ENG1004 - Electronics & Sensors

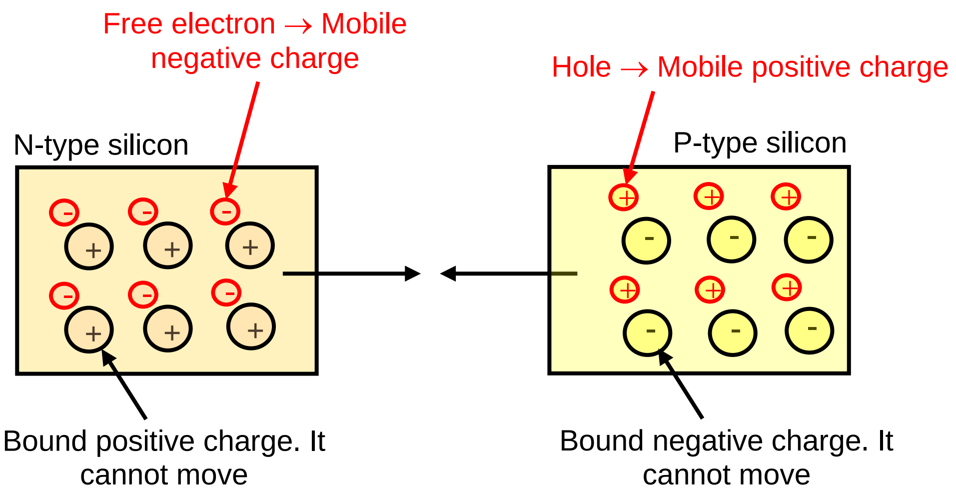
Part 2 – PN Junctions (Diodes)

S. Le Goff, School of Engineering, Newcastle University

1. Basic Operation of PN Junctions (Diodes)

We have seen that a crystal of pure silicon can be turned into a relatively good electrical conductor by adding an impurity such as arsenic/phosphorus for an N-type semiconductor or boron/aluminium/gallium for a P-type semiconductor. By itself, a single type of semiconductor material is not very useful. But, something interesting happens when a single semiconductor crystal contains both P-type and N-type regions.

Hereafter, we examine the properties of a single silicon crystal which is half N-type and half P-type. The two types are shown separated, as if they were two separate crystals being put in contact. In the real world, such crystals cannot be joined together usefully. Therefore, a practical PN junction can only be created by inserting different impurities into different parts of a single crystal.



When we join the N- and P-type crystals together, an interesting interaction occurs around the junction.

Majority carriers (electrons) from the N region near the PN interface tend to diffuse into the P region. As electrons diffuse, they leave the bound positive charges in the N region. Electrons diffuse from an area where there is a high concentration of them (N-type material) to an area where there is a low concentration of them (P-type material).

Note that diffusion is a very natural process that is present in many physical phenomena (e.g., think about what happens when ink is injected into a glass of water).

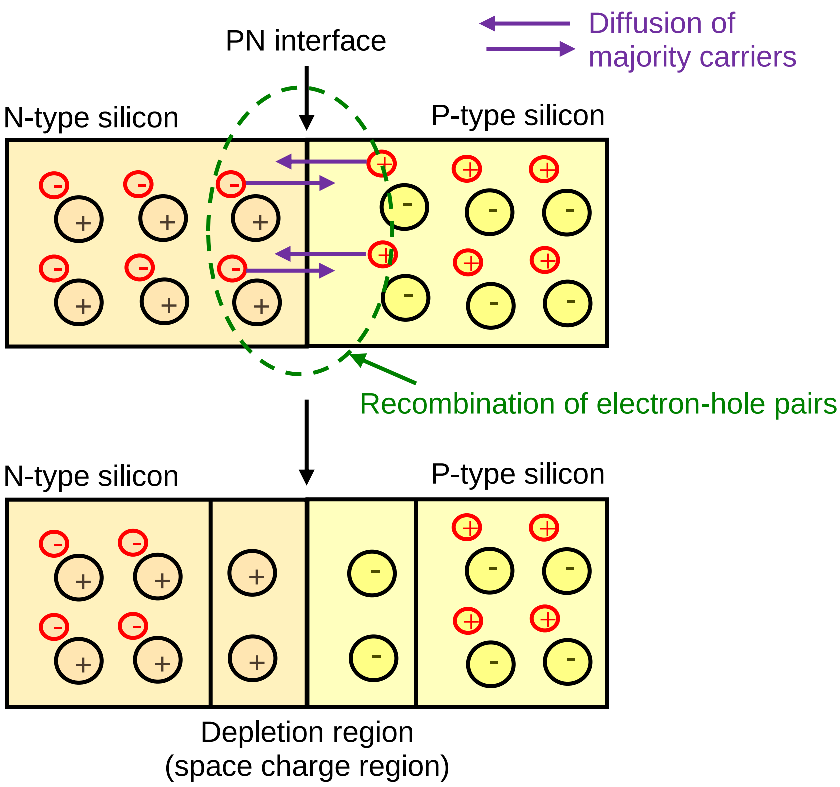
Likewise, majority carriers (holes) from the P-type region near the PN interface begin to diffuse into the N-type region, leaving the fixed negative charges. The reason why holes diffuse is that they move naturally from an area where there many of them (P-type material) to an area where their concentration is much lower (N-type material).

Around the interface, the majority carriers recombine (recombination of electron-hole pairs) and only the bound charges are left on both sides. The regions nearby the PN interface thus lose their neutrality and become charged, forming the space charge region also known as *depletion* region.

The *built-in* electric field created by the space charge region opposes the diffusion process for both electrons and holes. There are therefore two concurrent phenomena: the diffusion process that tends to generate more space charge and the built-in electric field generated by the space charge that tends to counteract the diffusion.

At some stage, the strength of the electric field generated by the space charge becomes sufficient to completely cancel the diffusion effect, and the majority carriers then stop moving towards the PN interface. The PN junction has reached a state of equilibrium.

The space charge region is a zone with a net charge provided by the fixed positive and negative charges that have been left “uncovered” by the diffusion of majority carriers. When equilibrium is reached, the region around the PN interface is completely depleted of majority carriers.



At equilibrium, there exist a voltage V_d across the junction called the *built-in potential*. Solid-state physics tell us that this voltage can be expressed as

$$V_d = \frac{kT}{q} \ln\left(\frac{N_A \cdot N_D}{n_i^2}\right)$$

with - q: Charge of an electron (≈ 1.602×10^{-19} C);

- T: Temperature in degrees Kelvin;

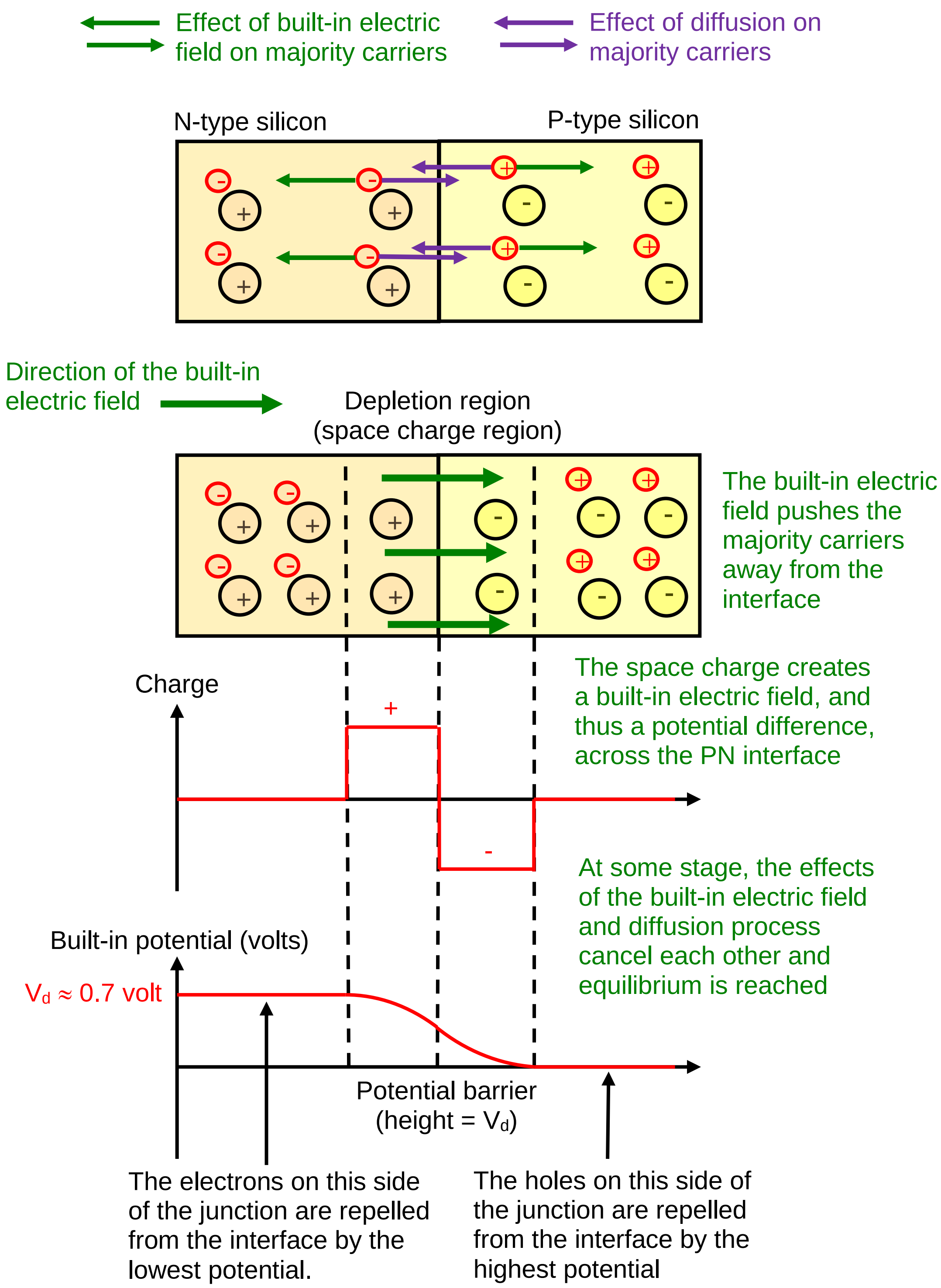
- k: Boltzmann's constant (≈ 1.38×10^{-23} J/K);

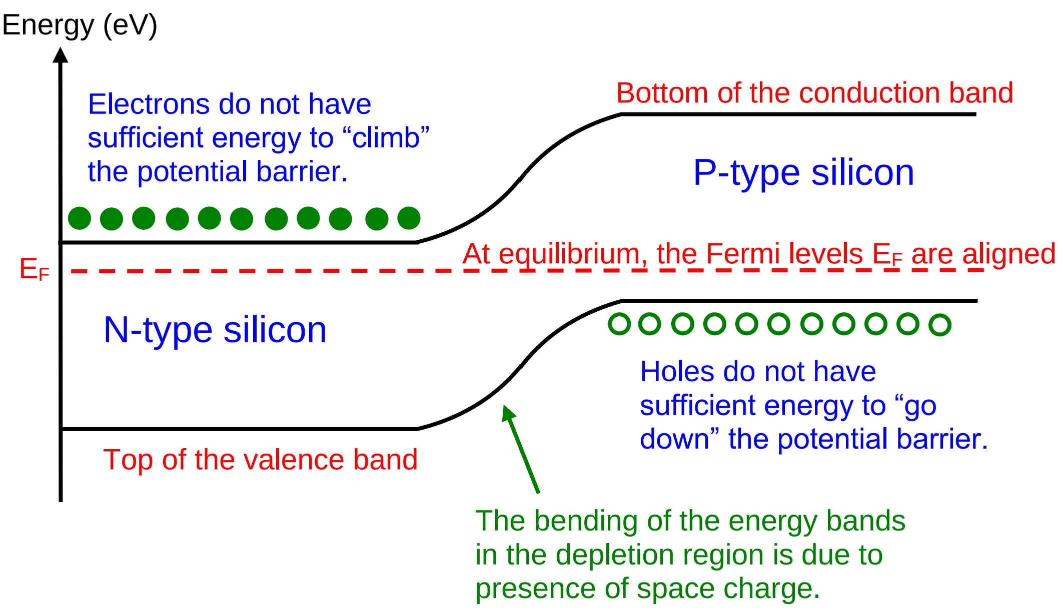
- n_i : Intrinsic carrier concentration in a pure sample of the semiconductor (≈ 1.5×10^{10} cm^3 at 300 K for silicon);

- N_A : Concentration of acceptor atoms in the P-type semiconductor, in atoms.cm^{-3};

- N_D : Concentration of donor atoms in the N-type semiconductor, in atoms.cm^{-3}.

For instance, for a silicon PN junction at 300 K, if $N_A = 10^{15}$ atoms.cm^{-3} and $N_D = 10^{16}$ atoms.cm^{-3}, we have $V_d = \frac{kT}{q}\ln\left(\frac{N_A \cdot N_D}{n_i^2}\right) = \frac{1.38\times10^{-23}\times300}{1.602\times10^{-19}}\ln\left(\frac{10^{15}\times10^{16}}{2.25\times10^{20}}\right) \approx 633$ mV.

