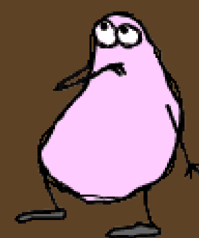
Nobelprize.org

The Official Web Site of the Nobel Prize

THE TRANSISTOR

BACK

Contents



THE TRANSISTOR

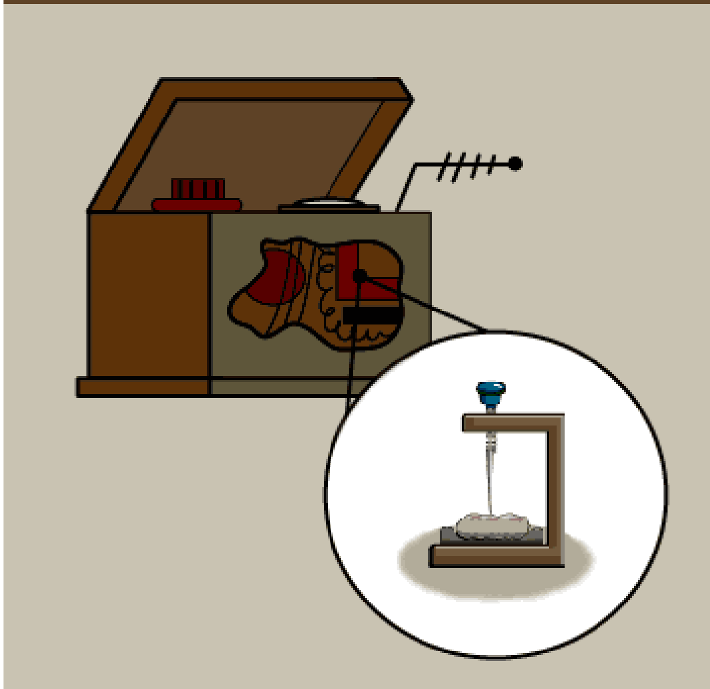
1:39 BACK

The Transistor - Function

The transistor is one of the most important discoveries of the 20th century. It has made it possible to send mankind to the moon, build small yet powerful computers, construct small and efficient hearing aids. These inventions are just a few examples of everyday items containing transistors, the list of such objects can be made very long. The transistor radio was one of the first transistorized items that became a big commercial success. Today, one of the most common uses of the transistor is as a part of the integrated circuit (IC) that is vital for a computer's function. Some even call it "the nerve cell of the information technology" and although the transistor is of great importance, few people know what it is and how it functions.

