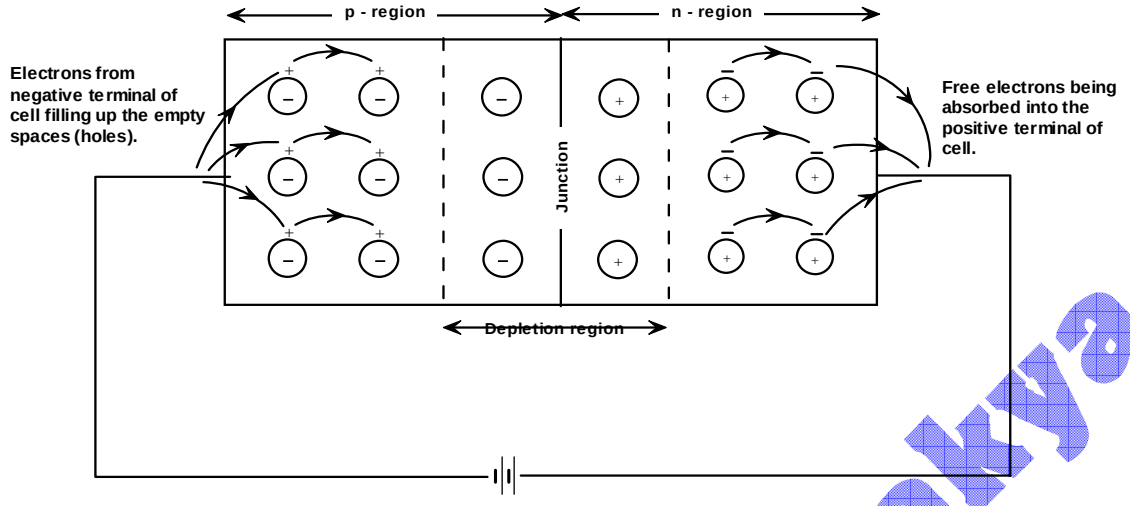
In a normal semiconductor junction, there is a depletion region of particular thickness in each p and n regions which gives rise to a barrier potential difference.



A normal semiconductor junction

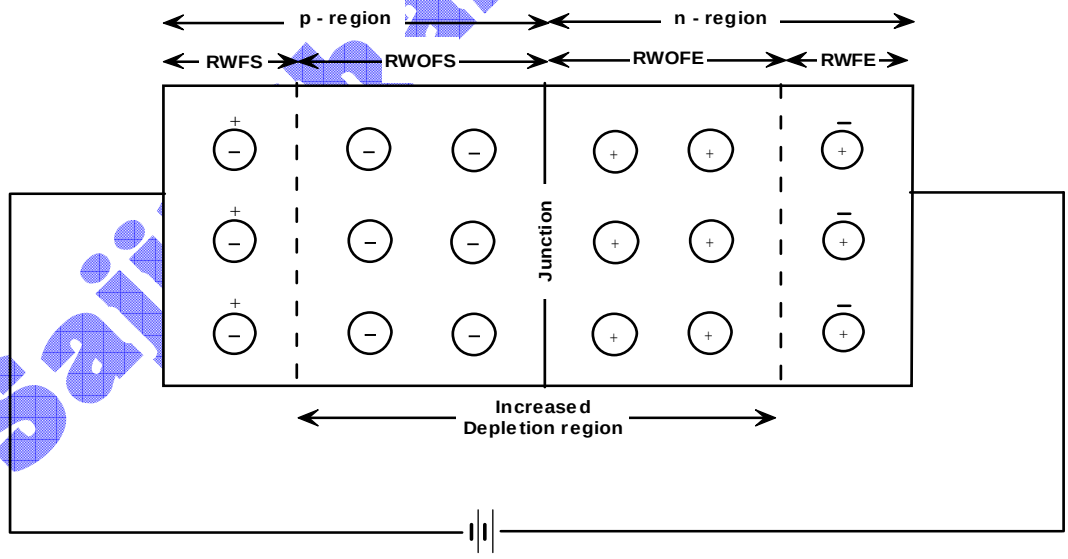
If the p side of the structure is connected to the negative terminal of a cell and n side to the positive, several changes occur in the structure. The first change is the absorption of the free electrons into the positive terminal of the cell, which makes more atoms of the n side depleted of free electrons, increasing the number of positively charged atoms at the right side of the junction. Similarly electrons from the negative terminal of the cell also starts to fill up the holes of the p side, making them also depleted of their holes. The electrons would rather fill up the holes of the inner atoms because they are slightly attracted by the increasing number of positively charged atoms at the other side. Therefore the number of hole-less atoms decrease. So the first effect is the increase in the span of the depletion region.

Hello: 4275803
Email: sajitcshakya@gmail.com
Web Address: http://sajitsir.tripod.com



A semiconductor junction soon after being reverse biased

In the depletion region there is a collection of negatively charged atoms at one side and positively charged atoms at the other. So if their span and amount goes on increasing, the barrier potential difference also goes on increasing. This process continues till the barrier potential difference becomes equal to the potential difference supplied by the cell. At that particular condition the whole circuit appears as if two cells with equal potential differences are at war with each other. So electrons can now flow neither way (get stuck), making conduction impossible and stopping the further increase of the depletion region. So this arrangement which does not allow the flow of electrons in a semiconductor junction by connecting and working in an opposite way compared to the forward bias is called "**Reverse Bias"**.

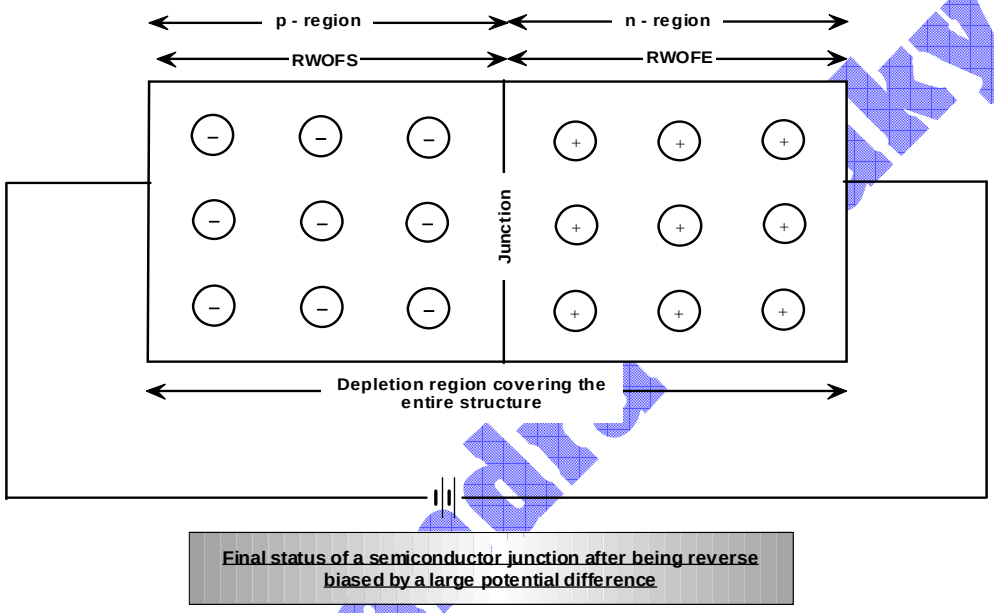


Final status of a semiconductor junction after being reverse biased by a small potential difference

However if the potential difference provided by the cell is very high, it will absorb all the free electrons and also fill up all the holes of the structure. This will fill the whole structure with the