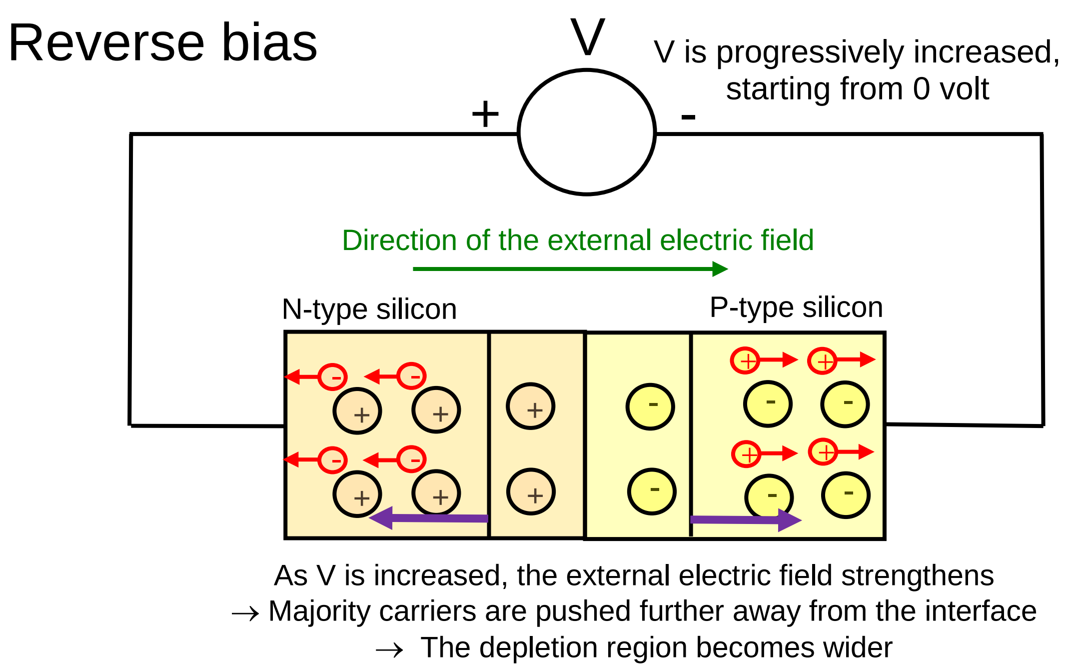




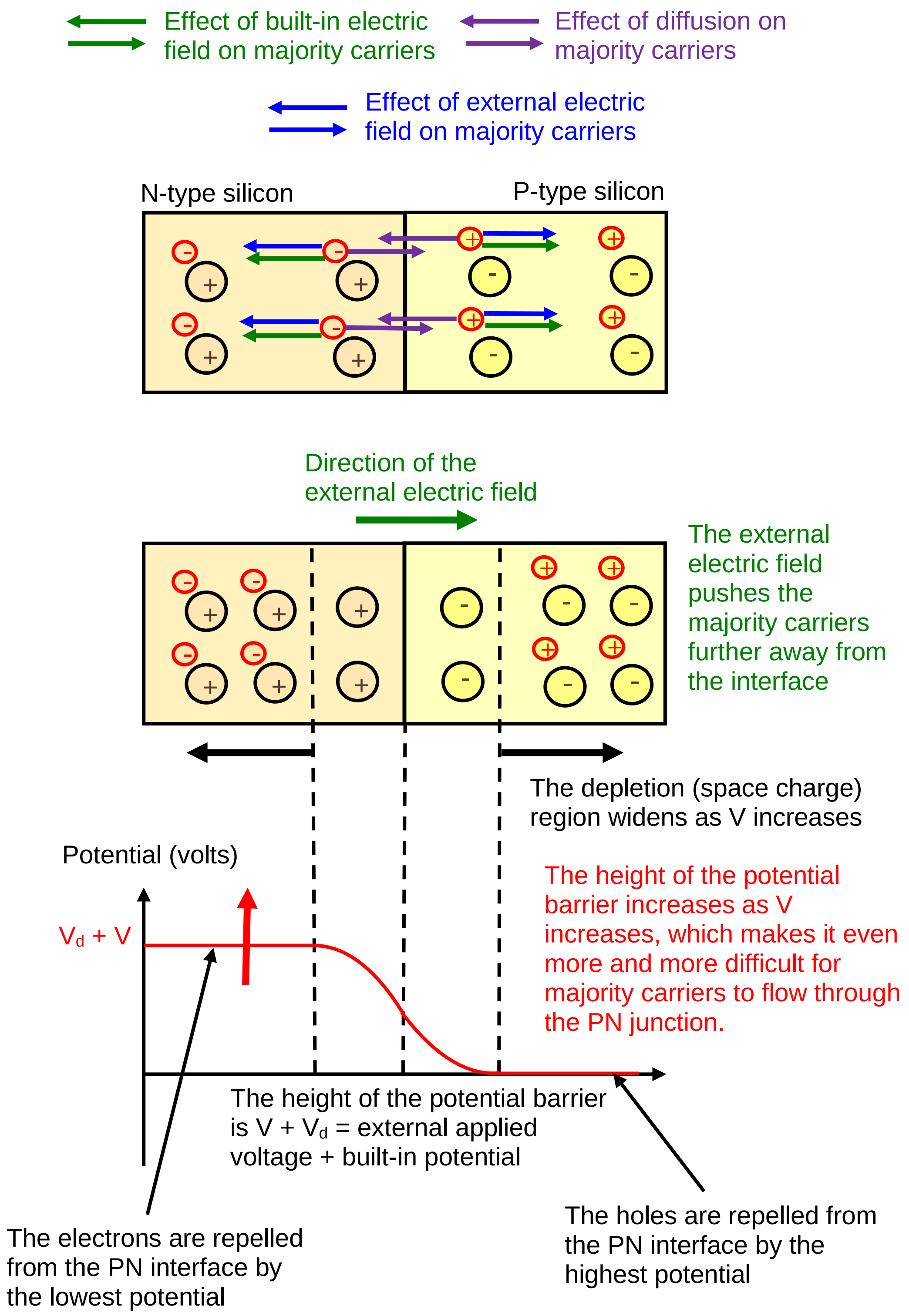
2. Reverse Bias

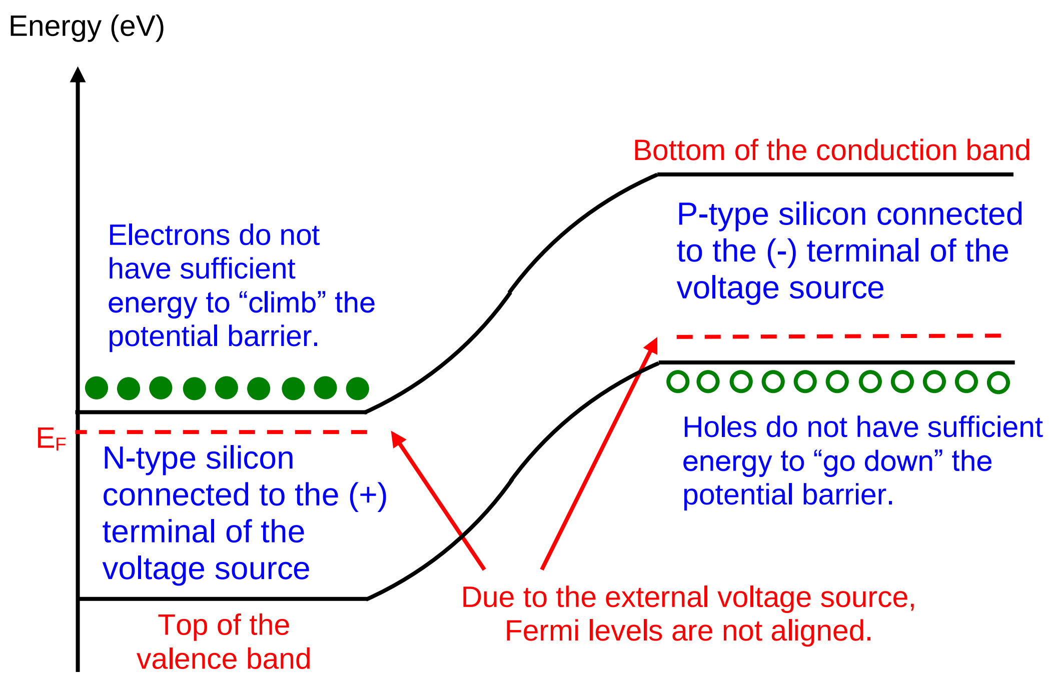
Suppose now that we apply a voltage across our PN junction. Assume first that the positive voltage is applied to the N-type material. This is known as *reverse bias*.

An electrical field is established across the PN junction due the voltage applied across it. Under reverse bias, the direction of this external electric field is such that the majority carriers are further attracted away from the PN interface.



Since all current majority carriers are attracted further away from the PN interface, the depletion region grows correspondingly larger. Therefore, despite the application of a voltage across the PN junction, there can be no significant current flowing through the crystal simply because no majority carriers can cross the junction. This is known as reverse bias applied to the semiconductor crystal.





Under reverse bias, the PN junction behaves like an insulator as no (significant) current can flow through it.

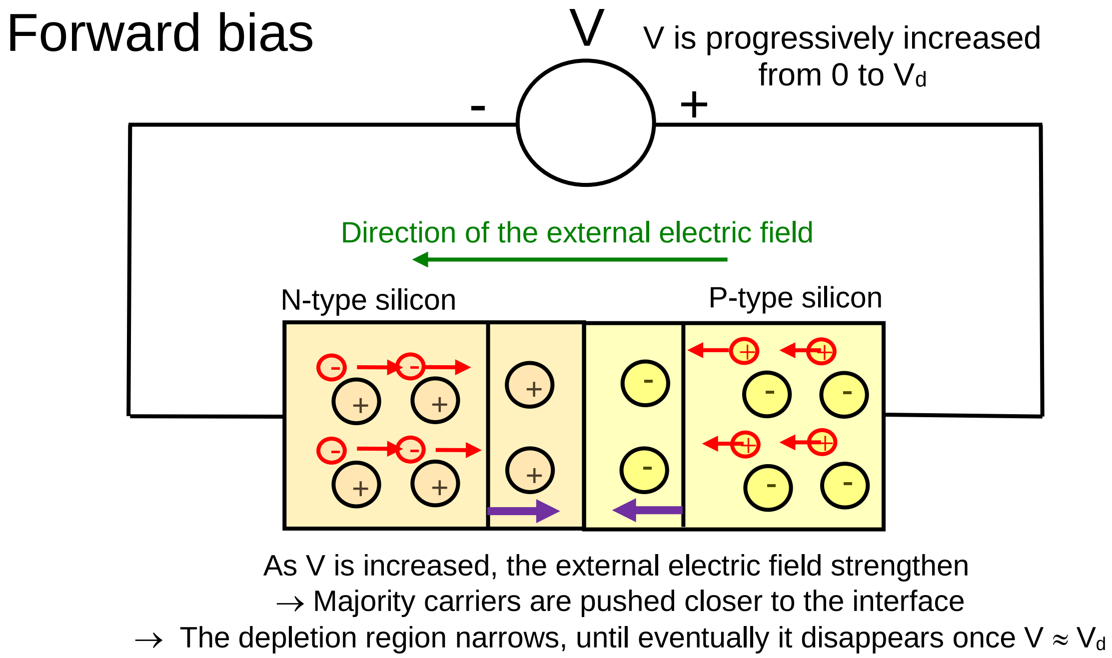
3. Forward Bias

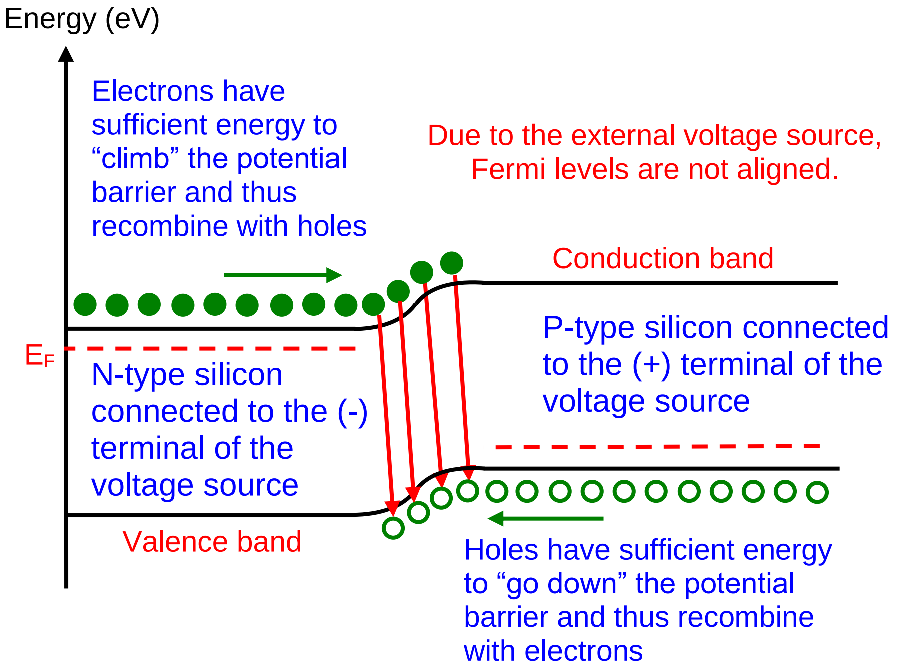
Assume now that the applied voltage polarities are reversed. The positive voltage applies to the P-type material. This is known as *forward bias*.

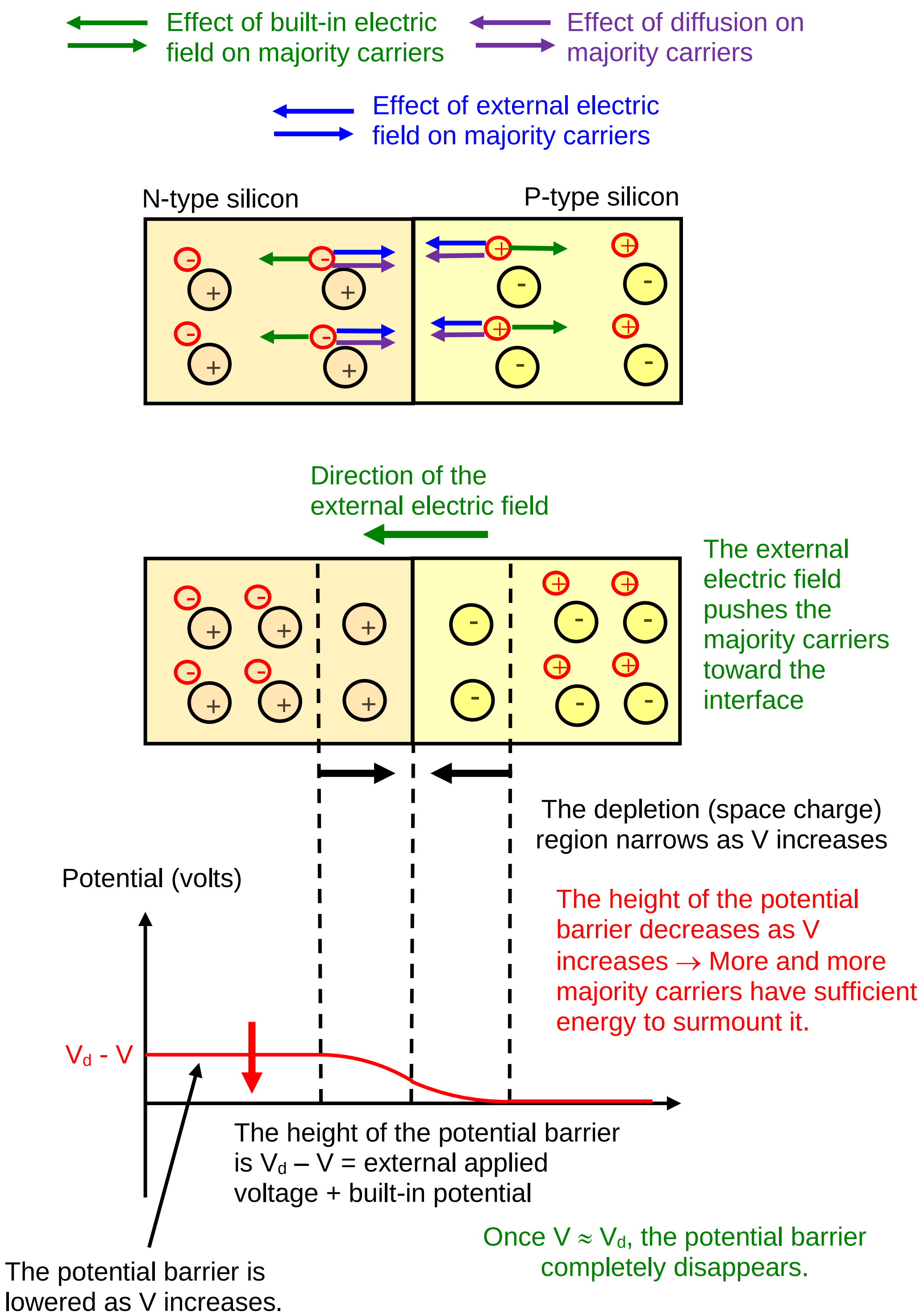
An electrical field is established across the PN junction due the voltage applied across it. Under forward bias, the direction of this external electric field is such that the majority carriers are pushed toward the PN interface. This has the effect of reducing the width of the depletion region.