THE TRANSISTOR

2:39 BACK

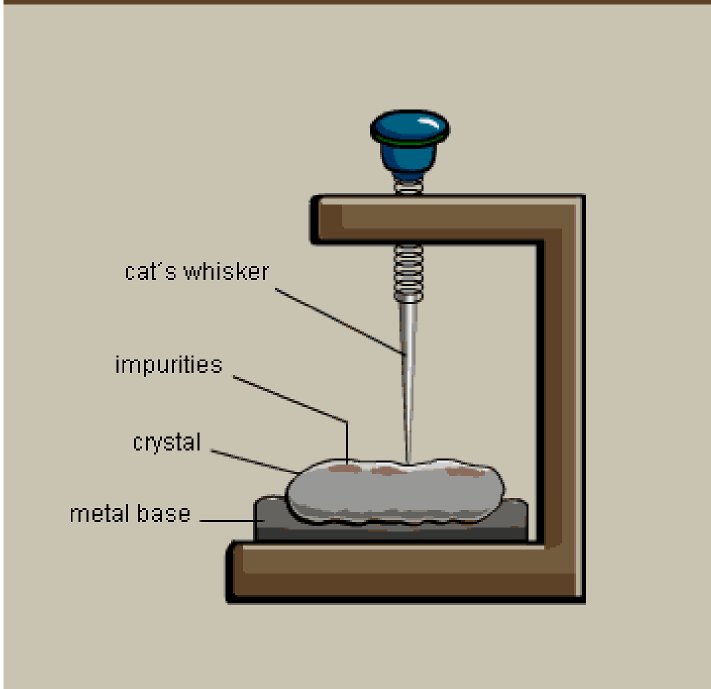


The Crystal Radio

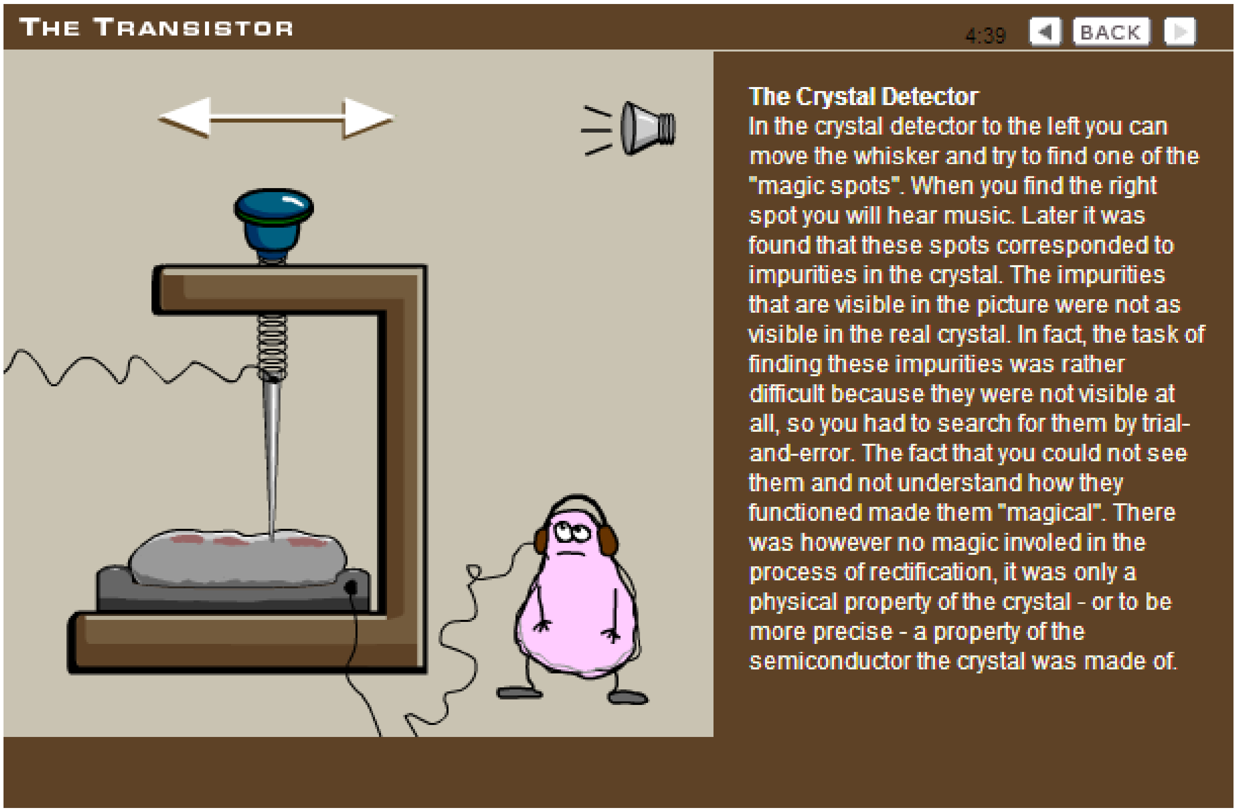
Before the transistor radio was invented there existed something called a crystal radio. In this radio there was a crystal that removed the noise from the radio signal. This crystal was very important to be able to use the radio. Although not known at the time, this crystal was constructed of a material called semiconductor, which is the key to the transistor's function.

In the crystal radio the crystal was often made of galena (lead sulfide), iron pyrite (fool's gold), zincite (zinc oxide) and sometimes the elements germanium or silicon. These substances are all examples of semiconductors.