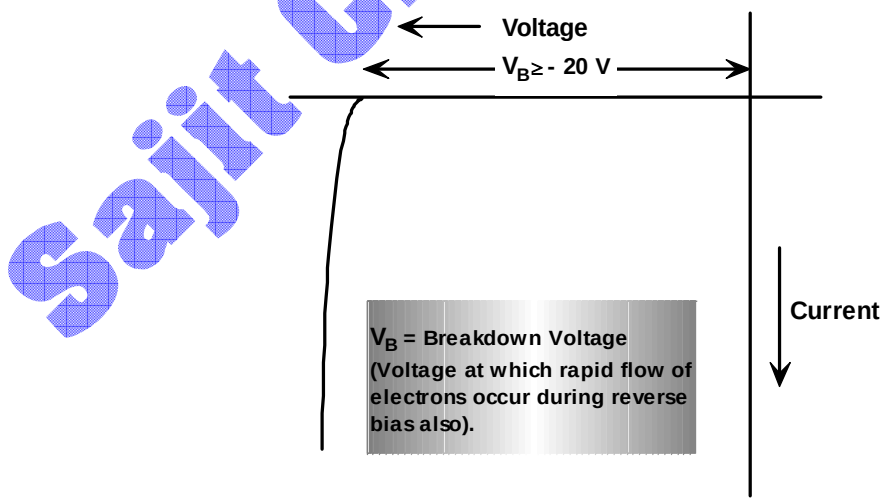
Hello: 4275803
Email: sajitcshakya@gmail.com
Web Address: http://sajitsir.tripod.com

depletion layer. Since no more free electrons or holes are available, that the force of the cell directs to the electrons in the valence orbits of the atoms of the structure. Because of this some of them get dislodged from their orbits and accelerate towards the positive side of the cell. On the way they knock out more electrons from other atoms and give a sudden rush of electrons within a very short time. This gives a quick rise of current. Such phenomenon is called as '**Avalanche Effect**' or '**Breakdown**'. The voltage at which this starts is called as "**Breakdown Voltage**". The current at that time is so high that there is a chance of excessive rise of temperature and burnout of the structure. Such conditions arise only at very large voltages and are generally avoided.



Final status of a semiconductor junction after being reverse biased by a large potential difference

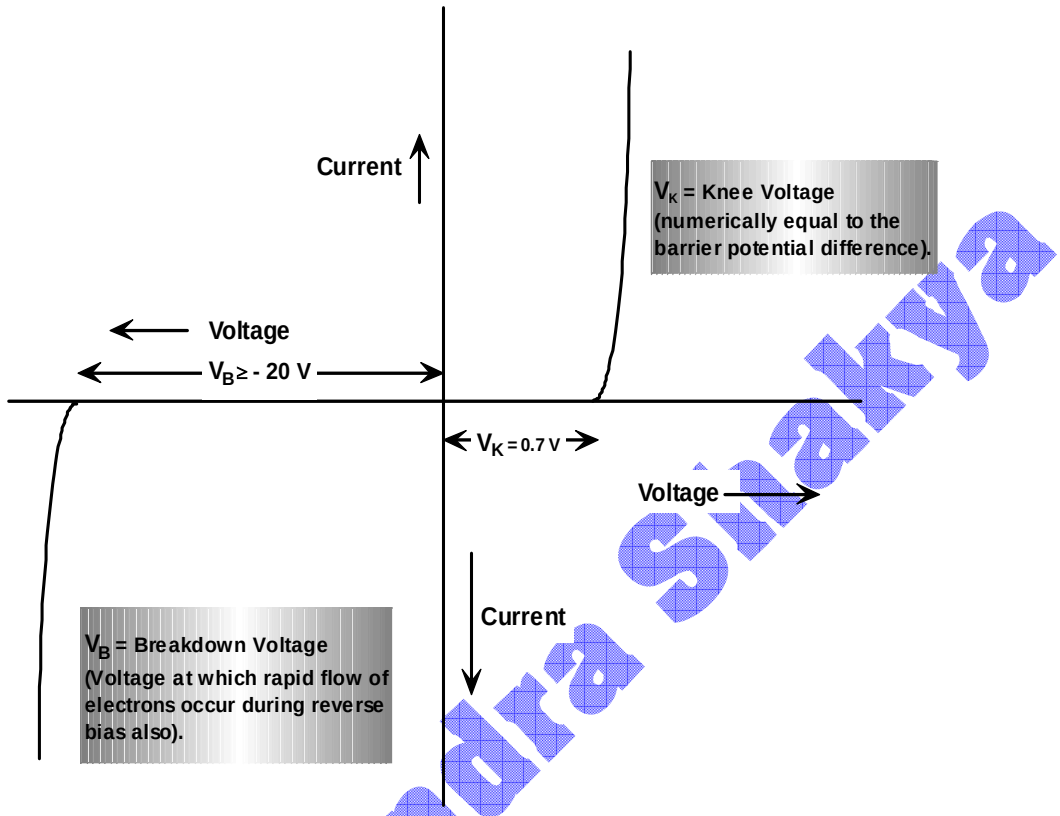
The variation of current with voltage in reverse biased condition would appear as:



It is to be noticed that after the breakdown voltage, the reverse current increases very rapidly. The study of variation of current with voltage at reverse bias is called as **Reverse Characteristics**.

Hello: 4275803
Email: sajitcshakya@gmail.com
Web Address: http://sajitsir.tripod.com

So the combined characteristics of a semiconductor junction would appear as:



Why is a semiconductor junction piece called a Semiconductor Valve?

As is evident from the two types of connections as forward bias and reverse bias, a semiconductor junction piece, made of n-type material at one side and p-type at the other, allows the flow of electrons from n-side to p-side, when connected respectively to the negative and positive terminal of a power source, but not in reverse direction. This action is much like of a valve. So the whole structure is called as **Semiconductor Valve**.

Why is a semiconductor junction piece called a Semiconductor Diode?

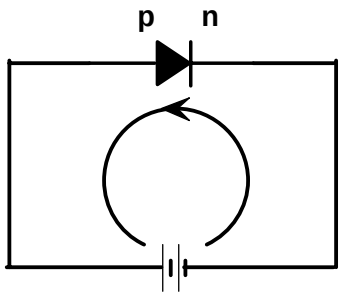
When a composite piece of semiconductor is constructed by using n-type and p-type material, one side of it has free electrons, eager to emerge and spread, whereas the other has holes, eager and thirsty for absorbing electrons. So, one side shows characters of a negative electrode (Cathode) and the other shows the property of positive one (Anode). So the composite piece contains both electrodes in a single structure. So it is called as Di-electrode, or Di-ode or **Diode**. Other nicknames for it are **Semiconductor Diode, or Junction Diode, or Semiconductor Junction Diode**.

Notation of a Semiconductor Diode

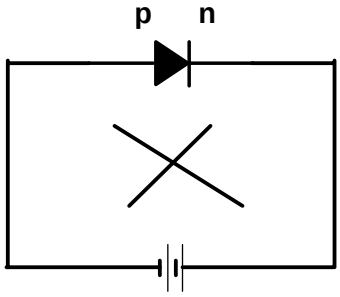
A semiconductor diode is denoted by the following symbol:

Hello: 4275803
Email: sajitcshakya@gmail.com
Web Address: http://sajitsir.tripod.com

The symbol consists of an arrow whose tip has a bar. The bar denotes the negative end (cathode) of the Diode. The base of the arrow denotes the positive end. When it is connected in forward biased and reverse biased mode it would appear as:



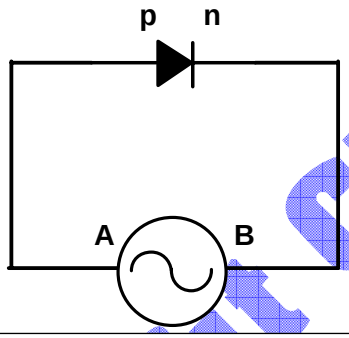
Forward Bias: Electrons flowing anticlockwise.