Once the applied voltage V has become large enough to make the depletion region sufficiently thin, i.e. once the value of V becomes approximately equal to the built-in voltage V_d of the PN junction, a significant number of majority carriers are finally able to cross the junction into the opposite ends of the crystal. This is the condition of forward bias.

Consider the flow of electrons across the junction. The forward bias causes a force on the electrons pushing them from the N side toward the P side.



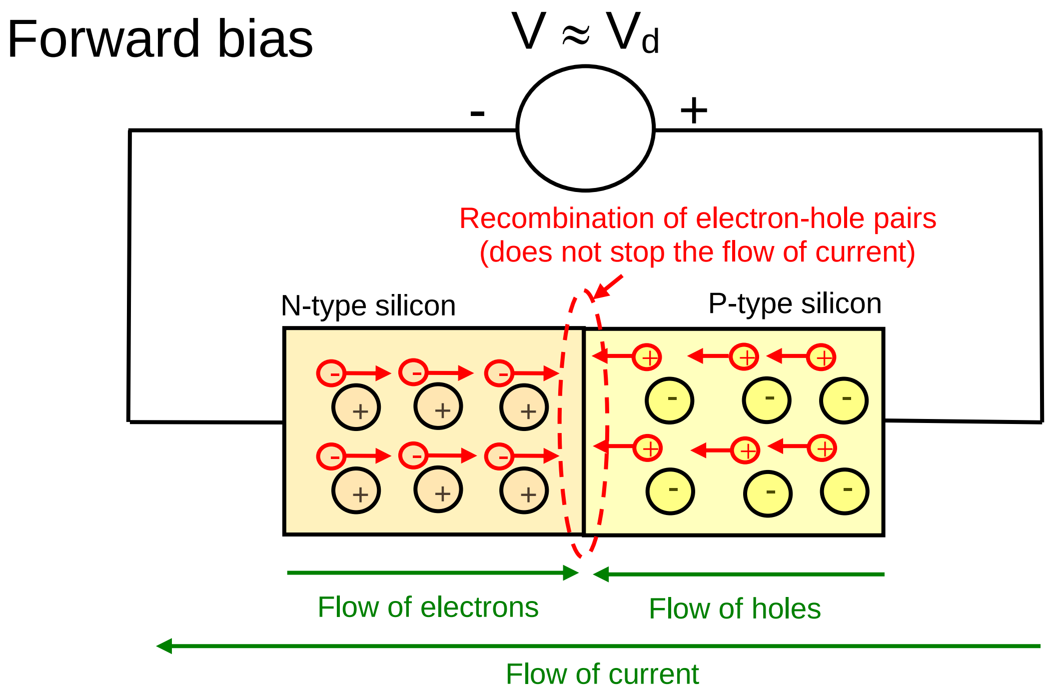




With sufficient forward bias ($V \approx V_d$), the depletion region is narrow enough that electrons can cross the junction and *inject* into the P-type material. However, they do not continue to flow through the P-type material indefinitely, because it is energetically favourable for them to recombine with holes. The average length an electron travels through the P-type material before recombining is called the *diffusion length*, and it is typically on the order of micro-meters.

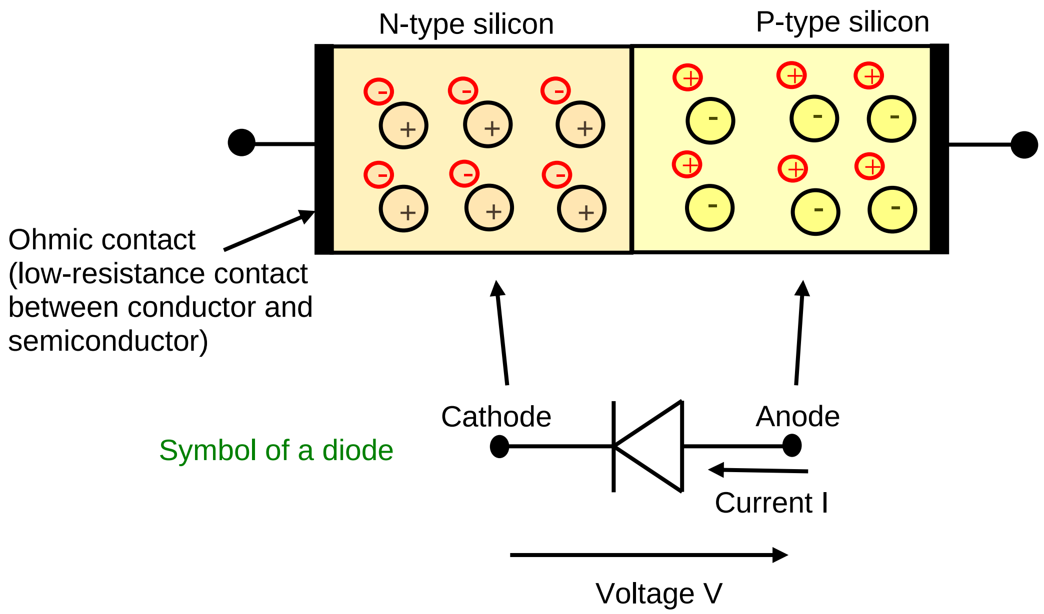
Although the electrons penetrate only a short distance into the P-type material, the electric current continues uninterrupted because holes (the majority carriers) begin to flow in the opposite direction. The flow of holes from the P-type region into the N-type region is exactly analogous to the flow of electrons from N to P (electrons and holes swap roles and the signs of all currents and voltages are reversed).

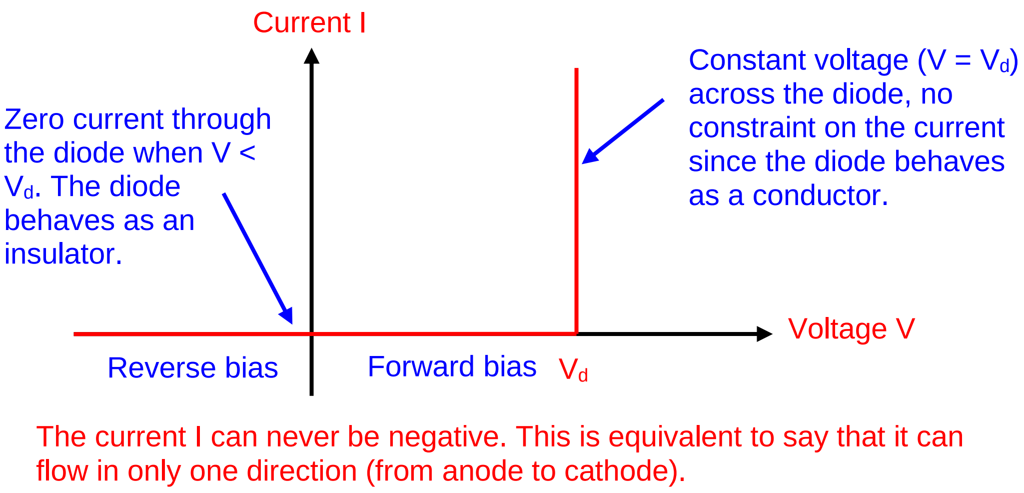
Therefore, the macroscopic picture of the current flow through the diode involves electrons flowing through the N-type region toward the junction, holes flowing through the P-type region in the opposite direction toward the junction, and the two species of carriers constantly recombining in the vicinity of the junction.



The electrons and holes travel in opposite directions, but they also have opposite charges, so the overall current is in the same direction on both sides of the diode.

The conclusion is that an electric current can flow through the PN junction in the forward direction as long as the applied voltage V is approximately equal to the built-in voltage V_d. When this happens, the resistance to the flow of current is actually negligible and the diode somewhat behaves as a conductor. The term “somewhat” is important here because a conductor material does not introduce a constant voltage drop of V_d when a current flows through it.



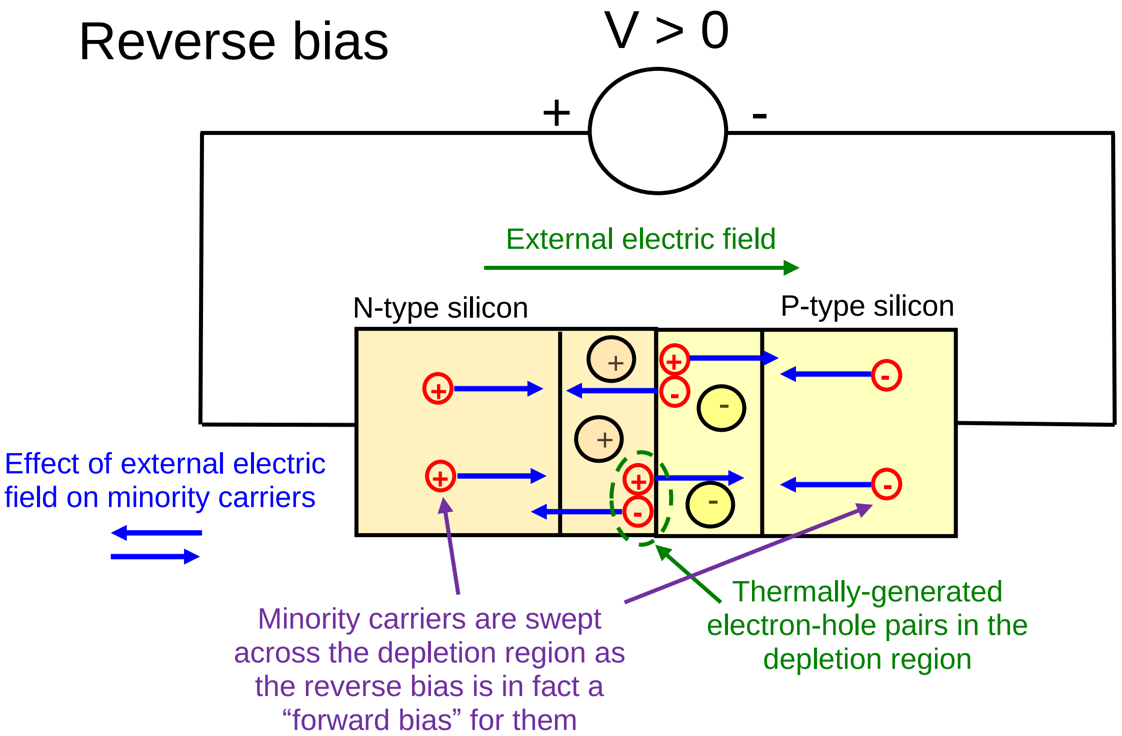


Finally, we can state that, in a PN junction, the current can flow in only one direction (forward direction). This is the basic property of a semiconductor diode which is the electronic component designed using a PN junction.

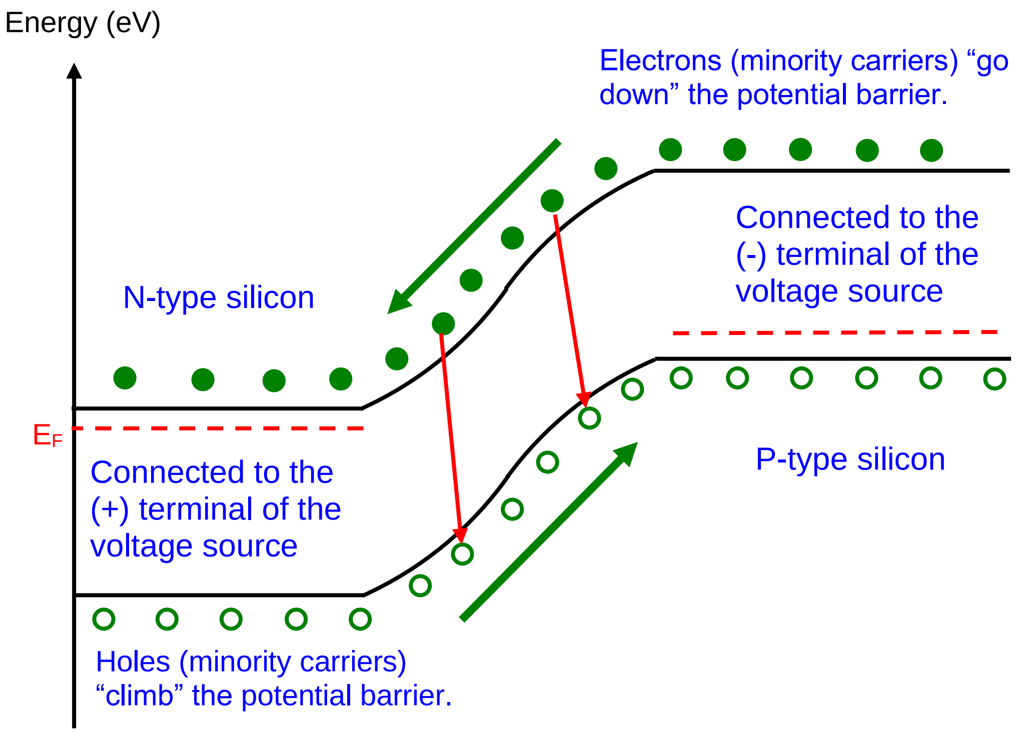
It is worthwhile mentioning that a small current (in practice, < 1 nA) does actually flow through the diode under reverse bias. This current, called *saturation current*, is a combination of the current caused by thermal generation of electron-hole pairs within the depletion region and the diffusion current due to minority carriers in the N and P regions diffusing across the depletion region. The electric field that is present under reverse bias sweeps these minority carriers out of or through the depletion region.

Although the saturation current is voltage independent, it does depend on temperature since both the current contributions depend on thermally stimulated carriers. It is also proportional to the area of the diode, and a function of the doping levels and width of the neutral regions. For typical silicon PN junctions, the saturation current is nominally in the range of 10^{-17} A/μm^2.

Electrons in the P region which diffuse into the depletion region are swept toward the N region by the electric field. Likewise, holes in the N region which diffuse into the depletion region are swept to the P region by the electric field. This diffusion current is in the reverse bias direction.



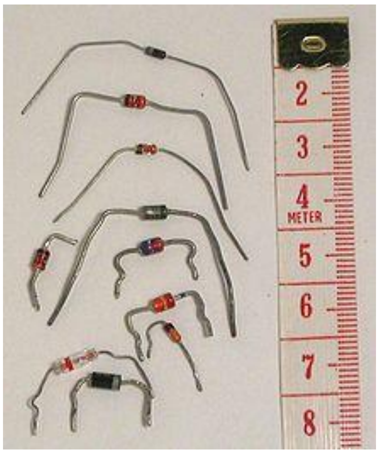
However, there are so few minority carriers in both N-type and P-type regions that this current is very small (smaller than 1 nA) and can often be ignored.