THE TRANSISTOR 3:39 BACK

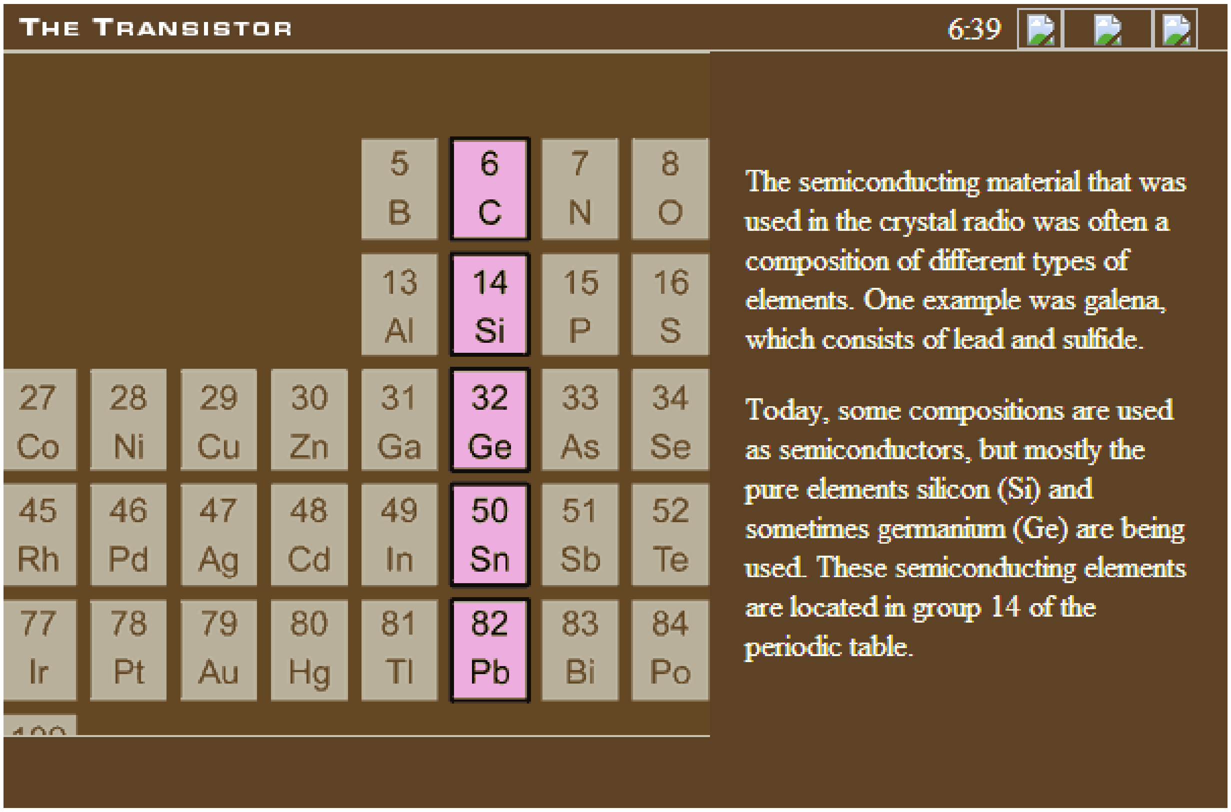


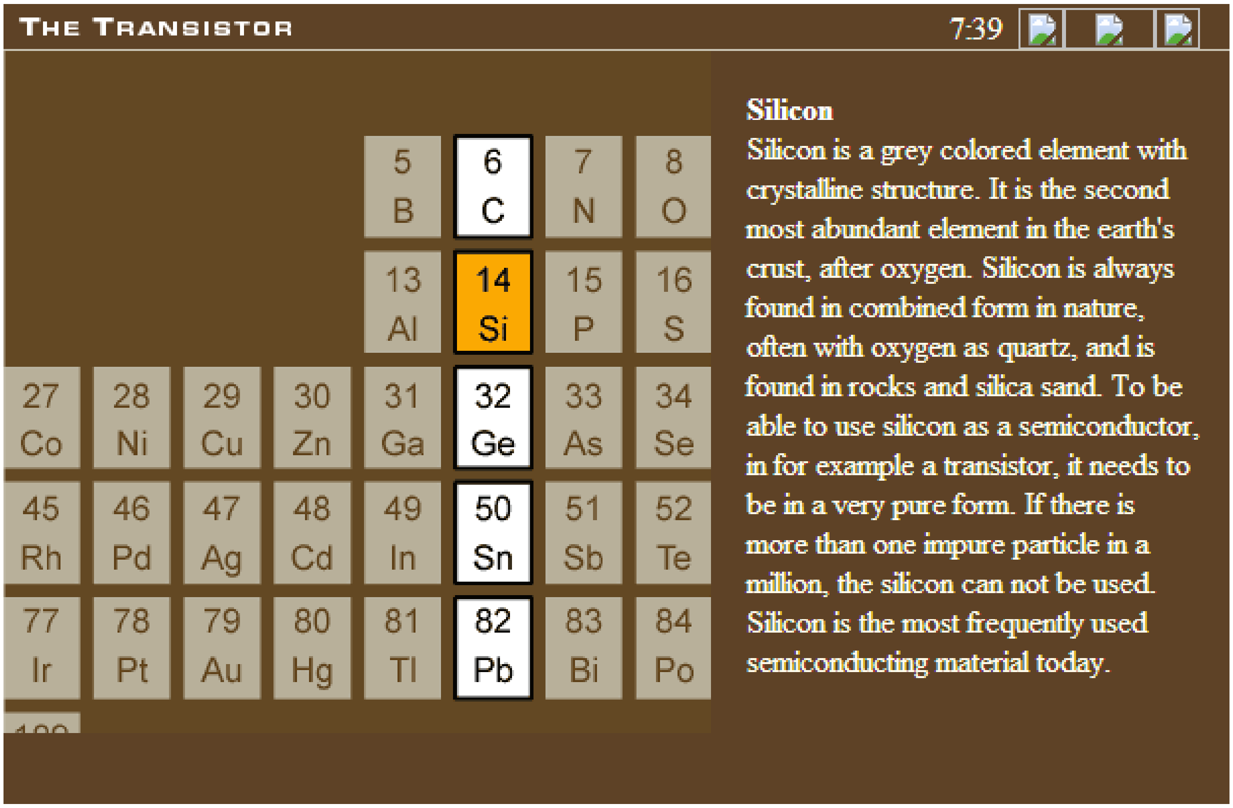
Touching against the crystal was something called a cat's whisker, a tiny metal wire often composed of steel or even gold. By moving it, the radio user tried to find one of the so-called "magic spots." The reason for this name was that only small areas on the crystal would make the radio function - just like magic. When the whisker was positioned over a "magic spot" an electrical signal travelled down the metal wire and through the crystal. When this happened you could hear the music being transmitted from the radio station. This was an example of a phenomenon called rectification. Rectification means that an electrical current is only allowed to pass in one direction and not in the other. In the crystal radio the cat's whisker and the crystal worked as a rectifier.