Reverse Bias: Electron flow not possible.

As is evident from the figure, electron flow is possible only from right to left through the diode, not reverse. However, if the flow of positive charges is taken into account, as in conventional concept, current flows from left to right through the diode. The direction of the arrow shows this possible direction of the positive charges.

Half wave rectifier (Behavior of Semiconductor to AC)

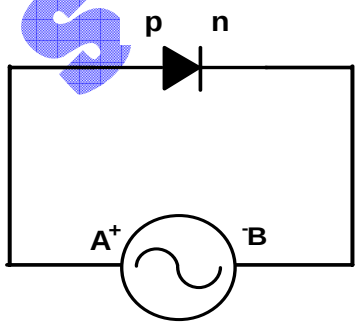


Connection of diode to AC source

When a diode is connected to DC source, it might or might not allow the flow of electrons, depending upon whether it is forward biased or reverse biased. If it is forward biased, it allows the flow of electrons, whereas if it is reverse biased, it does not.

However if the diode is connected to an AC source, the behavior is somewhat different. An AC source emits electrons from both terminals but the emission is alternate (occurs turn by turn). If electrons are coming from first terminal, they are entering the other. However, if the emission is from the second, they enter from the first.

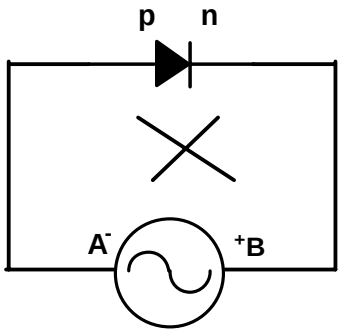
In the figure, such an AC source is connected across the diode. At a certain instant, electrons are coming from terminal B. So B is negative and A is positive at that moment. So the diode becomes forward biased and conduct at that time. The variation of current will follow the variation of voltage and in the graph the shape formed by its curve will be the same as that of the voltage.



Conduction possible.

Hello: 4275803
Email: sajitcshakya@gmail.com
Web Address: http://sajitsir.tripod.com

After some time, electrons start emerging from A. So A becomes negative and B becomes positive. The diode will be reverse biased at that time and so will not allow the flow of electrons, making the value of current zero. It will remain zero till A is negative.

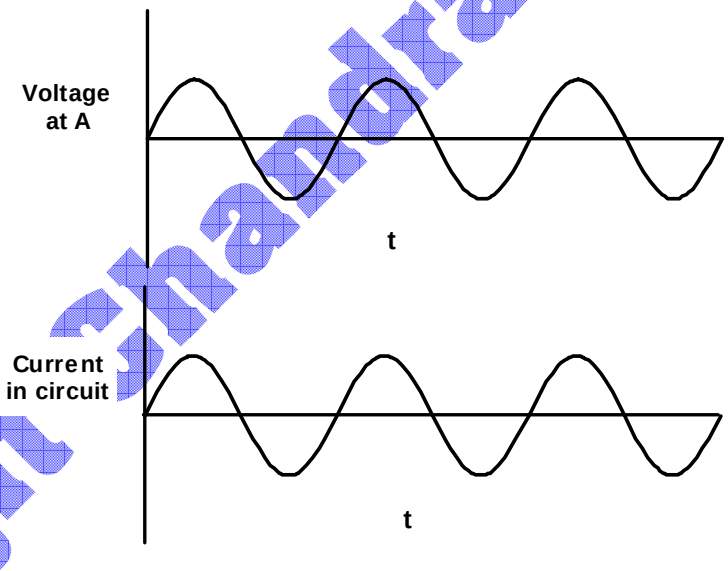


Conduction not possible.

Again after some time, the initial conditions prevail and the diode starts conducting. The whole process goes on till the source is operating.

Had the diode been absent in the circuit, the electrons would have been performing to and fro motion under the influence of alternating voltage. But due to the presence of the diode, it travels for some time, pauses for equal interval of time, again travels in the same direction for same amount of time, again pauses and so on. So a diode makes the flow of electrons "unidirectional" instead of bidirectional.

As is observed from the graph, the diode uses only the positive voltage to create current and cuts off the negative voltage. So this arrangement is also called as a "**Rectifier**" (rectify means correction or removal of incorrect things or phenomena). But it can give current out of half cycle of voltage only, so it is further called "**Half Wave Rectifier**".



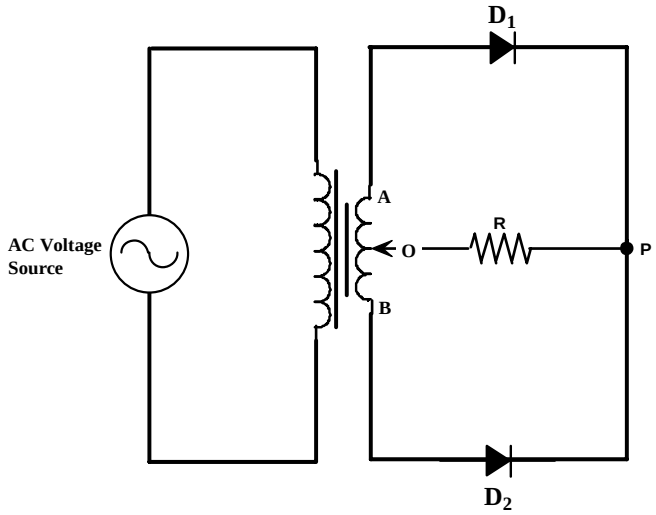
So a diode gives a unidirectional flow of electrons out of alternating one. But it causes the waste of half of the voltage giving a low average current.

Sajit Chandra Shakya
Physics Educator
Swarnim Basti, Chagal, Kathmandu

Hello: 4275803
Email: sajitcshakya@gmail.com
Web Address: http://sajitsir.tripod.com

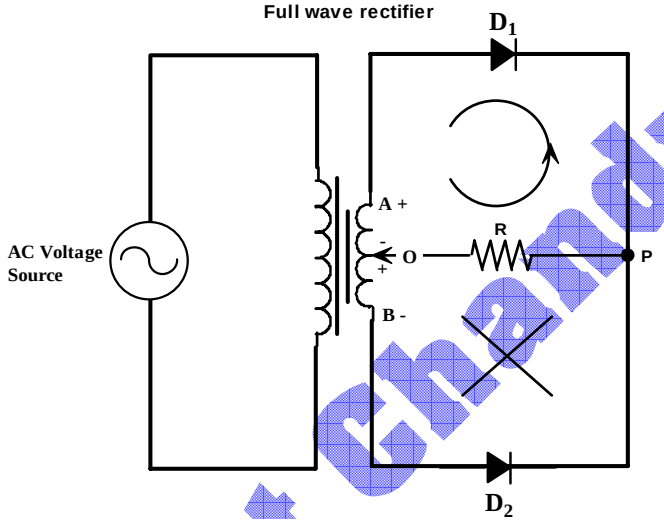
Full Wave Rectifier

The half wave rectifier makes the flow of current unidirectional, but there are a lot of losses and the average current is also very low. Only half of the voltage (positive voltage) can create current. So a method was sought for to obtain current out of the negative voltage also, that too in the same direction. This required an extra diode and a center tapped transformer (which is the transformer with a connecting terminal available through the center of the secondary coil as well).