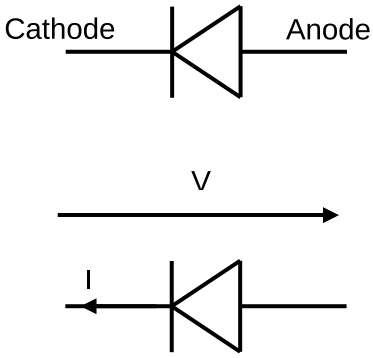
However, the fact that, unlike majority carriers, minority carriers are able to cross a depletion region is something to be remembered as this point is crucial to the understanding of more advanced semiconductor devices such as bipolar and field-effect transistors.

4. Current-Voltage Characteristic of a Diode

A more detailed analysis of the PN junction would provide us with an actual expression, namely the Shockley equation, linking the current flowing through the diode and the voltage across it.



Symbol of a diode



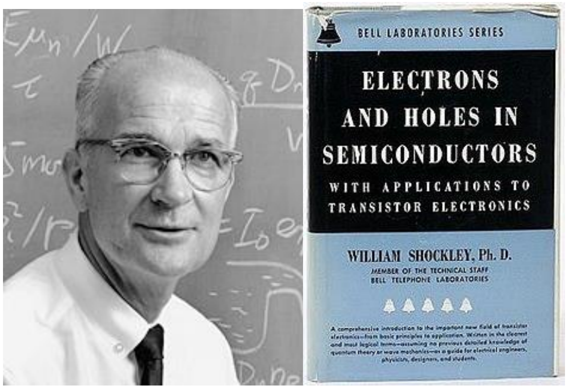
The Shockley diode equation, named after transistor co-inventor William Shockley, gives the current–voltage characteristic of a diode in either forward or reverse bias. The equation is given by

$$I = I_S\left(\exp\left\{\frac{qV}{\eta kT}\right\}-1\right) = I_S\left(\exp\left\{\frac{V}{\eta V_T}\right\}-1\right) \approx I_S \exp\left\{\frac{V}{\eta V_T}\right\},$$

where
- I_S: Saturation current of the diode (in the range from 10^{-8} to 10^{-16} A, typically);
- η: Emission coefficient. This is an empirical constant that varies from 1 to 2 depending on the fabrication process and semiconductor material and in many cases is assumed to be approximately equal to the unit (and thus omitted) for simplicity sake.
- q: Electron charge ($\approx 1.602\times10^{-19}$ C);
- T: Temperature in degrees Kelvin;
- k: Boltzmann's constant ($\approx 1.38\times10^{-23}$ J/K);
- V_T: Thermal voltage defined as $V_T = \frac{kT}{q}$.

At room temperature (T = 300 K), the thermal voltage V_T is equal to

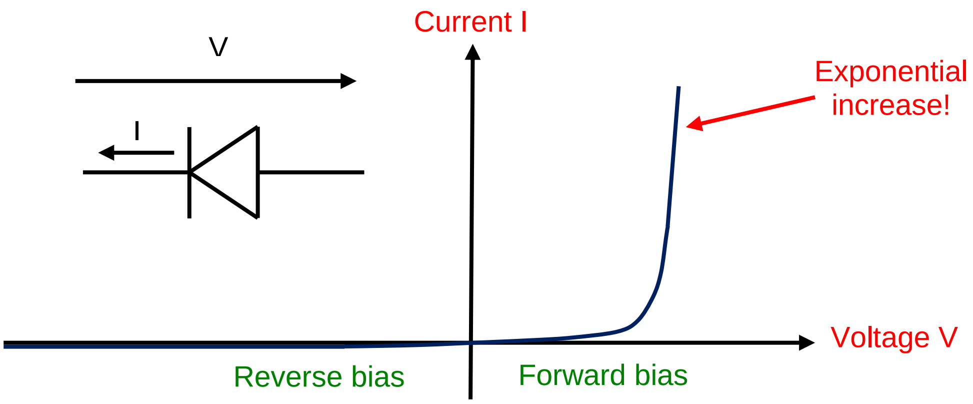
$$V_T = \frac{kT}{q} \approx \frac{1.38\times10^{-23}\ \mathrm{J/K}\times 300\ \mathrm{K}}{1.602\times10^{-19}\ \mathrm{C}} \approx 25\ \mathrm{mV}.$$

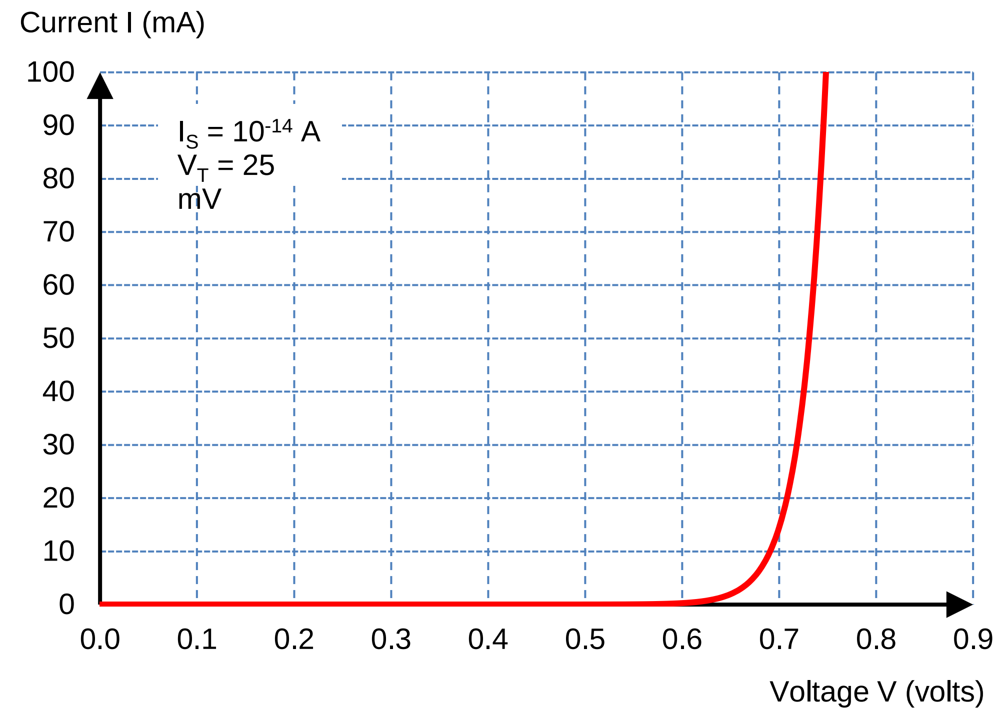


William Bradford Shockley (1910 - 1989)

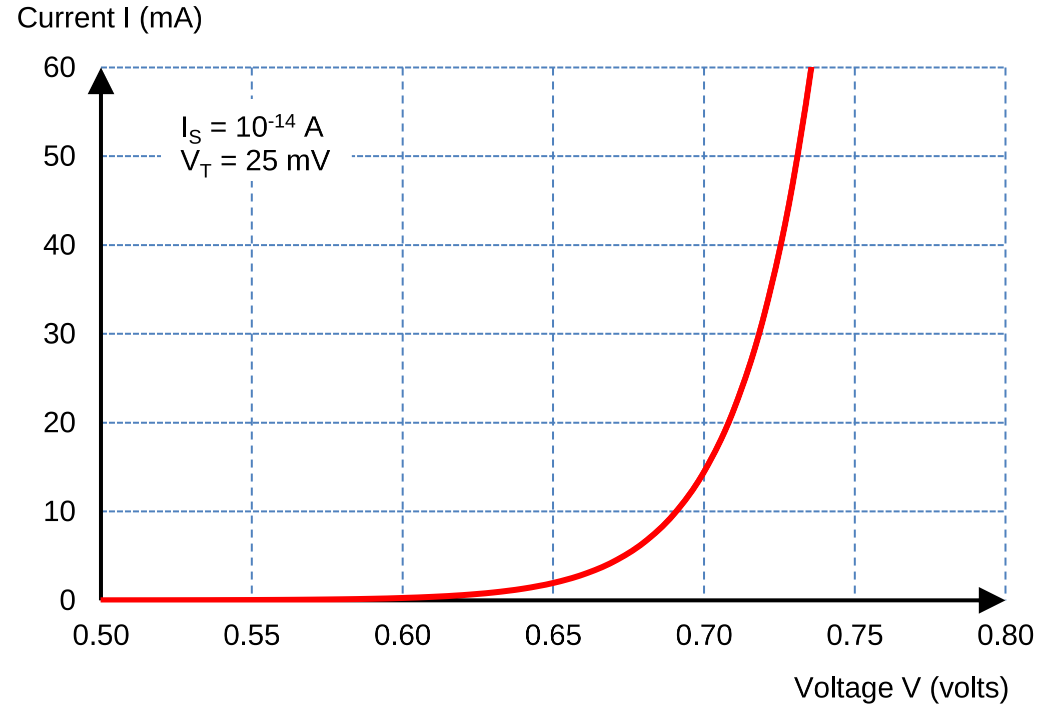
The Shockley equation was first presented in a 558-page treatise entitled "Electrons and Holes in Semiconductors" published by Shockley in 1950.

The Shockley expression means that the current flowing through a diode varies exponentially with the applied voltage.

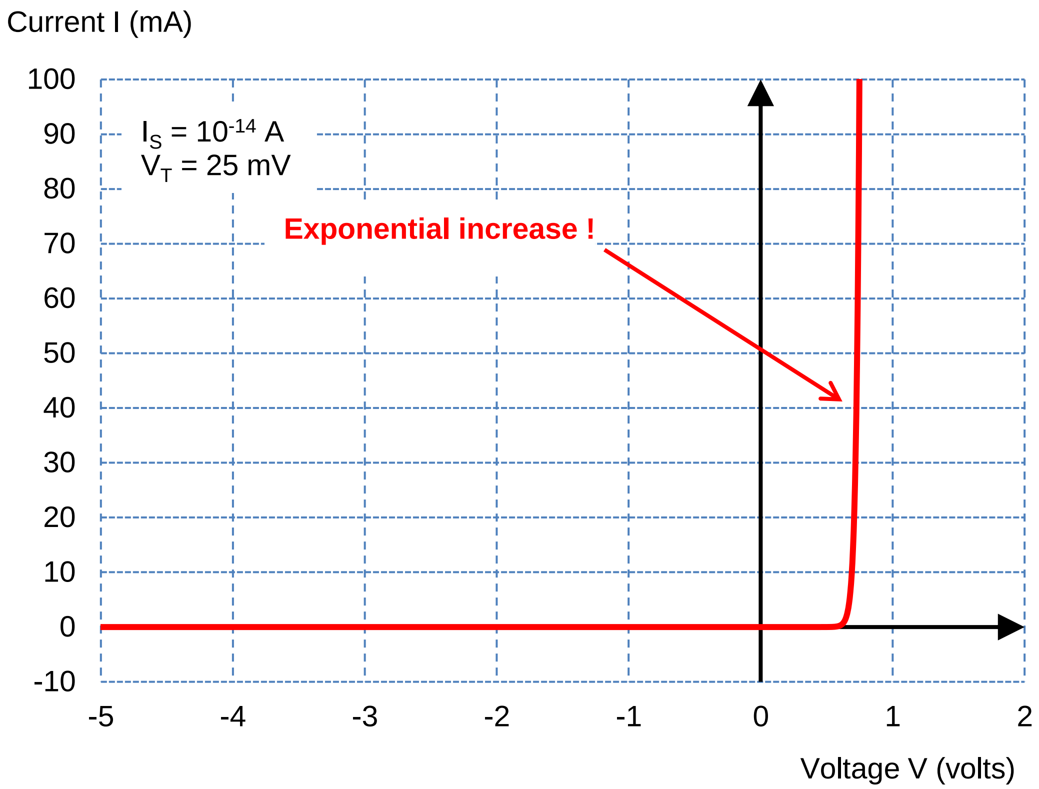




Current I (mA)
100
90
80
70
60
50
40
30
20
10
0
$I_S = 10^{-14}$ A
$V_T = 25$
mV
0.0
0.1
0.2
0.3
0.4
0.5
0.6
0.7
0.8
0.9
Voltage V (volts)

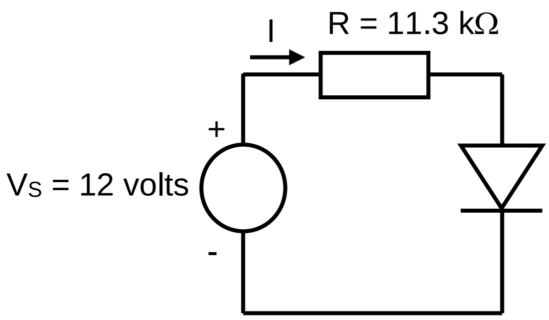


Current I (mA)
60
50
40
30
20
10
0
$I_S = 10^{-14}$ A
$V_T = 25$ mV
0.50
0.55
0.60
0.65
0.70
0.75
0.80
Voltage V (volts)



The Shockley expression should be useful for the analysis of diode circuits since it provides us with a direct link between current and voltage, which is similar to what Ohm's law does for a resistance.

For instance, consider the simple circuit depicted below. It is composed of a DC voltage source V_S = 12 volts, a resistance R = 11.3 kΩ, and a silicon diode.



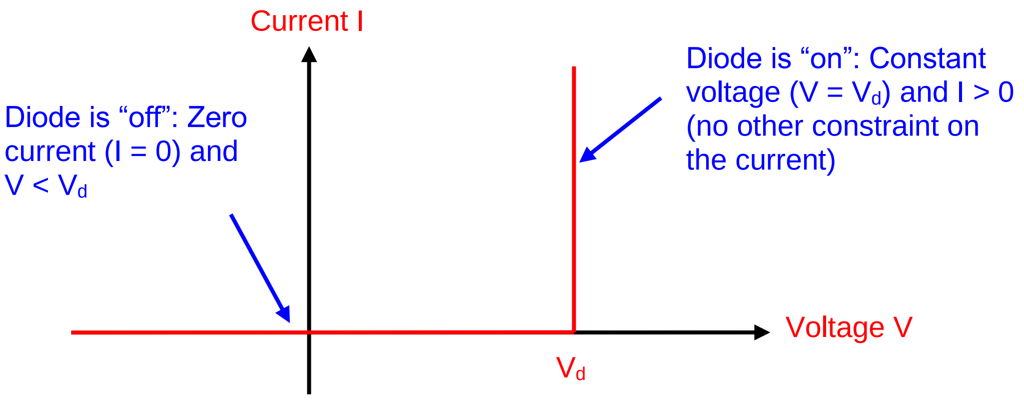
What is the expression and the value of the current I flowing though this circuit?