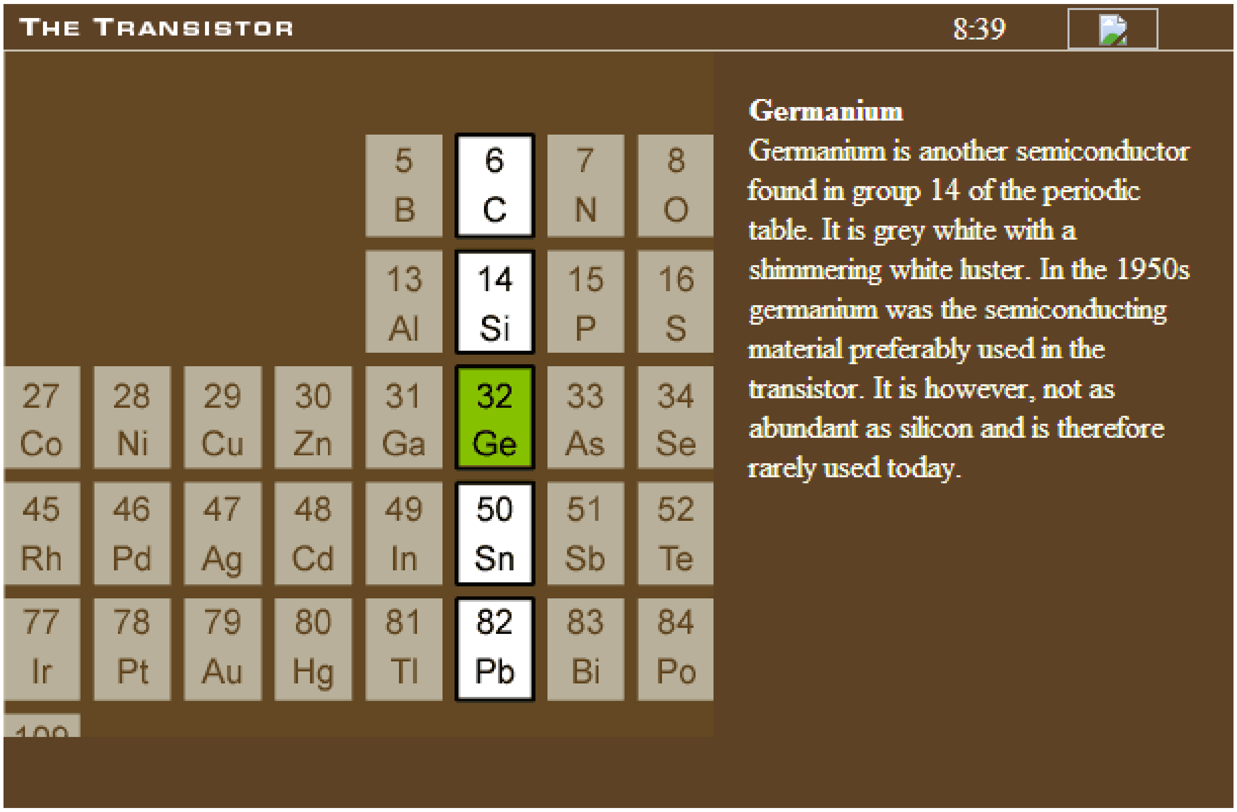

THE TRANSISTOR
4:39
BACK
The Crystal Detector
In the crystal detector to the left you can move the whisker and try to find one of the "magic spots". When you find the right spot you will hear music. Later it was found that these spots corresponded to impurities in the crystal. The impurities that are visible in the picture were not as visible in the real crystal. In fact, the task of finding these impurities was rather difficult because they were not visible at all, so you had to search for them by trial-and-error. The fact that you could not see them and not understand how they functioned made them "magical". There was however no magic involed in the process of rectification, it was only a physical property of the crystal - or to be more precise - a property of the semiconductor the crystal was made of.

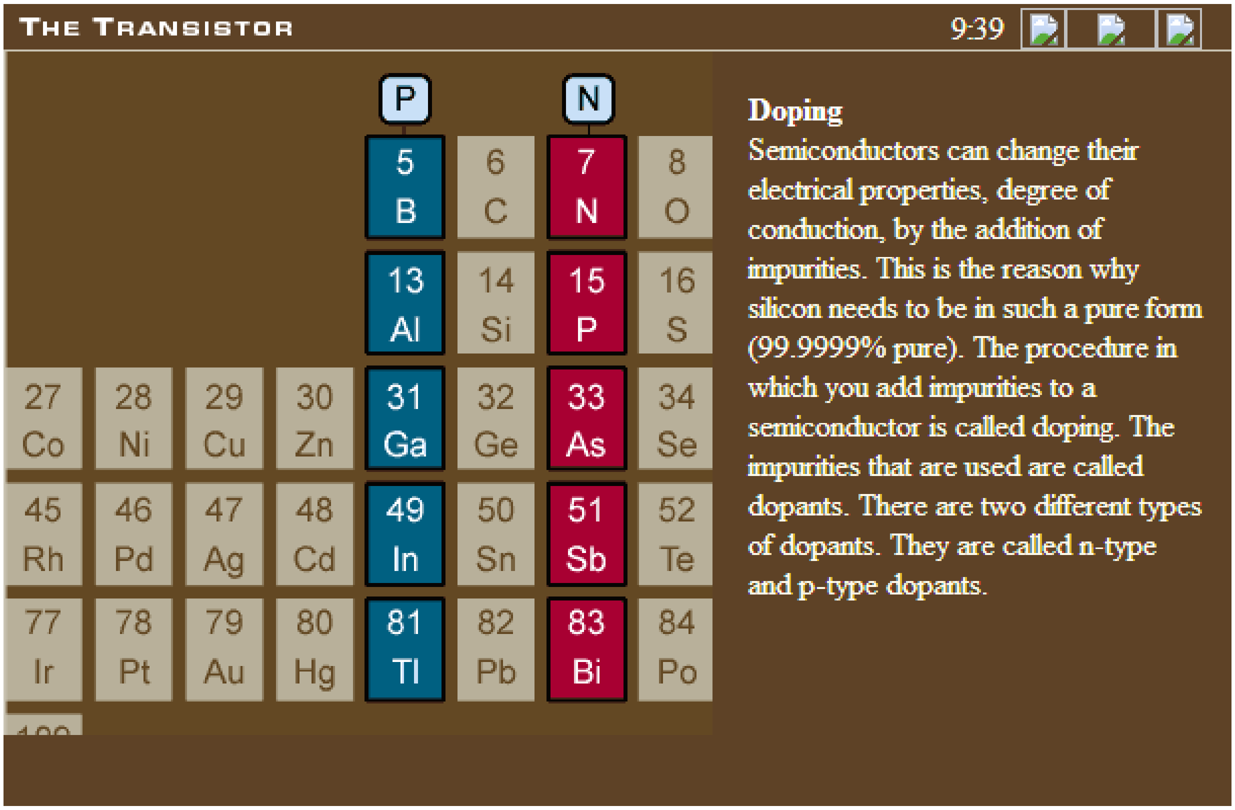
Semiconductor

One of the fundamental concepts in electronics is conduction. Different materials conduct an electrical current in as many diverse ways. One example of a good conductor is a metal, such as gold or copper. An example of a bad conductor is glass or an item made of wood. These bad conductors are called insulators. A semiconducting material has conducting properties between these good/bad conductors. It is called a semiconductor, an intermediate between a conductor and an insulator.