Full wave rectifier

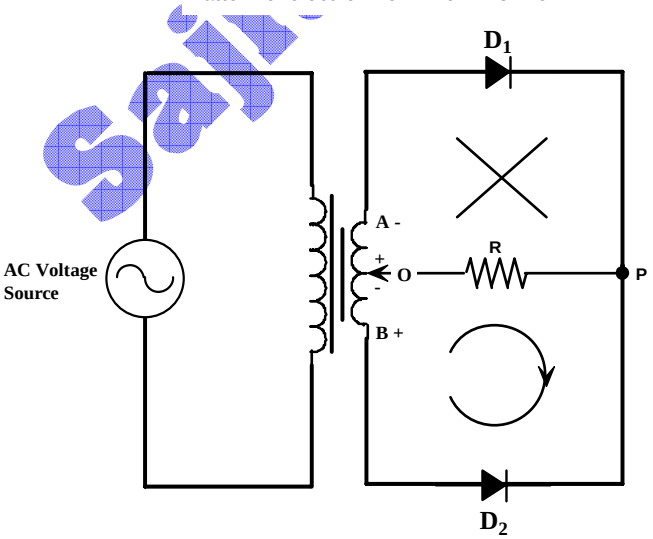
This could be achieved by connecting the p-sides of two diodes D_1 and D_2 to the ends of the secondary coil of a center tapped transformer. The n-sides are connected to a common point P and then led to a resistor (or any power consuming device) and then to the central terminal O of the secondary coil. The ends of the primary coil of the transformer are connected to the main AC source.



Pattern of electron flow when A is +ve.

When the AC source operates, it starts to induce varying magnetic fields, hence varying magnetic lines of forces. These oscillating magnetic lines of forces sweep through the secondary coils and thus produce oscillations in its electrons. So electrons start emerging from A and B alternately.

At a certain instant, electrons are emerging from end B of the secondary coil. At that time B will be negative and A will be positive. The center O will be at neutral, i.e. zero. But for A, O will be negative (because it is at lower potential level, lower meaning negative), and for B, O will be positive (because it is at higher potential level, higher meaning positive). This will forward bias diode D_1 and reverse bias D_2. So electrons will flow from O, resistance, P, D_1, and then to A, causing circuit completion. However electrons do not go to D_2.



Pattern of electron flow when A is - ve.

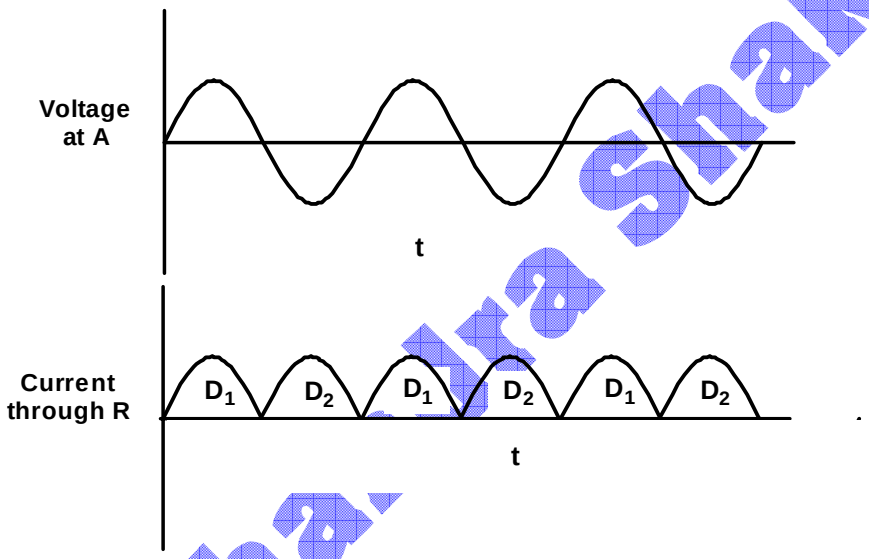
After some time, electrons start emerging from A, which will make A negative and B

Hello: 4275803
Email: sajitcshakya@gmail.com
Web Address: http://sajitsir.tripod.com

positive. The center O will again be at neutral, i.e. zero. Now for A, O will be positive and for B, O will be negative. This will forward bias diode D_2 and reverse bias D_1. So electrons will flow from O, resistance, P, D_2, and then to B, causing circuit completion. However electrons do not go to D_1.

So whether A is positive or negative, electrons will keep flowing through the resistance R and the direction of electrons will be the same. The current is therefore unidirectional and the average value will be also high, since it has got a certain positive value whether the voltage is positive or negative. The speed of the electrons is not uniform yet; they alternate between higher and lower speeds, but move in the same direction.

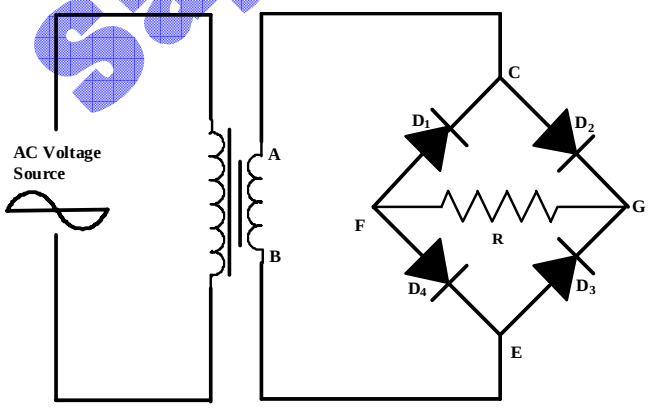
The variation of voltages at A and the variation of current through the resistance are as shown in the figure. The current is contributed by D_1 and D_2 alternately, but never simultaneously.



Que: Why is full wave rectifier more useful than half wave rectifier?

Bridge rectifier

A normal full wave rectifier needs the use of a center tapped transformer, which is often expensive. In order to prevent its use, two additional diodes are needed. The four diodes form a bridge like arrangement, much like in Wheatstone Bridge.



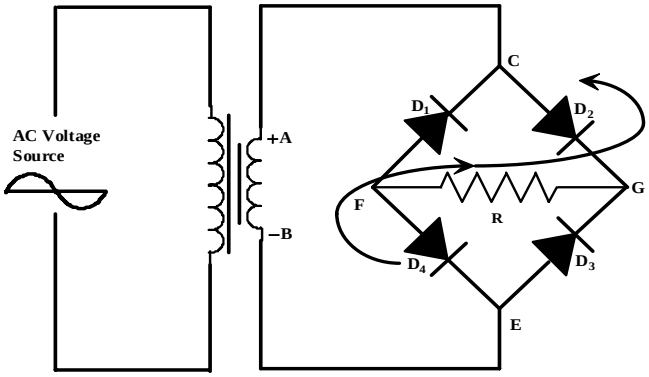
Full wave rectifier

The arrangement is as shown in the 1st figure. The ends of the secondary coil of a transformer are connected to the Wheatstone bridge arrangement at points C and E. When the AC source is operated, it will induce voltage in the secondary coil also and electrons start to emerge from the ends of the coil alternately.