At first glance, it appears that the answer to this question can be obtained by using the following equation:

$$I \approx I_S \exp\left\{\frac{V_S - R \cdot I}{V_T}\right\}.$$

This is an equation with the current I as the only unknown quantity (assuming of course that the values of the diode parameters I_S and V_T are known). Yet, we are not able to proceed any further because we cannot extract an expression of the current I out of this expression.

We thus conclude that, for the manual analysis of (almost) all circuits containing diodes, Shockley equation is actually too complicated to handle. Electronic engineers deal with this problem by simplifying things and using the much simpler model of the diode given below. This model is suitable for the manual analysis of most diode circuits.



The constant V_d is called *threshold voltage* (or *forward voltage drop*) of the diode. We have $V_d \approx 0.7$ volt for a silicon diode, $V_d \approx 0.25$ - 0.30 volt for a germanium diode, $V_d \approx 1$ volt for a gallium-arsenide diode, and $V_d \approx 0.2$ – 0.4 volt for an aluminium-silicon diode (a metal – semiconductor junction is known in electronics as a *Schottky diode*).

In this model, the current is zero for any voltage below the threshold voltage V_d. In effect, the diode is viewed as a switch which is open when we apply low or negative voltages across it but which closes when we apply a voltage equal to V_d across it. It is important to understand that, with this model, it is strictly impossible to get a voltage larger than V_d across the diode.

Now, with this new simplified diode model, let us consider once again the simple diode circuit introduced earlier and attempt to find the expression of the current I flowing through it. We can write the following general expression:

$$V_S = V + R \cdot I,$$

where V denotes the voltage across the diode.

There are two possibilities at this stage: the diode is either off or on. Since we do not know more than that for now, the best thing is simply to consider both modes of operation and see what the result is in each case.

- 1st possible mode of operation: The diode is off. In this case, we have $V < V_d$ and $I = 0$. The general expression then leads to $V_S = V + R \cdot I = V_d$. This result does not make sense since $V_S = 12$ volts and $V_d \approx 0.7$ volt. We conclude that the diode cannot be off.

- 2nd possible mode of operation: The diode is on. In this case, we have $V = V_d$ and $I > 0$. The general expression then leads to $V_S = V + R \cdot I = V_d + R \cdot I$. There is nothing wrong with this result since $I > 0$ and $V_S > V_d$. We conclude that the diode can be on.

By considering both possibilities, the only conclusion that can be reached is that the diode is on, and the current flowing through the circuit is then given by

$$I = \frac{V_S - V_d}{R} \approx \frac{12\,V - 0.7\,V}{11.3\,k\Omega} = 1\,mA\,.$$

The simplified diode model has allowed us to answer the question and this is clearly a very useful tool to perform the manual analysis of most diode circuits.

5. Effect of Temperature on the Current-Voltage Characteristic of a Diode

The current–voltage characteristic of a diode depends on the temperature T for two reasons:

- The thermal voltage $V_T = \dfrac{kT}{q}$ is proportional to the temperature;

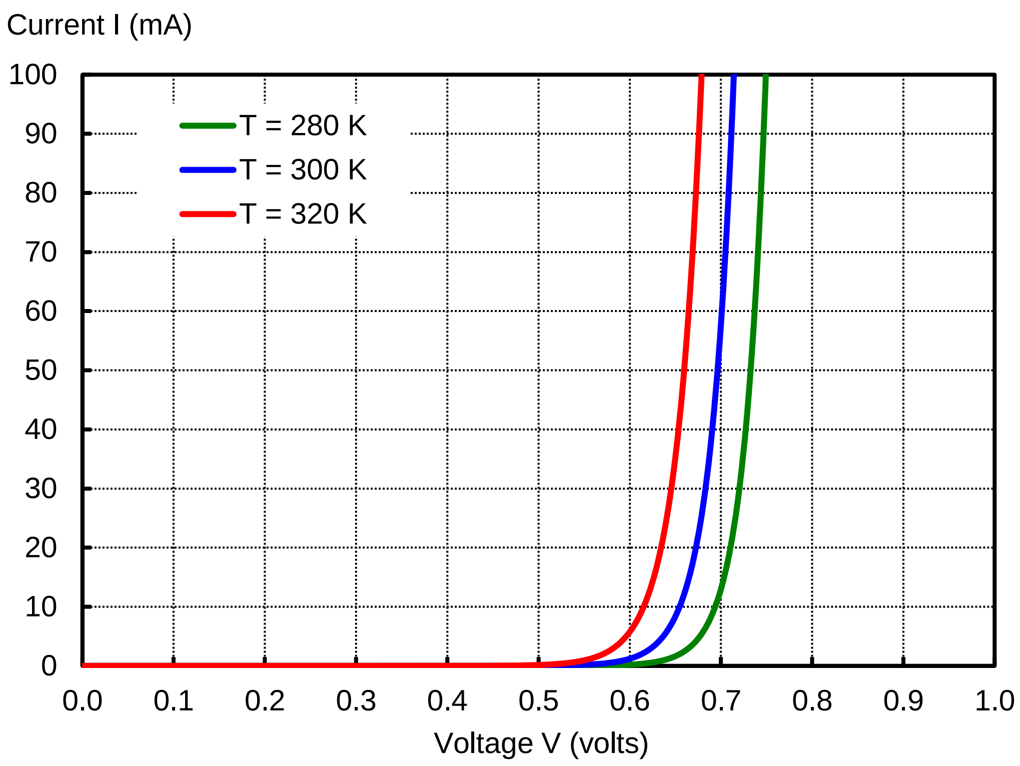
- The saturation current I_S is a strong function of temperature. In fact, we could show that

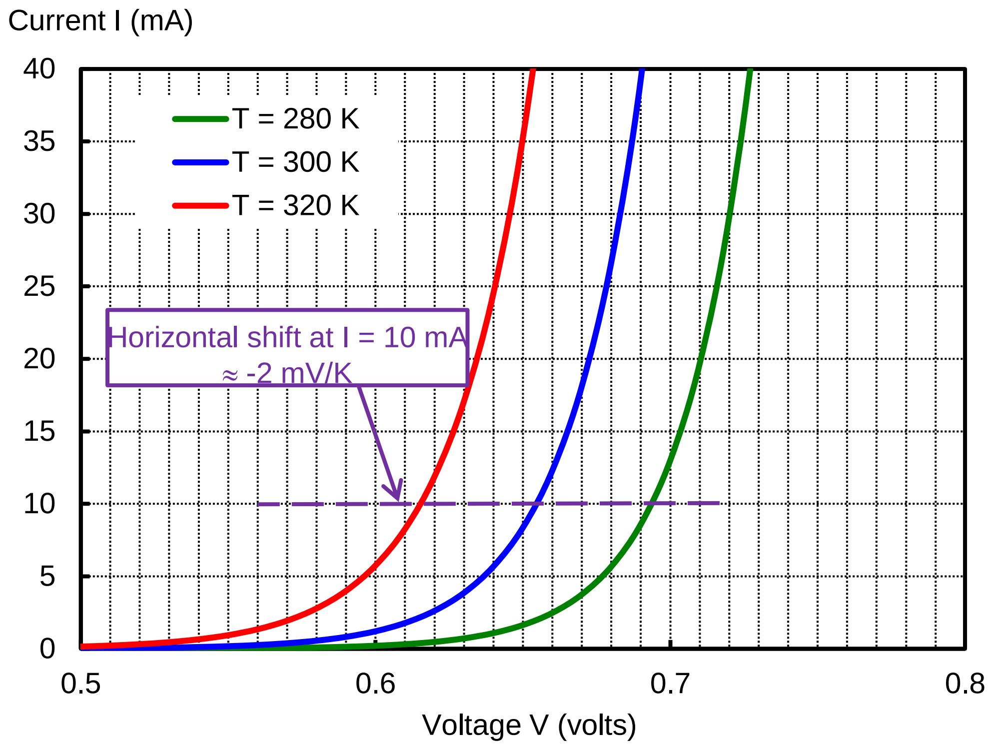
$$I_S = C \cdot T^3 \cdot \exp\left(-\frac{E_{g,0}}{kT}\right),$$

where C is a constant, $E_{g,0}$ is the zero-temperature band gap ($E_{g,0} \approx 1.17$ eV for silicon, 0.74 eV for germanium, 1.52 eV for gallium arsenide), and k denotes the Boltzmann constant expressed here in $eV.K^{-1}$ ($k \approx 8.62 \times 10^{-5}$ $eV.K^{-1}$).

We have plotted below the current-voltage characteristics obtained at three different temperatures (T = 280, 300, and 320 K). Here, the constant C has been assumed to be approximately equal to 0.141 $A.K^{-2}$ so that $I_S \approx 10^{-13}$ A at T = 300 K (room temperature).

It is observed that, for given current value of 10 mA, the measured shift in voltage across the diode is approximately equal to -2 $mV.K^{-1}$. In other words, the voltage decreases by about 2 mV for each additional degree kelvin.





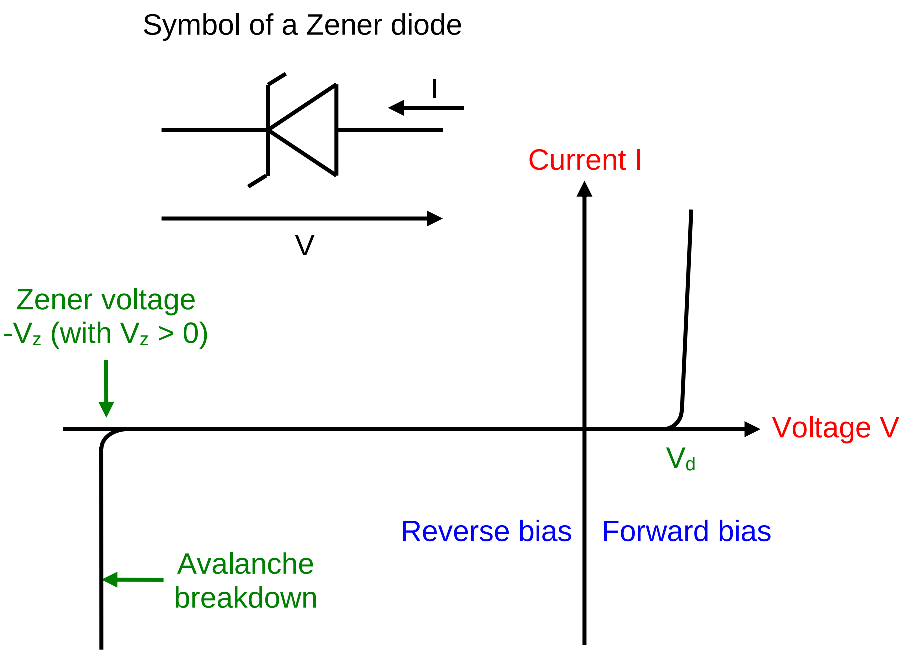
Such temperature dependence can be an issue in some applications and must thus be dealt with in the design of actual semiconductor circuits. On the other hand, it is a very interesting feature that can be exploited for designing temperature sensors.

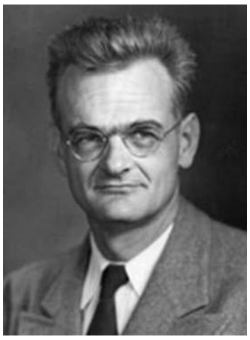
6. Zener Diode

With the application of sufficient reverse voltage, any diode experiences a breakdown and conducts current in the reverse direction. Electrons which break free under the influence of the applied electric field can be accelerated enough that they can knock loose other electrons and the subsequent collisions quickly become an avalanche.

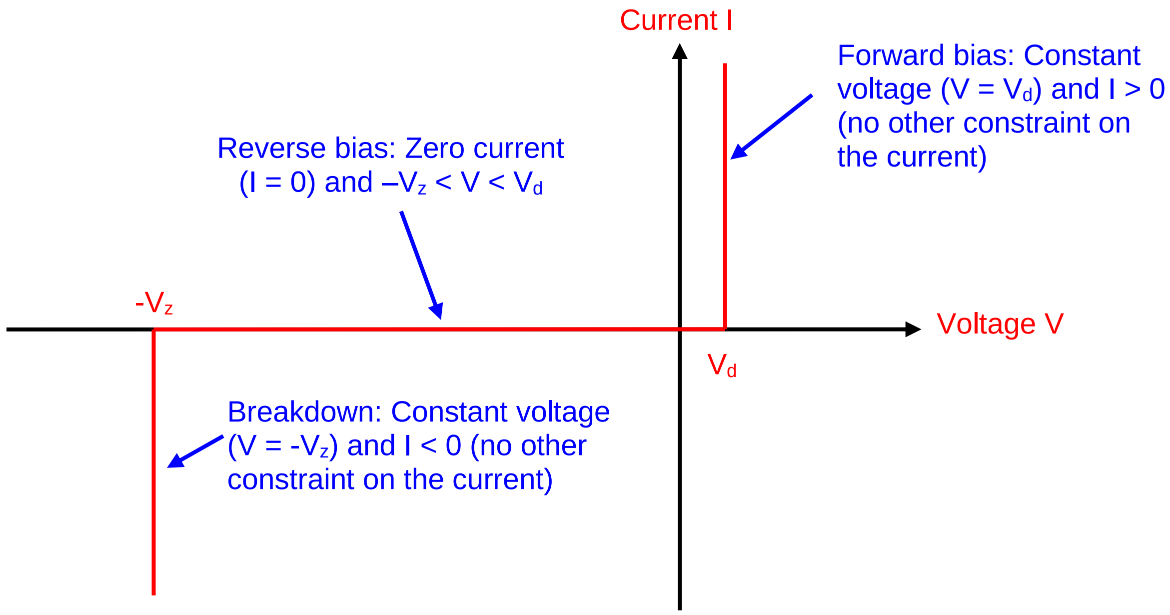
When this process takes place, very small changes in voltage can cause very large changes in current. The breakdown process depends upon the applied electric field. Thus, by changing the thickness of the layer to which the voltage is applied, we can change the value of the breakdown voltage.

This principle is exploited to design a specific type of diodes named Zener diodes, after the American physicist Clarence Zener, which break down at well-defined voltage values ranging from about a few volts to several hundred volts.