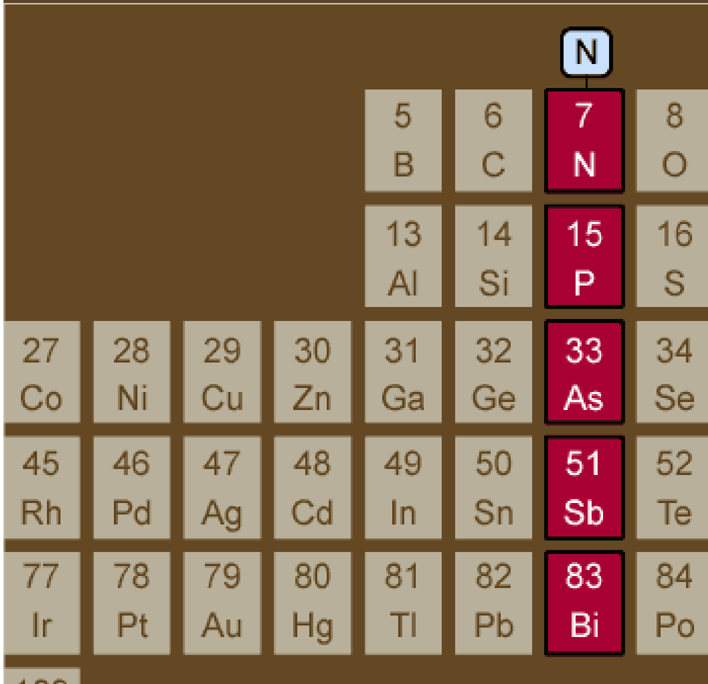

The Transistor
6:39
5 B
6 C
7 N
8 O
13 Al
14 Si
15 P
16 S
27 Co
28 Ni
29 Cu
30 Zn
31 Ga
32 Ge
33 As
34 Se
45 Rh
46 Pd
47 Ag
48 Cd
49 In
50 Sn
51 Sb
52 Te
77 Ir
78 Pt
79 Au
80 Hg
81 Tl
82 Pb
83 Bi
84 Po
The semiconducting material that was used in the crystal radio was often a composition of different types of elements. One example was galena, which consists of lead and sulfide.
Today, some compositions are used as semiconductors, but mostly the pure elements silicon (Si) and sometimes germanium (Ge) are being used. These semiconducting elements are located in group 14 of the periodic table.


THE TRANSISTOR
7:39
5 B
6 C
7 N
8 O
13 Al
14 Si
15 P
16 S
27 Co
28 Ni
29 Cu
30 Zn
31 Ga
32 Ge
33 As
34 Se
45 Rh
46 Pd
47 Ag
48 Cd
49 In
50 Sn
51 Sb
52 Te
77 Ir
78 Pt
79 Au
80 Hg
81 Tl
82 Pb
83 Bi
84 Po
Silicon
Silicon is a grey colored element with crystalline structure. It is the second most abundant element in the earth's crust, after oxygen. Silicon is always found in combined form in nature, often with oxygen as quartz, and is found in rocks and silica sand. To be able to use silicon as a semiconductor, in for example a transistor, it needs to be in a very pure form. If there is more than one impure particle in a million, the silicon can not be used. Silicon is the most frequently used semiconducting material today.


The Transistor
8:39
5 B
6 C
7 N
8 O
13 Al
14 Si
15 P
16 S
27 Co
28 Ni
29 Cu
30 Zn
31 Ga
32 Ge
33 As
34 Se
45 Rh
46 Pd
47 Ag
48 Cd
49 In
50 Sn
51 Sb
52 Te
77 Ir
78 Pt
79 Au
80 Hg
81 Tl
82 Pb
83 Bi
84 Po
Germanium
Germanium is another semiconductor found in group 14 of the periodic table. It is grey white with a shimmering white luster. In the 1950s germanium was the semiconducting material preferably used in the transistor. It is however, not as abundant as silicon and is therefore rarely used today.