Hello: 4275803
Email: sajitcshakya@gmail.com
Web Address: http://sajitsir.tripod.com

Let at an instant, electrons are emerging from point B. At that time,

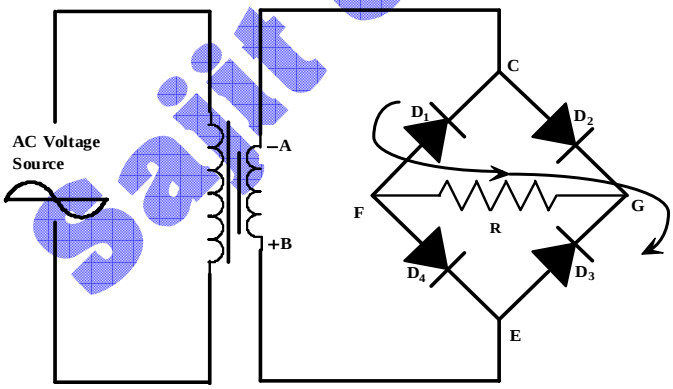
1. A is positive and B is negative.
2. E also becomes negative.
3. D_4 will conduct (because of being forward biased) whereas D_3 will not (because of being reverse biased).
4. Then electrons will reach point F. They have two choices – to flow through R or through D_1.
5. They can not flow through D_1 because it will not conduct (because of being reverse biased already).
6. So electrons are forced through R to G.
7. At G also they have two choices – to go through D_2 or D_3.
8. They can not go through D_3 because it is already non-conducting because of electron pressure from the other side.
9. So electrons are forced through D_2.
10. Then they reach point C, where they again have two choices – to go through D_1 or to go to A.
11. They can not go through D_1 because it is already non-conducting due to the electron pressure from the other side.
12. So electrons go to point A, completing a circuit. So current flows through the arrangement as well as the power consuming resistor used as shown in the 2nd figure.



Full wave rectifier

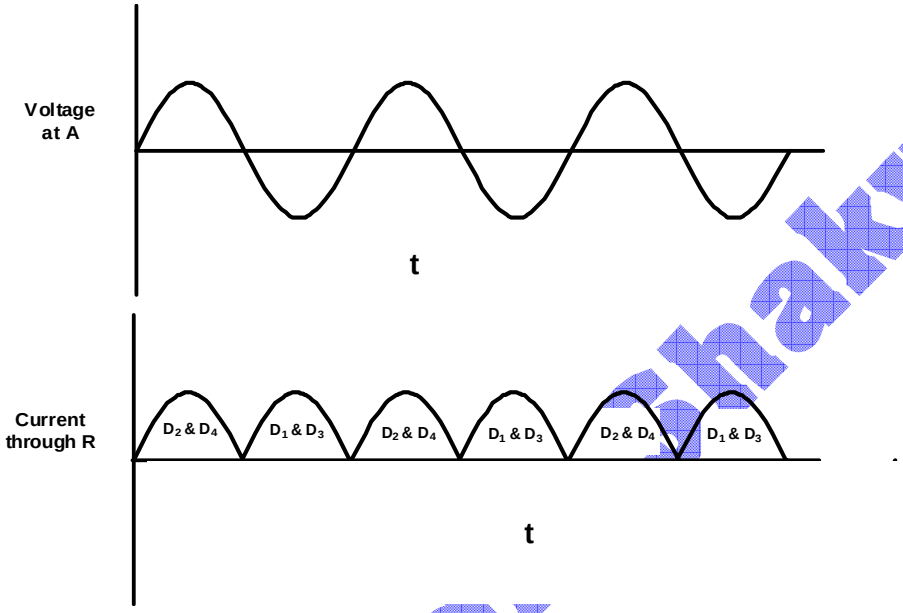
After some time, electrons start emerging from point A. At that time,

1. B is positive and A is negative.
2. C also becomes negative.
3. D_1 will conduct (because of being forward biased) whereas D_2 will not (because of being reverse biased).
4. Then electrons will reach point F. They have two choices – to flow through R or through D_4.
5. They can not flow through D_4 because it will not conduct (because of being reverse biased).
6. So electrons are forced through R to G.
7. At G also they have two choices – to go through D_2 or D_3.
8. They can not go through D_2 because it is already non-conducting because of electron pressure from the other side.
9. So electrons are forced through D_3.



Hello: 4275803
Email: sajitcshakya@gmail.com
Web Address: http://sajitsir.tripod.com

10. Then they reach point E, where they again have two choices – to go through D_4 or to go to B.
11. They can not go through D_4 because it is already non-conducting due to the electron pressure from the other side.
12. So electrons go to point B, completing a circuit. So current flows through the arrangement as well as the power consuming resistor used as shown in the 3rd figure.

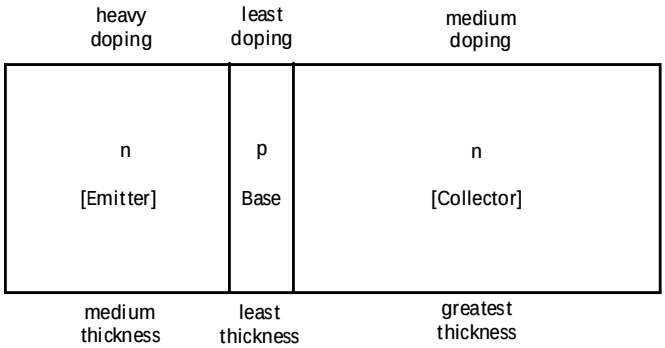


Whatever be the direction of the voltage coming from the source or from points A or B, the current through R is always in the same direction, from left to right. The graphs showing the initial voltage waveform (with point A as the reference point) and the output current observed through the resistor R is as shown in the figure. In each half current wave, one half is contributed combinedly by D_1 and D_4, whereas the other half by D_2 and D_3 respectively.

Transistor

The development of a transistor started with the addition of more doped layers into already existing two layers of semiconductor diodes. A diode has two doped layers – n and p, each respectively with free electrons and holes. So a third layer was added to such structure. The thickness of the layers, and their doping levels were varied in innumerable ways. In addition, several types of connections were tried on those three regions and at last a connection was finalized which could fulfill specific purposes.

Such a three region structure appears as:



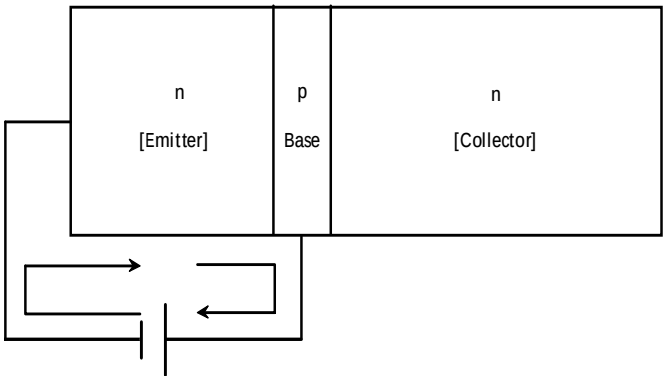
The three layers are of varied thicknesses. The middle layer is the thinnest. The two layers at the ender are thicker, with one layer thicker than the other. Besides the middle layer has opposite doping compared to the other two layers. For example, if the middle layer is doped positively (p type), the

Hello: 4275803
Email: sajitcshakya@gmail.com
Web Address: http://sajitsir.tripod.com

others will be doped negatively (n type). Similarly if the middle is doped negatively, the others are doped positively. They are respectively called 'npn' and 'pnp' structures. The middle layer (the thinnest one) is the least doped. The layer with medium thickness is doped heavily and the thickest layer has the medium doping.

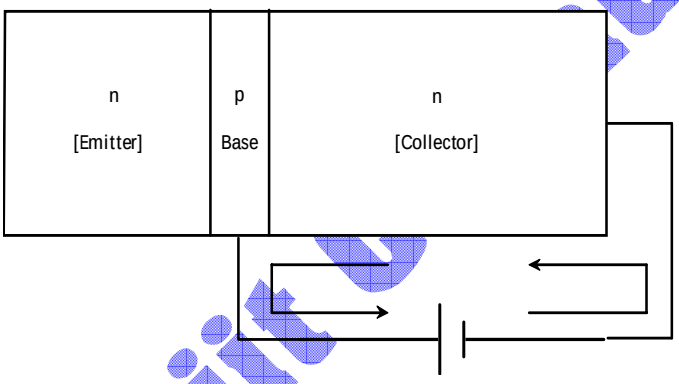
Then several types of connections are tried to study the behavior of the structure.