Clarence Melvin Zener (1905 - 1993)

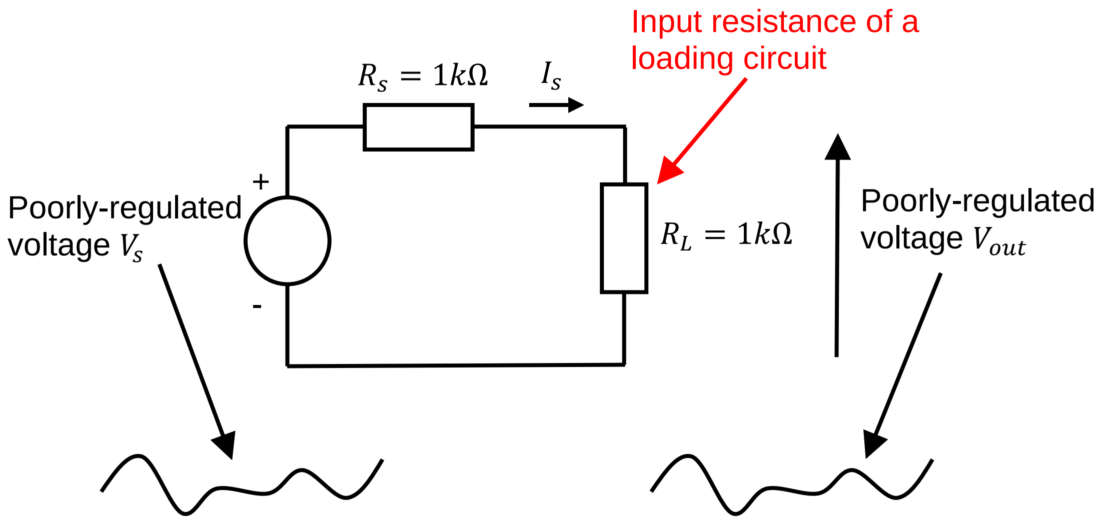


The current-voltage characteristic shown above implies that a Zener diode with a threshold voltage V_d and a breakdown voltage V_Z can be seen as an electronic component with three modes of operation:

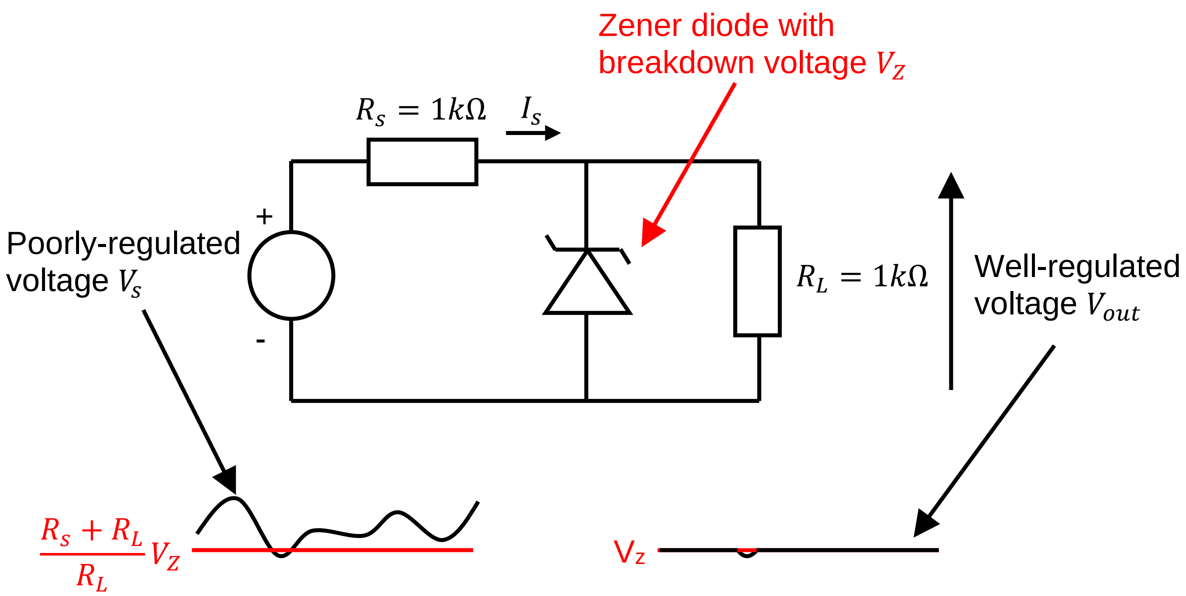
- The Zener diode is in the on (forward-bias) mode if $V = V_d$ and $I > 0$;
- The Zener diode is in the off (reverse-bias) mode if $-V_Z < V < V_d$ and $I = 0$;
- The Zener diode is in the breakdown (avalanche) mode if $V = -V_Z$ and $I < 0$.

The useful feature is that the voltage across the diode remains nearly constant even with large changes in current through the diode. Such diodes find a wide range of applications in electronic circuits. For instance, to illustrate how a Zener diode can be employed for voltage regulation, let us consider the circuit shown below.

This circuit consists a poorly regulated (i.e., unstable) source voltage V_s in series with a source resistance $R_s = 1k\Omega$. This circuit is connected a load circuit modelled by its input resistance $R_L = 1k\Omega$. The voltage delivered to the load circuit is given by $V_{out} = \frac{R_L}{R_s+R_L}V_s = \frac{V_s}{2}$. As the source voltage V_s is unstable, it turns out that the voltage V_{out} is also unstable.

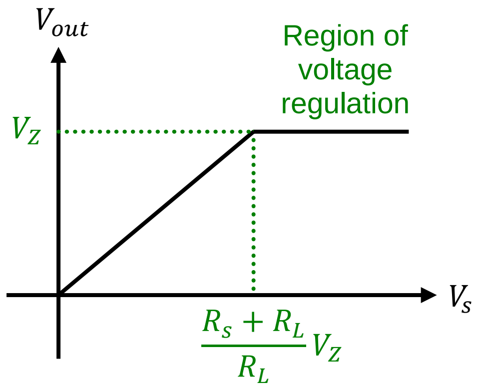


It is possible to stabilise (regulate) the output voltage V_{out} by connecting a Zener diode across the load resistance R_L. If we want V_{out} to be equal at all times to a certain value V_Z, we should employ a Zener diode with a breakdown voltage V_Z and make sure that the source voltage V_s is at all times greater than $\frac{R_s+R_L}{R_L}V_Z$.



By using the three-mode simplified model described previously, we can easily show that, as long as $V_s > \frac{R_s+R_L}{R_L}V_Z$, the Zener diode is in the breakdown mode and we thus have $V_{out} = V_Z$, which implies that the output voltage is then perfectly regulated.

However, as soon as the source voltage V_s falls below the threshold given by $\frac{R_s+R_L}{R_L}V_Z$, the Zener diode turns off and, as a result, we have $V_{out} = \frac{R_L}{R_s+R_L}V_s$, which clearly means that the voltage regulation stops.



This is a very simple technique to achieve voltage regulation. It has however a major drawback: a significant proportion of the current I_s delivered by the source voltage has to flow through the Zener diode instead of flowing into the load circuit, which is a waste of power since, ideally, all the current produced by the voltage source V_s should be made available by the load circuit.

It can be shown (try it as an exercise) that, when the Zener diode is in the breakdown mode, the current flowing into the load circuit is given by $I_L = \frac{V_Z}{R_L}$, whereas the current flowing through the diode is expressed as $I_Z = \frac{V_s-V_Z}{R_s} - \frac{V_Z}{R_L}$.

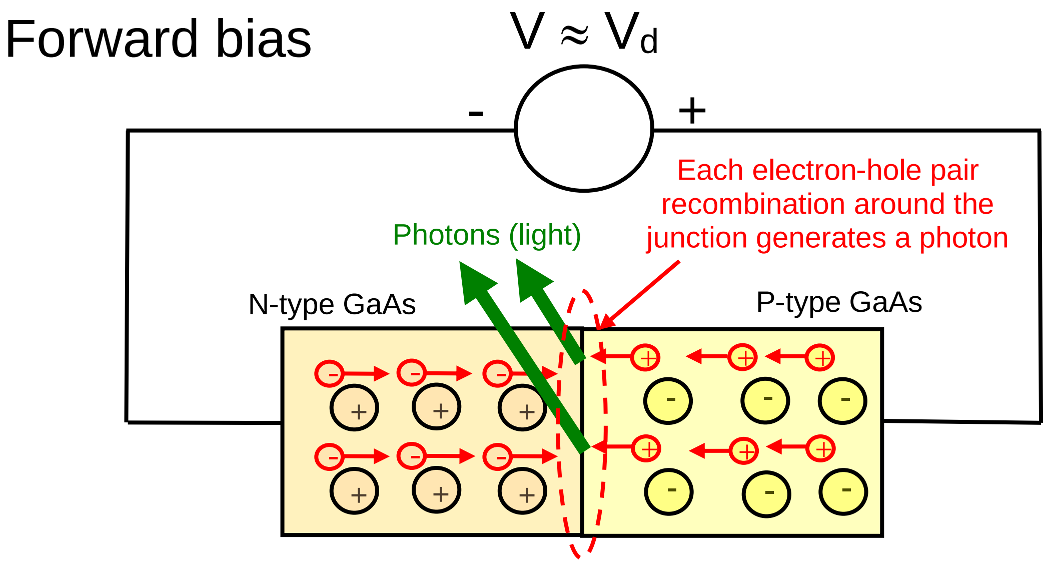
The power efficiency, η, of this voltage regulation method can be defined as $\eta = \frac{I_L}{I_Z+I_L} = \frac{R_s}{R_L} \cdot \frac{V_Z}{V_s-V_Z}$. If $V_s = \frac{R_s+R_L}{R_L}V_Z$, the power efficiency η is equal to 1, i.e. 100%. However, if $V_s = n\frac{R_s+R_L}{R_L}V_Z$, where n designates a real number greater than the unit, we can show that the power efficiency η is then given by $\eta = \frac{100}{2n-1}$ %.

Assuming that $R_s = R_L$ and $V_Z = 10$ volts, the previous result indicates that voltage regulation occurs for any $V_s > 20$ volts. At $V_s = 20$ volts, the power efficiency η is 100%. However, if $V_s = 21$ volts, the power efficiency η falls to approximately 91%, whereas it falls to roughly 83% for $V_s = 22$ volts. The power efficiency falls to only 50% when $V_s = 30$ volts.

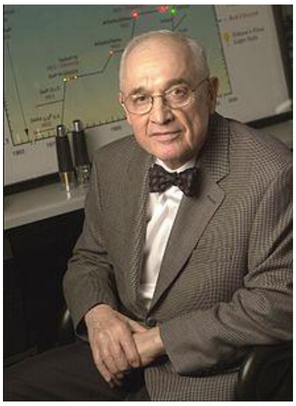
From these results, we conclude that the voltage regulation technique relying on a Zener diode has an acceptable power efficiency provided that the input voltage V_s to be regulated does not rise too much above the regulation threshold given by $V_s = \frac{R_s + R_L}{R_L} V_Z$.

7. Light Emitting Diodes (LED)

Diodes made from a combination of certain elements, such as gallium arsenide, can emit light under forward bias.



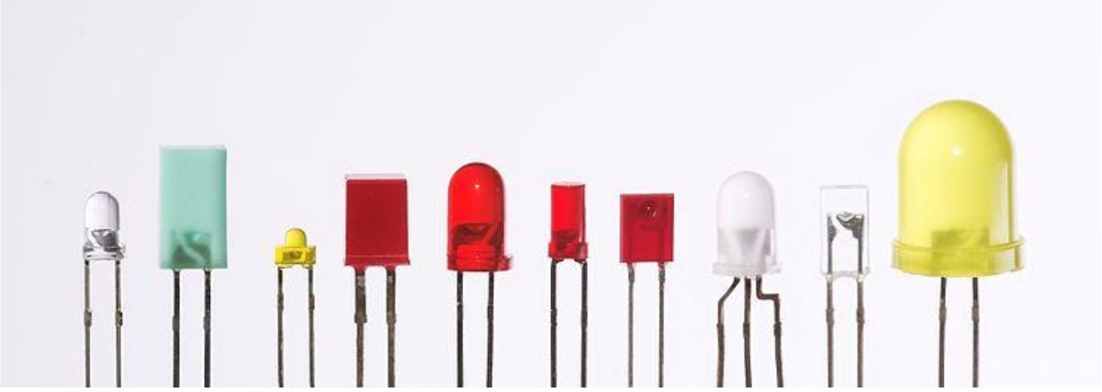
The first practical visible-spectrum (red) LED was developed in 1962 by Nick Holonyak, while working at General Electric Company in Syracuse, New York. LED development began with infrared and red devices made with gallium arsenide.



Nick Holonyak (1928 -)

Advances in materials science have enabled making devices with ever-shorter wavelengths, emitting light in a variety of colours. Some of the currently available colours are red, orange, yellow, green, blue, and infra-red (invisible light at a frequency lower than red).

As in other diodes, current can flow easily from the P-side to the N-side under forward bias, but not under reverse bias. With forward bias, when an electron meets a hole, it falls into a lower energy level (from the conduction energy band to the valence energy band) and releases energy in the form of a photon.



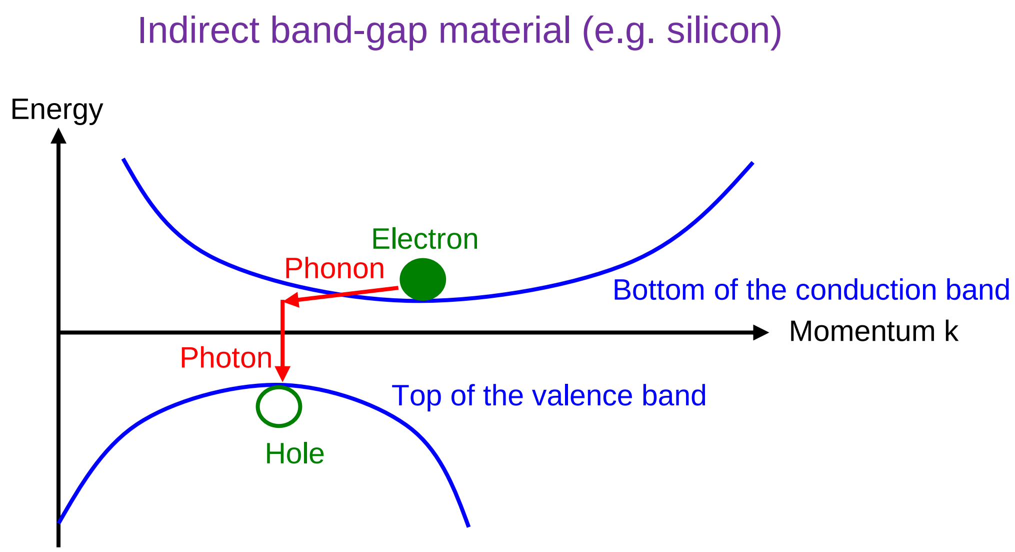
LEDs are produced in a variety of shapes and sizes

The wavelength of the light emitted, and thus its colour, depends on the energy gap between conduction and valence bands of the materials forming the PN junction. In silicon or germanium diodes, the recombination of electrons and holes produces no optical emission because these are indirect band-gap materials. The materials used for the LED have a direct band-gap with energies corresponding to near-infrared, visible, or near-ultraviolet light.

In a semiconductor, if an electron goes from the bottom of the conduction band to the top of the valence band, it may have to undergo a significant change both in momentum and energy.

Whenever something in physics changes state, one must conserve not only energy, but also crystal momentum. To conserve both energy and momentum, the electron-hole recombination can produce a photon (particle of light with little momentum compared to its energy) and/or a phonon (particle associated with lattice vibrations, i.e. heat, with small energy and large momentum compared to that of a photon).