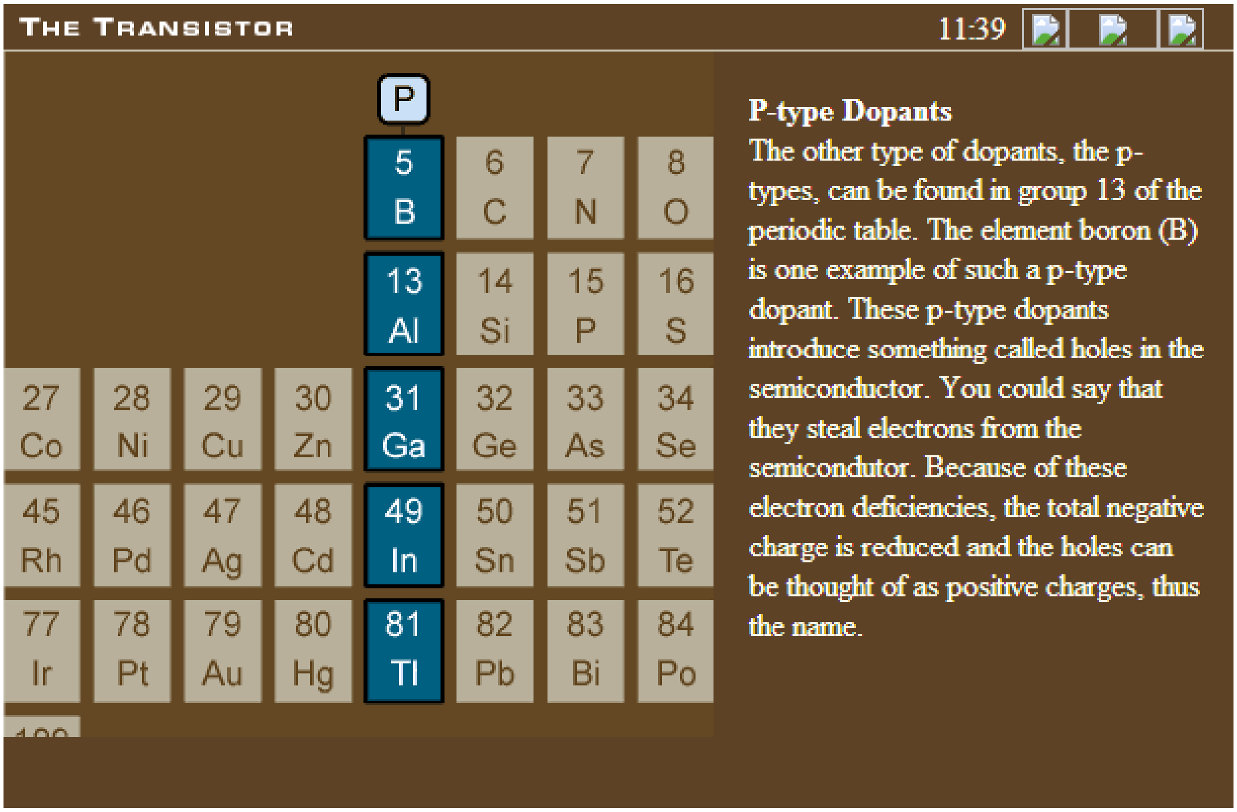

THE TRANSISTOR
9:39
P
N
5 B
6 C
7 N
8 O
13 Al
14 Si
15 P
16 S
27 Co
28 Ni
29 Cu
30 Zn
31 Ga
32 Ge
33 As
34 Se
45 Rh
46 Pd
47 Ag
48 Cd
49 In
50 Sn
51 Sb
52 Te
77 Ir
78 Pt
79 Au
80 Hg
81 Tl
82 Pb
83 Bi
84 Po
Doping
Semiconductors can change their electrical properties, degree of conduction, by the addition of impurities. This is the reason why silicon needs to be in such a pure form (99.9999% pure). The procedure in which you add impurities to a semiconductor is called doping. The impurities that are used are called dopants. There are two different types of dopants. They are called n-type and p-type dopants.

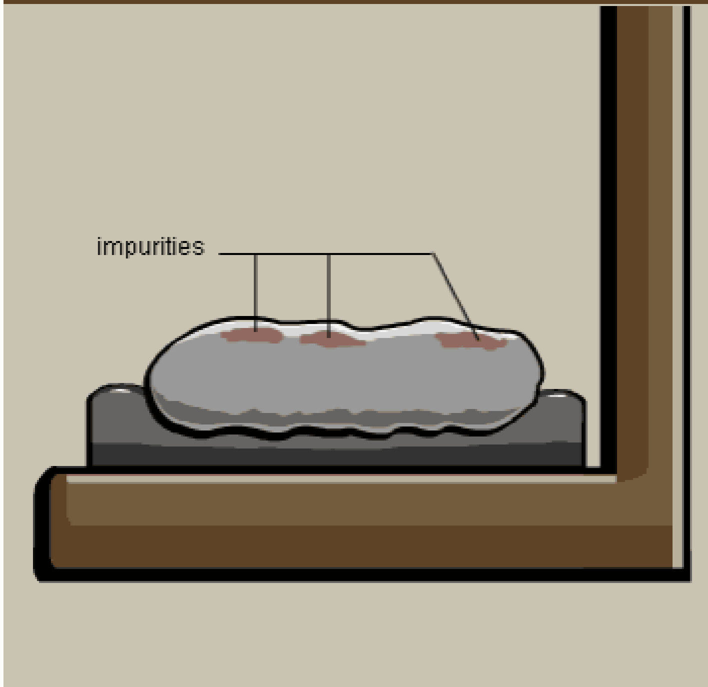
THE TRANSISTOR 10:39



N-type Dopants

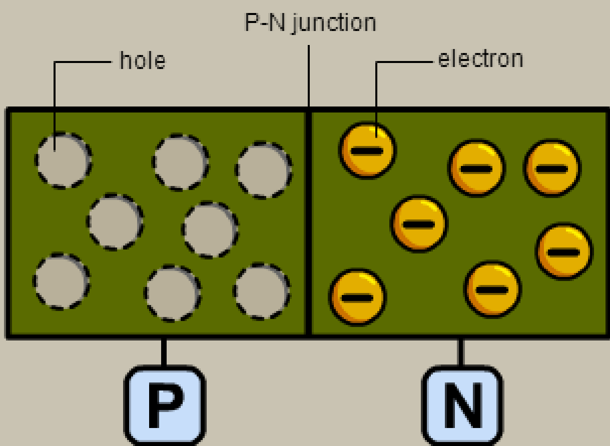
Some of the most common n-type dopants are located in group 15 of the periodic table. One example of such an dopant is the element phosphorus (P). The n-type dopant adds negatively charged electrons to the semiconductor, thus the name. Sometimes these dopants are called donators because they donate electrons to the semiconductor.