1. When the connection is as shown in the figure:



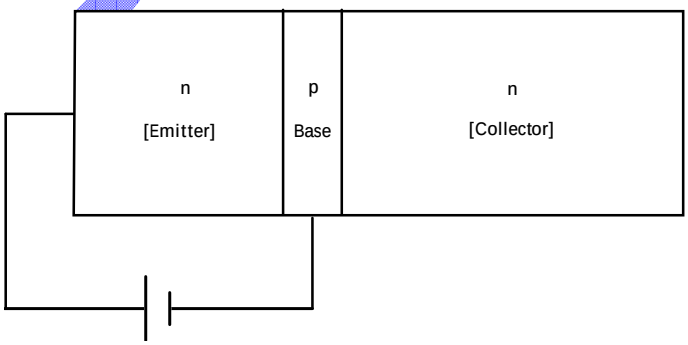
In such condition, the two layers at the left form a diode, whose n side is connected to the negative terminal of the cell and p side to the positive, making the diode forward biased. So the only effect that is possible is the flow of electrons in clockwise direction. The third layer at the right does not play any part in the flow ar blockade of electrons in the circuit.

2. When the connection is as shown as in the following figure:



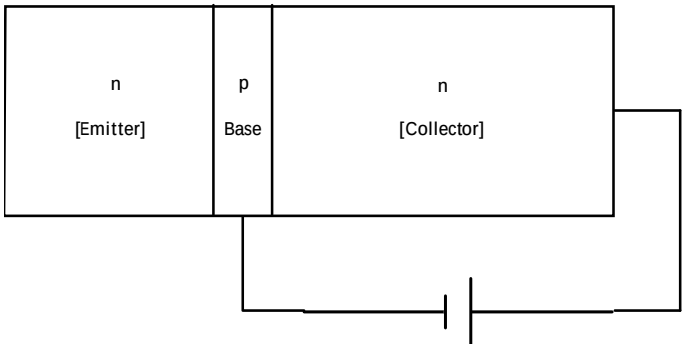
In such condition, the two layers at the right form a forward biased diode once again. So electrons will flow in the anticlockwise direction. The layer at the left has no part to play in such condition.

3. When the connection is as shown in the figure:



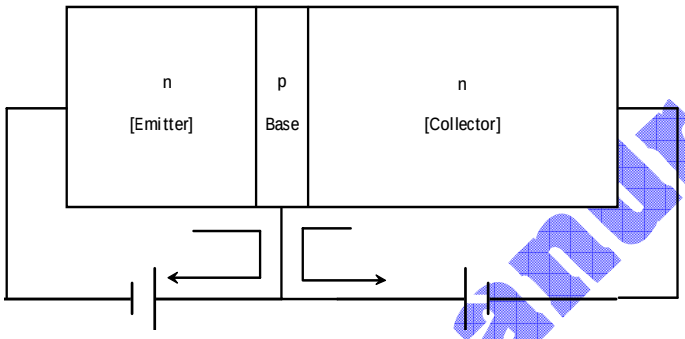
The two layers at the left which form a diode will be reverse biased. So under normal conditions, there is no flow of electrons at all.

4. When the connection is as shown in the figure:



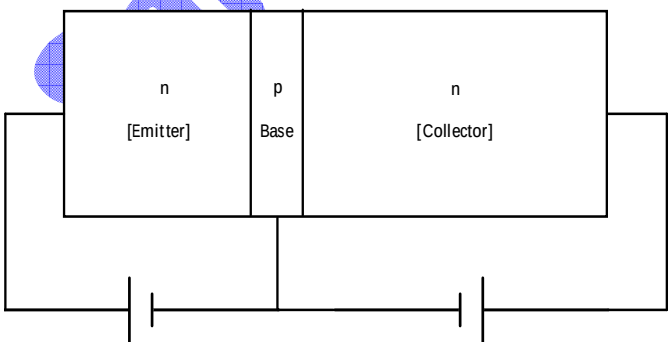
The two layers at the right which form a diode will be reverse biased. So under normal conditions, there is no flow of electrons at all.

5. When a combined connection is tried as shown in the following figure:



In such condition, the combination of middle-left layer and middle-right layer both form forward biased diodes. So electrons from both the layers at the ends will be pushed to the middle layer, come out of it and are redistributed to the respective circuits. So there is a heavy flow of current. However, electrons being pushed from the layer at the left can not go to the layer at the right due to the latter's connection to the negative potential. Similarly electrons being pushed from the layer at the right can not go to the layer at the left because of its connection to the negative potential. Electrons from each layer at the edges will accumulate at the middle layer and come out of it for proper distribution.

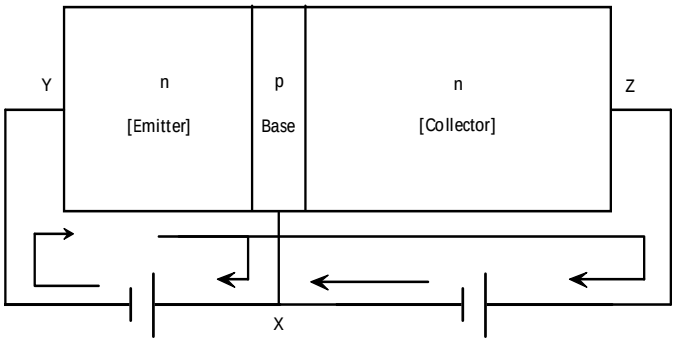
6. When a combined connection is tried as shown in the following figure:



In this condition both the diodes will be reverse biased. So there will be no flow of electrons at all.

Hello: 4275803
Email: sajitcshakya@gmail.com
Web Address: http://sajitsir.tripod.com

7. When a combined connection is tried as shown in the following figure:



The general conception will be that electrons will flow in the circuit at the left and there would be no current in the circuit at the left. But the special design of the structure (varied thicknesses, varied levels of doping and exact positioning), there are several events.

Since the layer at the left is very heavily doped and is in contact with the negative terminal of a cell, there will be a rush of free electrons from this layer to the right soon after connection. Their main target is the holes in the middle layer. But there are very less number of holes in the middle layer (due to its less doping and less thickness). So very few free electrons can be absorbed by the holes, which are then attracted by the positive terminal of the cell at the left.

The unabsorbed electrons now rush into the layer at the right. This event is made possible due to two factors. One is the extremely thin middle layer which can not retard the freely flowing electrons effectively. The other is the fact that the layer at the right is connected to the positive terminal of the cell at the right and this cell is generally of high potential difference. This will help attract the electrons as soon as they venture to this region. These electrons then reach point 'Z' and then recombine with the electrons from the middle layer at point 'X'. These electrons then reach Y, reenter the structure from the layer at the left and the whole process continues indefinitely or till the structure arrangement remains intact.

What's the advantage of such an arrangement?

Such structure and arrangement gives great control over current flowing in the circuit at the right by varying several parameters. The main control parameter is the thickness of the middle layer. If it is thick, fewer electrons reach the other side. If it is thin, more electrons will reach the other side. So this layer serves more as a control layer, so it is called as the '**BASE'** (remember the word "base station", or "base force" or "base layer"). The layer at the left (which is doped heavily and has medium thickness), provides a rich supply of electrons for the whole circuit. So it is called the '**EMITTER'**. The terminal at the right, which gathers the electrons that have come from the emitter and were spared by the base, is called the '**COLLECTOR'**. So the three layers can be shown in the following table featuring their overall aspects.