For an indirect band-gap semiconductor, such as silicon and germanium, the bottom of the conduction band is not vertically aligned to the top of the valence band. In this case, when an electron goes from the conduction band to the valence band, both energy and momentum can only be conserved if the transition creates both a photon and a phonon.

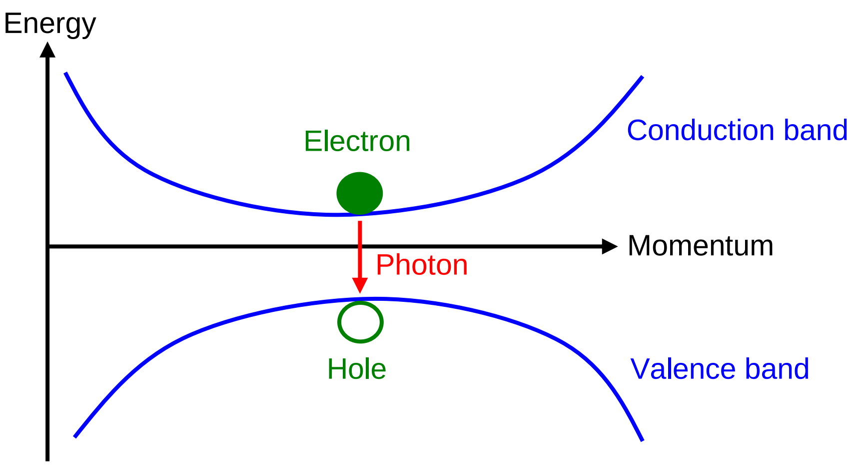


Since the transition process in an indirect band-gap semiconductor involves a phonon in addition to the electron and photon, the probability of having an interaction taking place that involves all three particles will be low, thus resulting in a small emission of photons. This explains why indirect band-gap materials have low electron-photon conversion efficiency.

In direct band-gap semiconductors, such as gallium arsenide, the bottom of the conduction band is vertically aligned with the top of the valence band. As a result, the transition of an electron from conduction band to valence band involves essentially no change in momentum. Thus, this transition can be made by creating only a photon, without the need for emitting a phonon.

The transition process in a direct band-gap semiconductor only involves an electron and a photon. Hence, the probability of having an electron-photon interaction taking place is higher than in the case of an indirect band-gap semiconductor (where three particles instead of two have to be involved). As a result, one finds that photon emission is maximised in a direct band-gap material, thus resulting in high electron-photon conversion efficiency.

Direct band-gap material (e.g. gallium arsenide)



All other optoelectronic devices such as photodiodes (where we are concerned with light absorption rather than emission), solar cells, and laser diodes are made using direct band-gap semiconductors for the reasons explained above.

LEDs are widely used in optical electronic devices. For example, in telecommunications, they are employed to transmit digital information at very high speed, as pulses of light, through empty space or fibre-optic cable.

8. Schottky Diodes

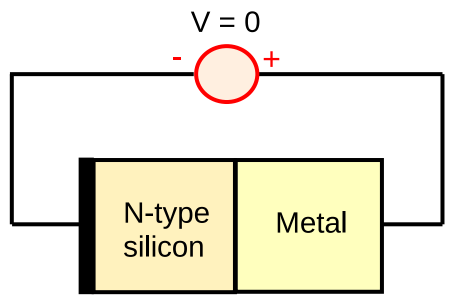
The Schottky diode, named after German physicist Walter H. Schottky, is a metal–semiconductor junction (instead of a semiconductor–semiconductor junction as in conventional diodes). Typical metals used are molybdenum, platinum, chromium or tungsten, whereas the semiconductor is typically N-type silicon. The metal side acts as the anode, and the N-type semiconductor acts as the cathode of the diode, thus meaning that the current can flow from the metal side to the semiconductor side, but not in the opposite direction.

The metal–semiconductor junction creates a Schottky barrier. The Fermi level in an N-type semiconductor is higher than in a metal, meaning that free electrons in an N-type semiconductor have higher energy than those in a metal.

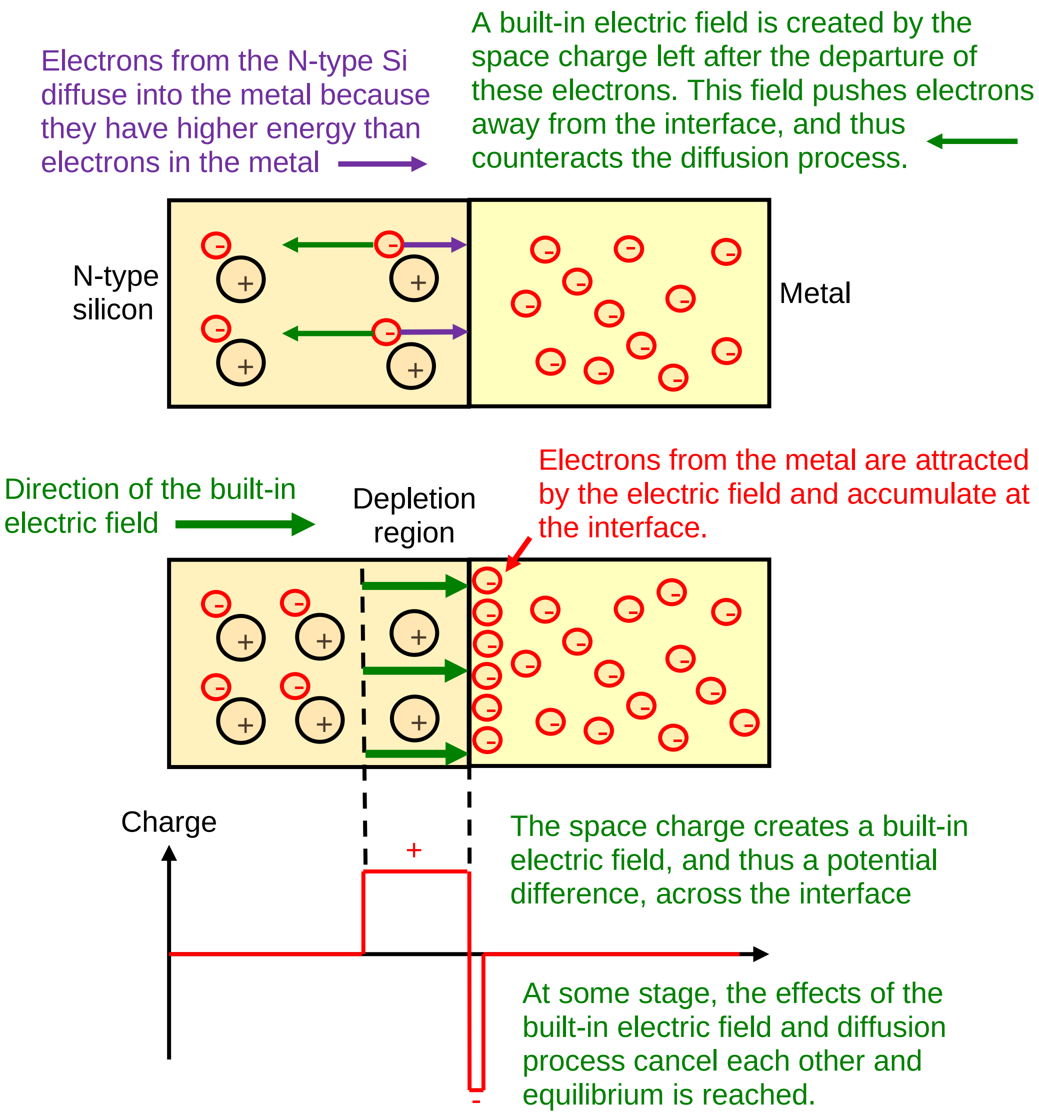
When both materials are put in contact, the high-energy electrons from the N-type semiconductor diffuse into the metal until the Fermi levels are aligned. As a general rule for junctions in solid-state physics, which Fermi level is higher dictates where electrons flow from.

A depletion region made of positive bound charges (space charge) is left near the interface after the departure of the high-energy electrons. The presence of these immobile positive ions generates, across the depletion region, a built-in electric field that counteracts the diffusion process.

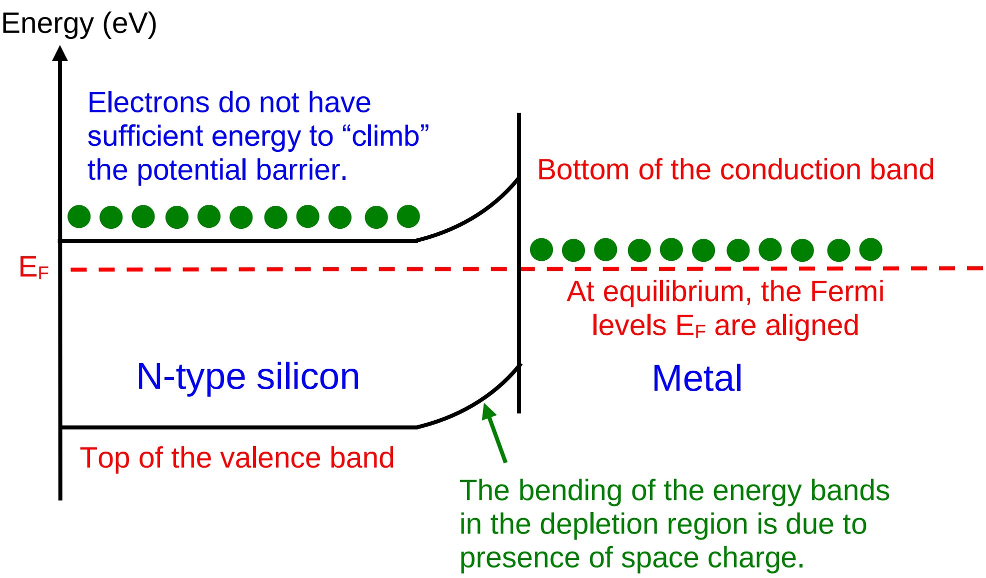
When the depletion region reaches a certain width, the electric field created by the space charge completely cancels the diffusion process, and a state of equilibrium is reached.



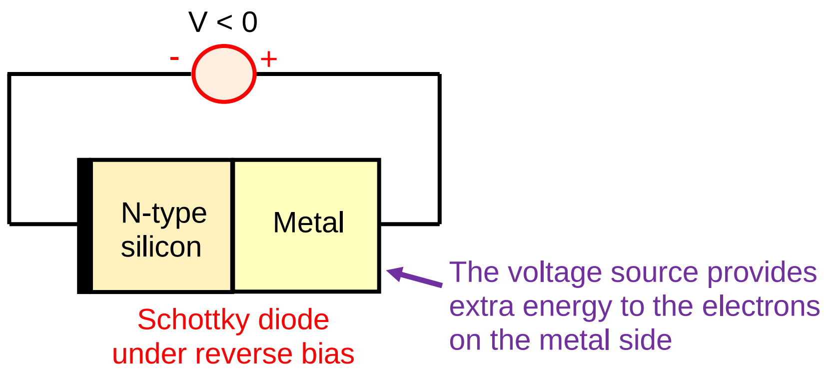
Schottky diode at equilibrium



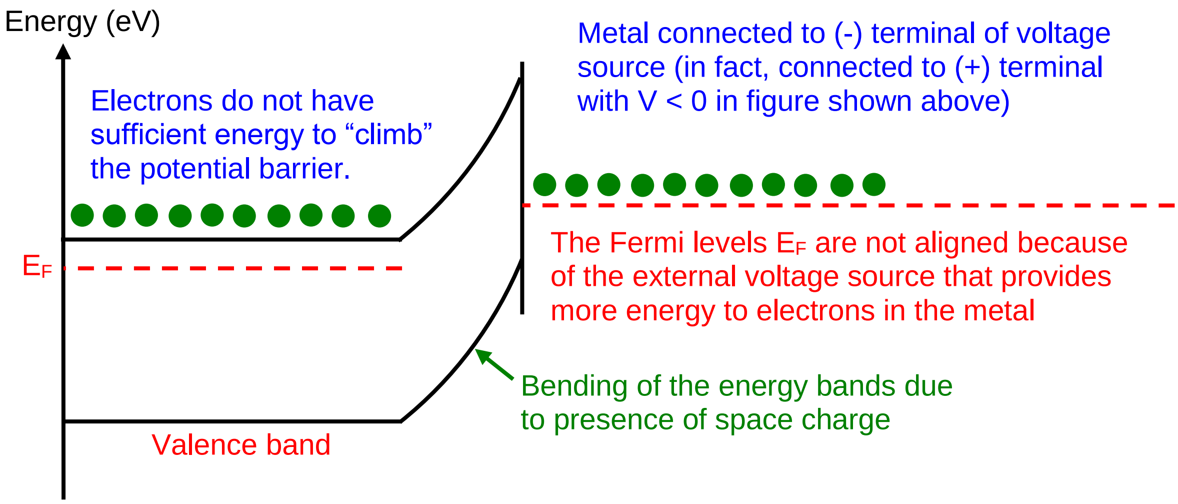
Due to the space charge, the energy bands in the N-type semiconductor are bent upwards, which creates a Schottky barrier that cannot be crossed by electrons from either side. This barrier can thus be viewed as a very high resistance when small voltage biases are applied to it.



A Schottky barrier can be reverse-biased by applying an external voltage across the junction with the highest voltage connected to the N-type semiconductor.

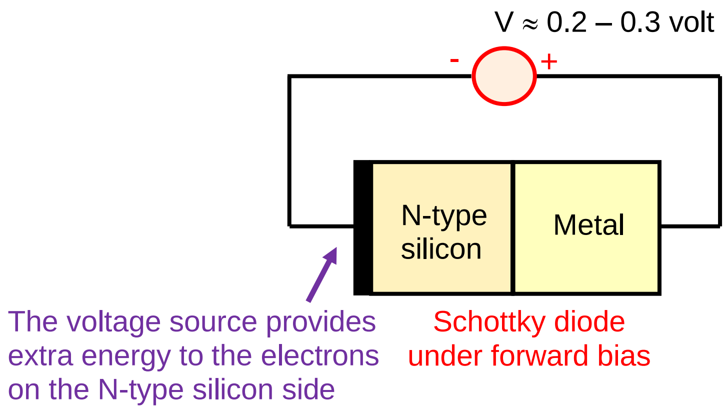


Under reverse bias, there is a small leakage current as some thermally excited electrons in the metal have enough energy to surmount the barrier. This leakage current rises gradually with reverse bias due to a weak barrier lowering. At very high biases, the depletion region breaks down.