THE TRANSISTOR
11:39
P
5 B
6 C
7 N
8 O
13 Al
14 Si
15 P
16 S
27 Co
28 Ni
29 Cu
30 Zn
31 Ga
32 Ge
33 As
34 Se
45 Rh
46 Pd
47 Ag
48 Cd
49 In
50 Sn
51 Sb
52 Te
77 Ir
78 Pt
79 Au
80 Hg
81 Tl
82 Pb
83 Bi
84 Po
P-type Dopants
The other type of dopants, the p-types, can be found in group 13 of the periodic table. The element boron (B) is one example of such a p-type dopant. These p-type dopants introduce something called holes in the semiconductor. You could say that they steal electrons from the semicondutor. Because of these electron deficiencies, the total negative charge is reduced and the holes can be thought of as positive charges, thus the name.



Dopants in the Crystal Radio
The impurities in the detector crystal in the crystal radio consisted of these different kinds of n- and/or p-type dopants. At the time of the crystal radio the nature of the impurities were not known. Although the radio users didn't understand the theory behind the phenomen of rectification, they used it in practice when they tried to find one of the "magic spots," where rectification occured.

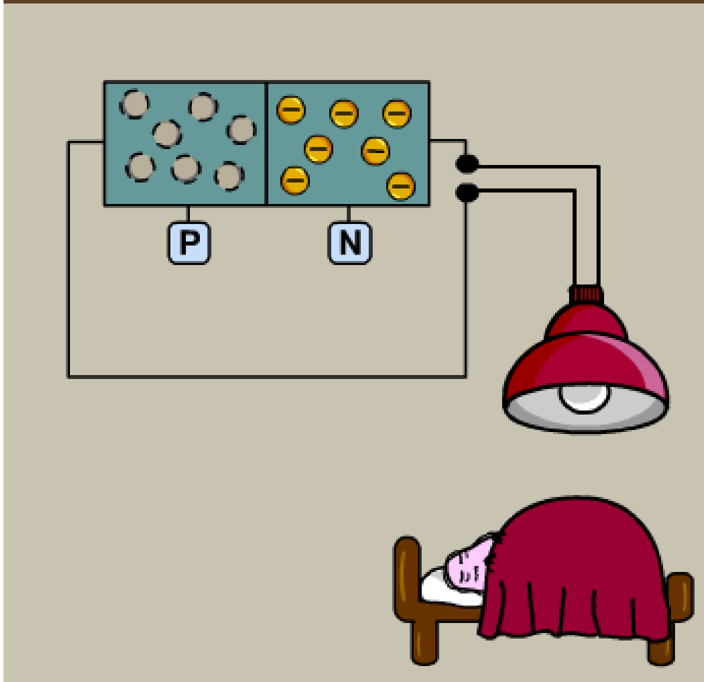
THE TRANSISTOR 13:39



Rectification

The theory of rectification is based on the special properties of the junction that is created between a semiconductor area of n-type and one of p-type. The junction is often referred to as a P-N junction. On one side (N) of this junction there is an excess of electrons because of the addition of n-type dopants to the semiconductor. On the other side (P) where the p-type dopants have been added, there is an excess of holes. This construction, with one n-type and one p-type area in one unit, is generally called a diode. The diode acts as an electrical rectifier, i.e. it only allows current to pass in one direction.

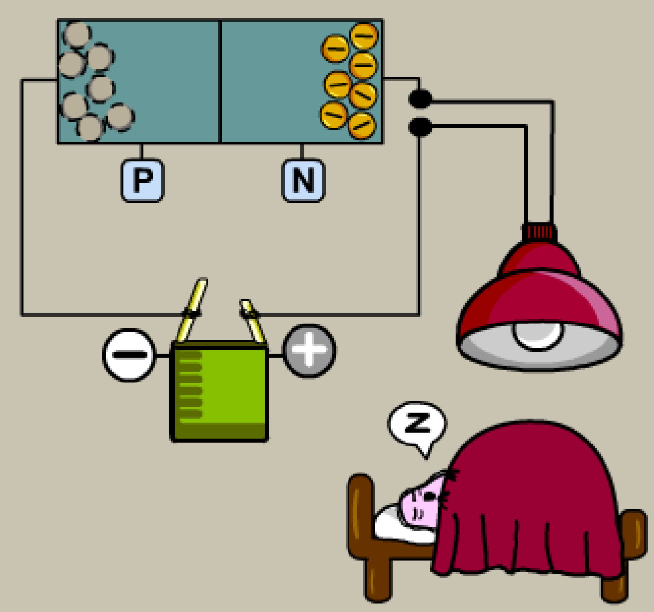
THE TRANSISTOR 14:39



The Diode

If we place the diode in an electrical circuit, such as the one to the left, the lamp will not shine. The reason for this is that there is no power source, such as a battery, and therefore there is no current present in the circuit.