Features	**Emitter**	**Base**	**Collector**
Doping level	Heavy	Least	Medium
Doping type	Same as the collector	Different from others	Same as the emitter
Thickness	Medium	Least	Largest
Positioning	Edge	Middle	Edge

Hello: 4275803
Email: sajitcshakya@gmail.com
Web Address: http://sajitsir.tripod.com

The free electrons going from the emitter distributes to the base and the collector, come out of them; combine at X and again enter the emitter. So the emitter current should be equal in magnitude to the sum of the current coming out of the base and collector. If they are respectively denoted by I_E, I_B and I_C,

Mathematically,

$I_E = I_B + I_C$ (do you see some use of Kirchoff's Current Law??)

Relationship between 'α' and 'β' parameter of a transistor

The 'α' parameter of a transistor is equal to the ratio of the collector to the emitter current.

i.e. $\alpha = \frac{I_C}{I_E}$ ..... [i]

Similarly,
The 'β' parameter of a transistor is equal to the ratio of the collector to the base current.

i.e. $\beta = \frac{I_C}{I_B}$ ..... [ii]

In any transistor, the emitter current is composed of the total of the collector and the base current.

i.e. $I_E = I_C + I_B$ ..... [iii]

Dividing both sides by I_C gives

$\frac{I_E}{I_C} = \frac{I_C}{I_C} + \frac{I_B}{I_C}$, or, $\frac{1}{I_C/I_E} = 1 + \frac{1}{I_C/I_B}$,

or, $\frac{1}{\alpha} = 1 + \frac{1}{\beta}$, or, $\frac{1}{\alpha} = \frac{\beta + 1}{\beta}$,

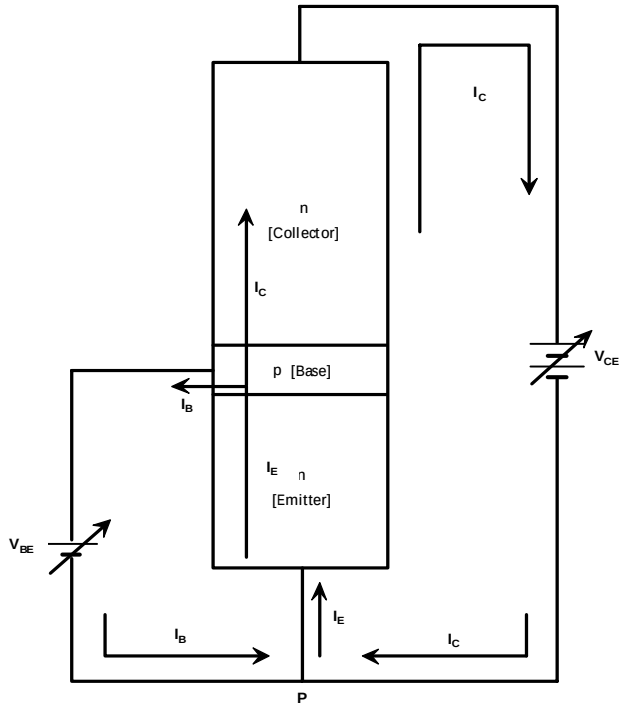
or, $\alpha = \frac{\beta}{\beta + 1}$ ..... [iv]

This relation helps determine the value of 'α' once the value of 'β' is known, and vice versa.

Common Emitter Characteristics

Sajit Chandra Shakya
Physics Educator
Swarnim Basti, Chagal, Kathmandu

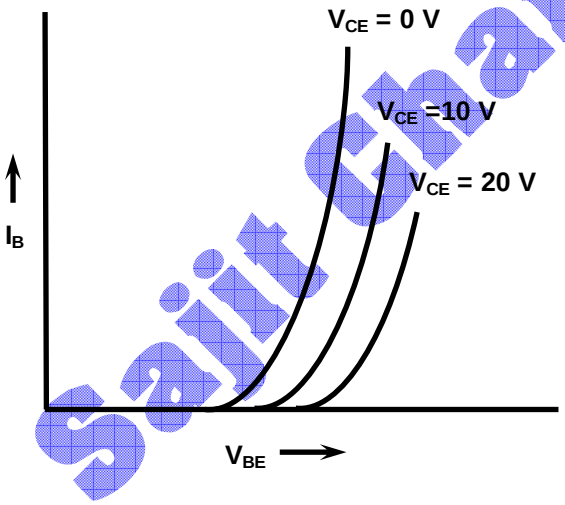
Hello: 4275803
Email: sajitcshakya@gmail.com
Web Address: http://sajitsir.tripod.com



The common emitter mode is the most used configuration of a transistor. In this connection, the emitter belongs to both the circuits.

The base and the emitter are connected across a smaller voltage source V_{BE}. Similarly the collector and the emitter are also connected to another voltage source V_{CE}, which has a relatively higher potential difference. The base-emitter circuit and the collector-emitter circuit meet at a common point P. In this configuration, both V_{BE} and V_{CE} are working together to push free electrons of the emitter. Most of these free electrons jump over to the collector while some are absorbed by the base, respectively forming the collector current I_C and the base current I_B. these two currents combine at point P to form the emitter current I_E. Because of the unique structure and configuration, the transistor in this mode shows different behaviors according to the controlling factors.

Input characteristics:



The variation of the base current I_B with the variation of the base-emitter voltage V_{BE} at fixed values of V_{CE} is called as the input characteristics. As per the definition, the voltage V_{CE} is kept constant at a certain value, and then V_{BE} is varied slowly to observe the behavior of I_B.

1. Let V_{BE} be set to zero volt. This means the collector region has no role to play whatsoever. The transistor will then act as a simple diode. So base current will not flow till V_{BE} becomes equal to 0.7 or 0.3 V depending upon the material used to construct the transistor. Then current starts to flow rapidly, similar to that as in a diode.
2. Let V_{BE} be set to a higher value of potential difference. This means increase in the electrons accelerating force of the positive terminal of V_{CE}. Therefore more electrons will rush to the collector causing a reduction in the share of I_B and also causing difficulty in its rise even if V_{BE} is increased to higher values.
3. As V_{BE} goes on increasing, the value as well as the rise of I_B becomes more and more difficult. The overall behavior is as shown in the figure.

Hello: 4275803
Email: sajitcshakya@gmail.com
Web Address: http://sajitsir.tripod.com

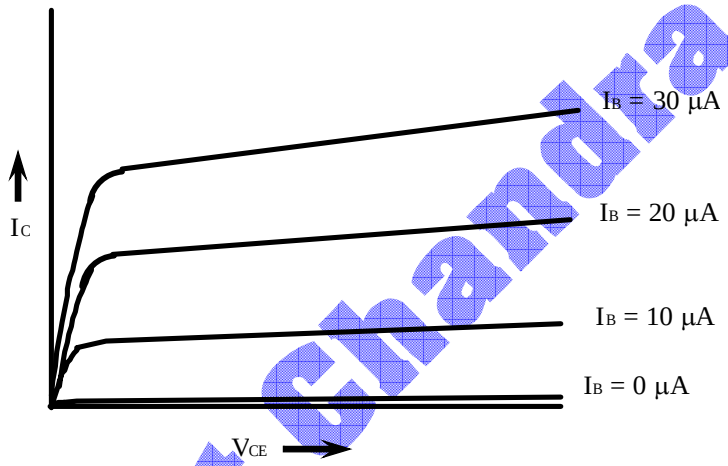
The graph showing the variation of base current with the base-emitter voltage gives a method of finding the input resistance of the transistor, which is the ratio of the base-emitter voltage to the base current at particular point. As the curve does not have constant slope, it is evident that the resistance varies a lot. If R_{in} denotes the input resistance, then

$$R_{in} = \frac{\Delta V_{BE}}{\Delta I_B}$$

Output characteristics:

The variation of the collector current I_C with the variation of the collector-emitter voltage V_{CE} at fixed values of I_B is called as the output characteristics. As per the definition, the current I_B is kept constant at a certain value, and then V_{CE} is varied slowly to observe the behavior of I_C.