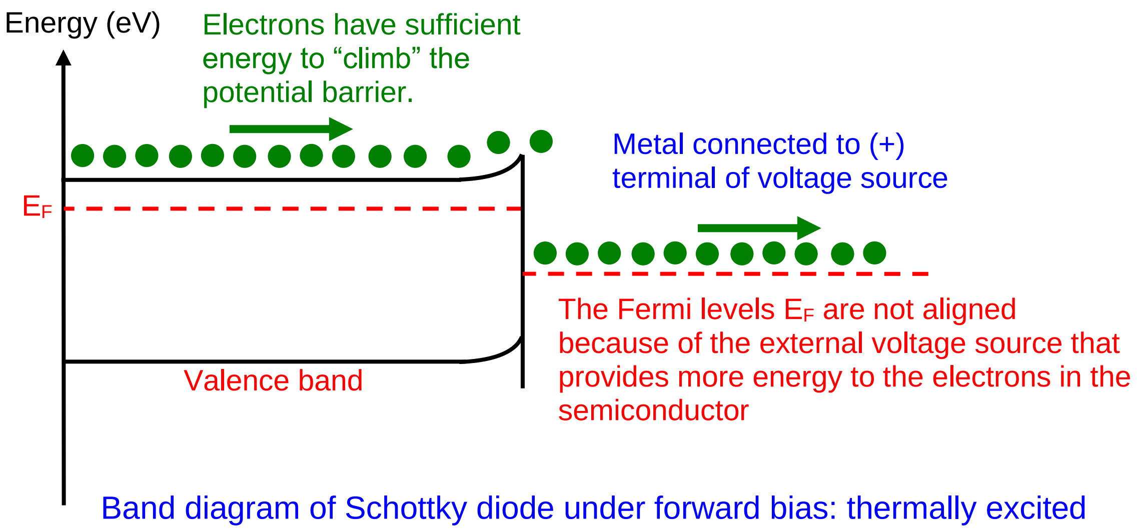


Band diagram of Schottky diode under reverse bias: The barrier is too high for (most) thermally excited electrons to enter the conduction band from the metal → Very small (negligible) current can flow from N-type semiconductor to metal.

A Schottky barrier can also be forward-biased by applying a small external voltage across the junction with the highest voltage connected to the metal.

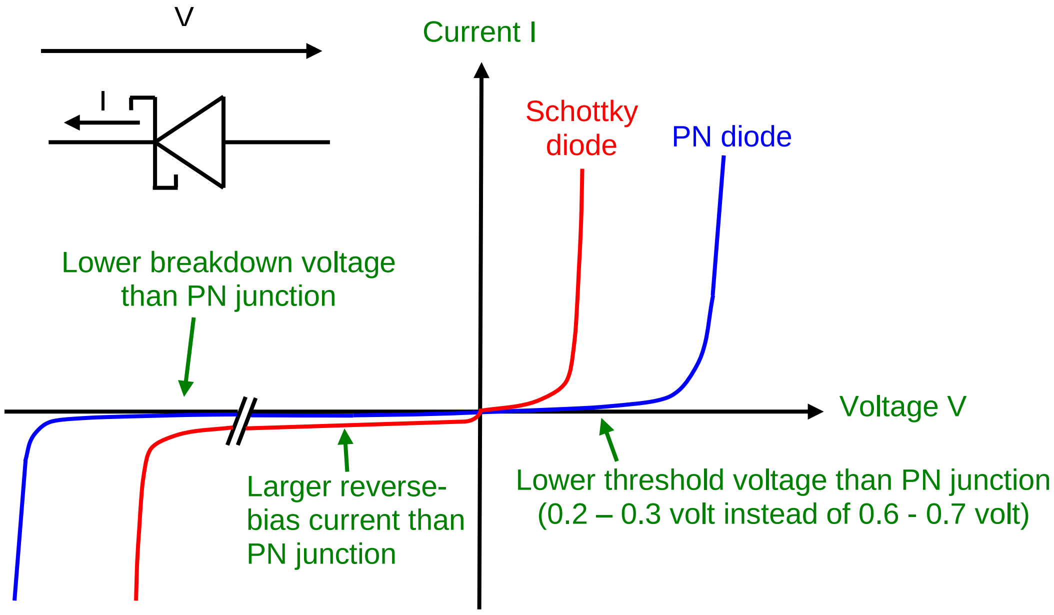


Under forward bias, there are many thermally excited electrons in the semiconductor that are able to pass over the barrier. The passage of these electrons over the barrier into the metal corresponds, as always in electronics, to a current in the opposite direction (from metal to semiconductor). The current rises very rapidly with the biasing voltage.



Band diagram of Schottky diode under forward bias: thermally excited electrons are able to spill from the semiconductor into the metal → A significant current can flow from metal to N-type semiconductor.

The Schottky junction behaves very much like a conventional PN junction, but it has a lower forward voltage drop V_d that typically ranges from 0.2 to 0.3 volt, depending on the type of metal and semiconductor used.

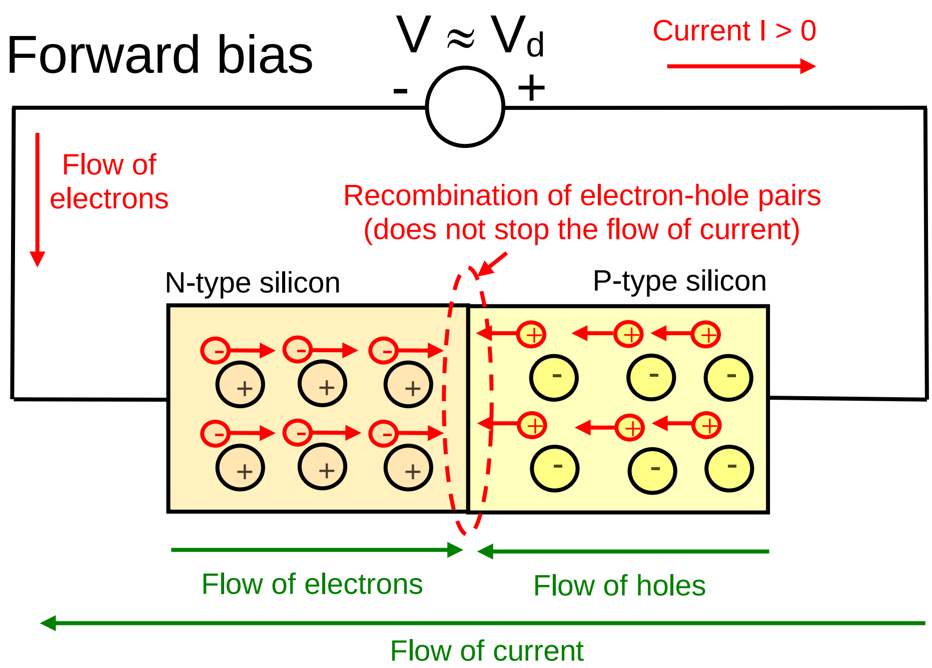


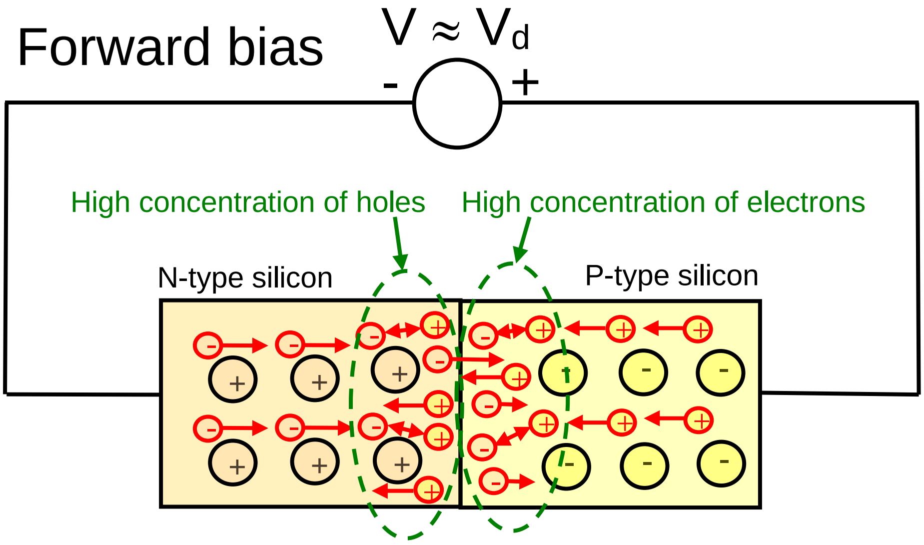
The Schottky junction can also switch on and off much faster than PN junctions, which makes Schottky diodes of great interest for the design of high-speed electronic circuits.

As far as speed is concerned, the most crucial difference between PN and Schottky diodes is the time, known as *reverse recovery time*, needed by the device to switch from the conducting to the non-conducting state.

We remember that, when a PN junction is under forward bias, electrons from the N-type material (majority carriers) cross the junction and inject into the P-type material. They then become minority carriers and start recombining with holes as soon as they move away from the junction to go deeper into P-type territory. The average length an electron travels through the P-type material before recombining (known as *diffusion length*) is typically on the order of micro-meters. Although the electrons penetrate only a short distance into the P-type material, the electric current continues uninterrupted because holes (the majority carriers) flow in the opposite direction.

The flow of holes from the P-type region into the N-type region is exactly analogous to the flow of electrons from N to P (electrons and holes swap roles and the signs of all currents and voltages are reversed).



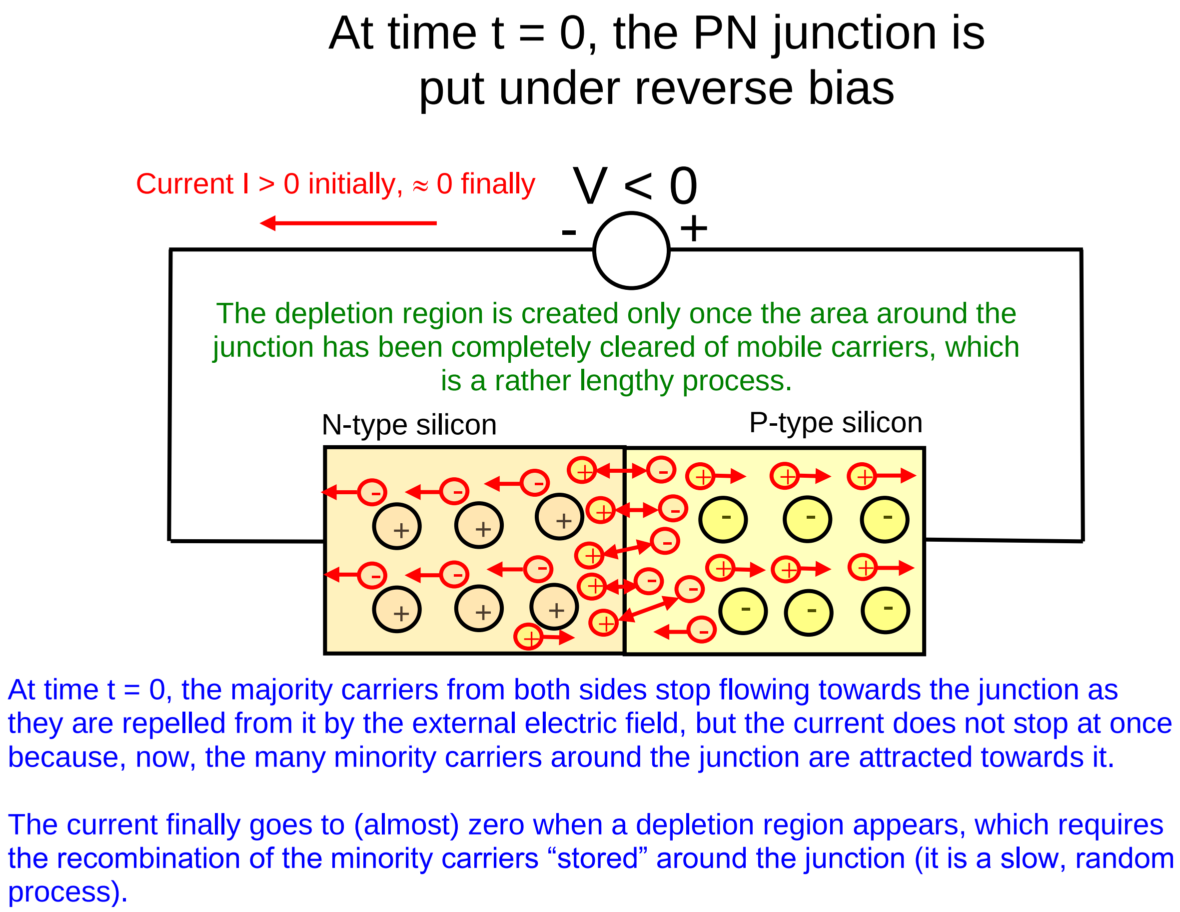


Under forward bias, there is a high concentration of minority carriers around the junction. This concentration decreases as these carriers move away from the junction due to the electron-hole pair recombinations constantly happening during the journey of minority carriers.

If the diode is suddenly put under reverse bias at time t = 0 (the external voltage V is assumed here to become smaller than zero), all charges suddenly start travelling in opposite direction, which is going to ultimately create a depletion region around the junction because the majority carriers move away from the PN junction.

In fact, the only carriers that are attracted towards the junction under reverse bias are the minority carriers. Under regular operation, there are so few of them that the resulting reverse-bias current (basically the saturation current I_s of the diode) is very small and thus negligible in most applications.

However, just after t = 0, there are still plenty of these minority carriers around the junction. This is the reason why the reverse-bias current is actually significant right after t = 0. This current will ultimately go to almost zero after some time, known as the reverse recovery time, as one would expect, but this will happen only once all the minority carriers around the junction have reached their equilibrium (very low) concentrations through electron-hole pair recombinations. Since these recombinations are random processes that happen rather slowly, the reverse recovery time is actually quite significant. In a PN junction, the reverse recovery time can be in the order of several microseconds to less than 100 ns for fast diodes.



As a conclusion, it takes quite a long time for a PN junction to switch from the on mode to the off mode, and this constitutes a major problem for applications which require high switching frequencies.

Schottky diodes do not have this issue as they only rely on an N-type semiconductor, meaning that only the electrons play a significant role in the normal operation of the device. The switching speed is thus not limited by the relatively slow and random recombination of electrons and holes that occurs with PN junctions. The switching time is ~100 ps for the small-signal diodes, and up to tens of nanoseconds for special high-capacity power diodes.

It is often said that the Schottky diode is a "majority carrier" semiconductor device. This means that if the semiconductor is a doped N-type, only the N-type carriers (mobile electrons) play a significant role in normal operation of the device.

When the Schottky diode is suddenly put under reverse bias, the high-energy majority carriers located near the junction are quickly injected into the conduction band of the metal contact on the other side of the diode to become free moving electrons. Therefore, no slow random recombination of electrons and holes carriers is required to create a depletion region on the semiconductor side of the junction, so that this diode can cease conduction faster than an ordinary PN diode.

Finally, it is worth mentioning that, with increased doping of the semiconductor, the width of the depletion region drops. Below a certain width, the majority charge carriers (electrons in the case of an N-type semiconductor) can tunnel through the depletion region. When this happens, the Schottky junction does no longer behave like a PN diode and instead becomes an ohmic contact (recall that an ohmic contact is a contact between a conductor and a semiconductor that behaves like a low-resistance conducting layer). This behaviour can be used for the simultaneous formation of ohmic contacts and diodes, as a diode will form between the metal and lightly doped N-type region, and an ohmic contact will form between the metal and the heavily doped N-type region.

- END -