1. Let I_B be first set to 0 A. It means the free electrons are no longer accelerated from the emitter to the collector. Then they also can not reach the collector, there fore the collector current will be zero whatever be the value of V_{CE}. However, due to the high value of V_{CE}, it might pull some electrons from various regions of the structure and might develop a small leakage current as shown in the figure.



2. Let the base emitter voltage V_{BE} be increased slightly so that some electrons will flow from the emitter. This will increase the value of the base current (say to 10 μA). It would be like giving an accelerating push to the electrons because of which the collector current I_C also increases. However, the rise of I_C is rapid in the beginning because the current is contributed by free electrons rushing from the highly doped emitter and the power sources. Later the electrons from the emitter would be finished and the current is composed only of the electrons coming from the power sources. So the rate of rise is not as steep as in the beginning. The pattern of rise of current would be same when current is set to other values also as $I_B = 10$ μA, $I_B = 20$ μA, $I_B = 30$ μA, etc.

The output characteristics help determine the output resistance. For this the ratio of change in V_{CE} to I_C has to be determined from the characteristics graph. Mathematically,

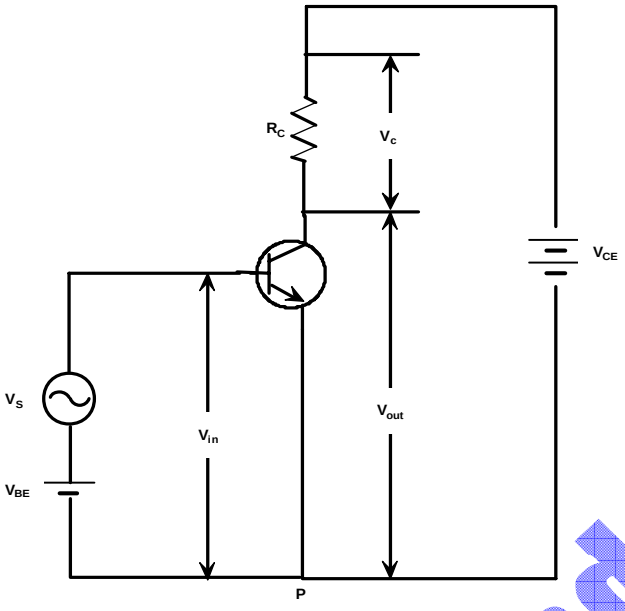
$$R_{out} = \frac{\Delta V_{CE}}{\Delta I_C}$$

Sajit Chandra Shakya
Physics Educator
Swarnim Basti, Chagal, Kathmandu

Hello: 4275803
Email: sajitcshakya@gmail.com
Web Address: http://sajitsir.tripod.com

Transistor as Amplifier

An amplifier is a device which can increase the amplitude of a signal (varying current supplied to it. The source of the varying current would be any ac signal generator, signal from a microphone, signal from waves received from the air, etc).



The amplifying action of a transistor is due to the overall structure and construction of the device itself. The emitter is highly doped, so produces a heavy flow of electrons even if the base emitter voltage rises a little, causing a small increase in base current I_B but larger increase in I_C. Similarly when the base emitter voltage falls a little, I_B shows a small decline but I_C decreases very rapidly.

It means,

When I_B rises a little, I_C rises heavily ($\because I_C = \beta I_B$)

&

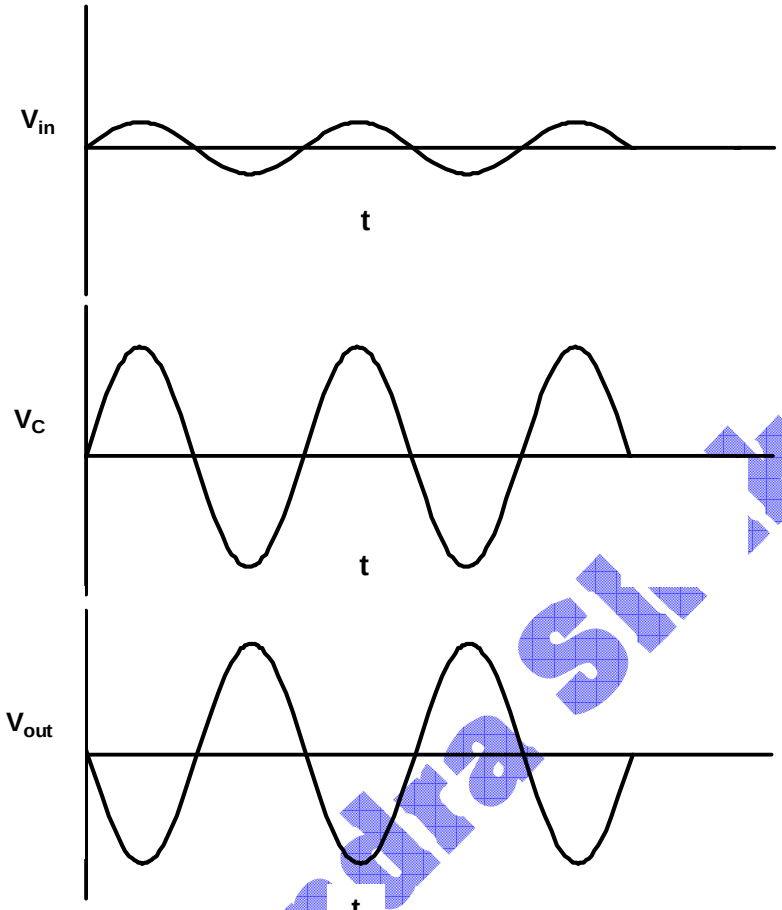
When I_B falls a little, I_C also falls heavily($\because I_C = \beta I_B$).

When a signal source is connected to the base emitter circuit and operated, it will constantly give an alternating voltage, sometimes high and sometimes low. This voltage V_S, when added up to V_{BE}, gives varying values of V_{in}.

So when,

V_{in} is high, I_B is high, I_C will be very high, V_C will also be very high ($\because V_C = I_C R_C$),
$\therefore$ V_{out} ($= V_{CE} - V_C$) will be low.

Hello: 4275803
Email: sajitcshakya@gmail.com
Web Address: http://sajitsir.tripod.com



Similarly, when

V_{in} is low, I_B is low, I_C will be very low, V_C will be very low ($\because V_C = I_C R_C$),

$\therefore$ V_{out} (= $V_{CE} - V_C$) will be high.

Therefore if the variation between the highest and the lowest values of the input voltage be denoted by v_{in}, the variation of V_C be denoted by v_c, and that of the output voltage V_{out} be denoted by v_{out}, then both v_c and v_{out} will be very large compared with v_{in}. That's why the amplitude of the voltage will be magnified by the device, resulting in the amount of power they can deliver. Besides, the variation of V_C and V_{in} follow a same pattern, so the wave forms representing them would be of same phase. However, the variation of V_{out} is opposite compared to that of V_{in}, so they will of opposite phase compared to each other, as shown in the graph.

The amplifier gain (A) of such amplifier is given by taking the ratio of the output voltage to the input voltage.

i.e., $A = \frac{v_{out}}{v_{in}}$