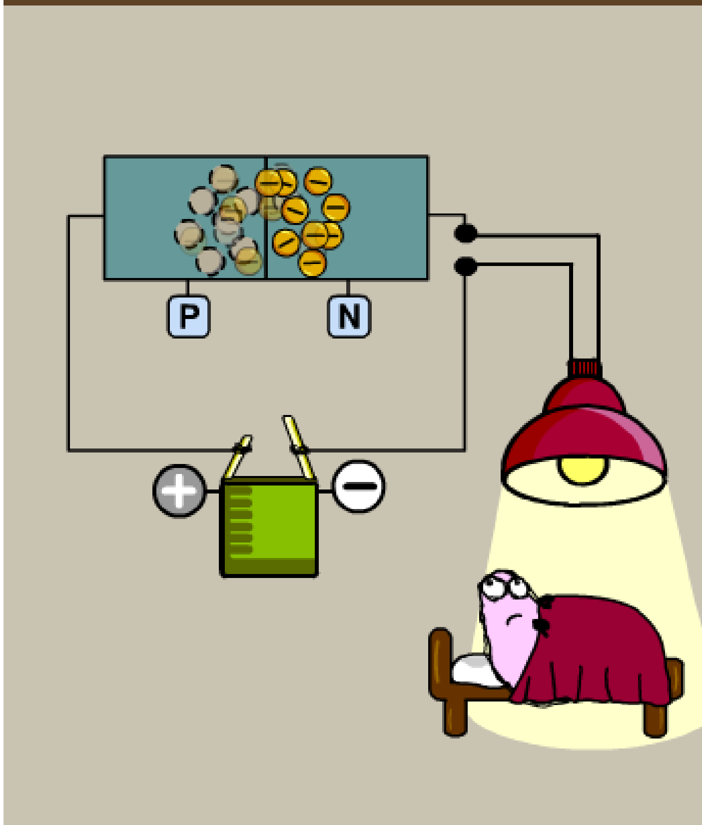
THE TRANSISTOR 15:39



Reversed Bias

If we introduce a battery, the properties of the n-type and the p-type area will change. If the polarity of the battery is placed as in the picture, the negative electrons will move to the right towards the positive pole of the battery. The reason for this movement is because electrical charges of opposite sign attract. In the p-type area the holes will move to the left towards the other, the negative pole of the battery. The gap that is created in the proximity of the P-N junction prevents the current from flowing in the circuit and the lamp will not shine. In this case, when the battery is applied with the positive pole versus the n-type area, the diode is said to be reversed biased. This means that no current is present in the circuit.

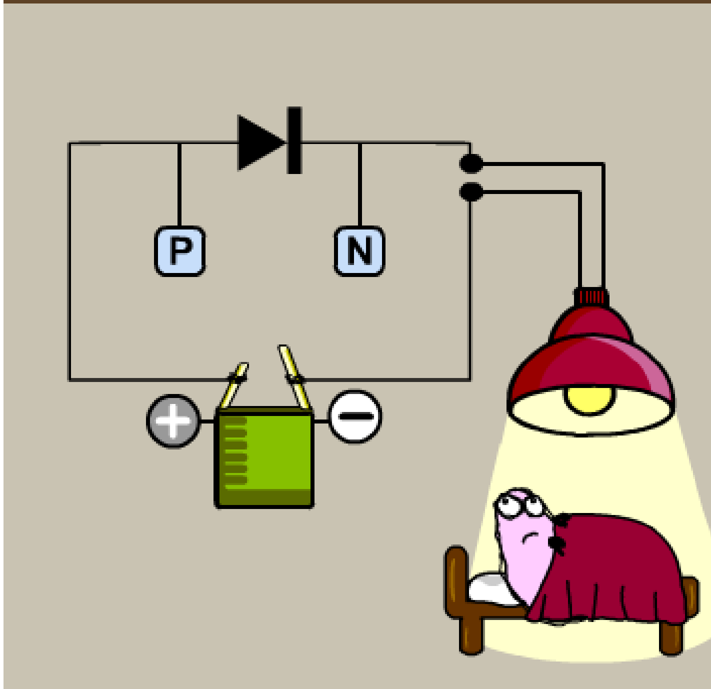
THE TRANSISTOR 16:39



Forward Bias

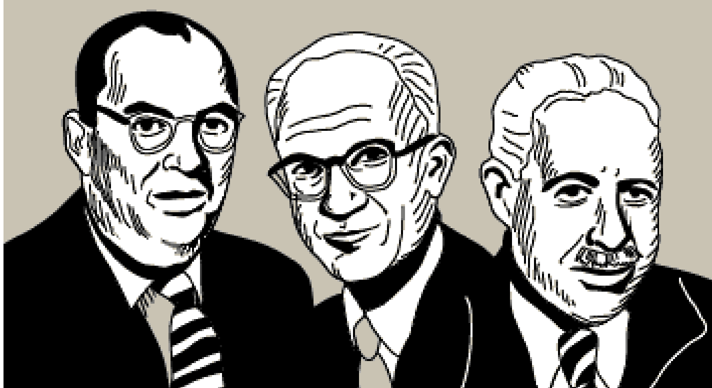
If we reverse the polarity of the battery, both the electrons and the holes will move towards the P-N junction. When the electrons are in such a close vicinity to the holes they will be able to jump into the holes. The movement of the electrons will enable the flow of the current in the circuit and the lamp will shine on our little friend. You could think of this polarity change as if the current changed to the opposite direction compared to the reversed biased diode. In this circuit, where the positive pole is placed versus the p-type area, the current will flow through the diode and the diode is said to be forward biased.

THE TRANSISTOR 17:39



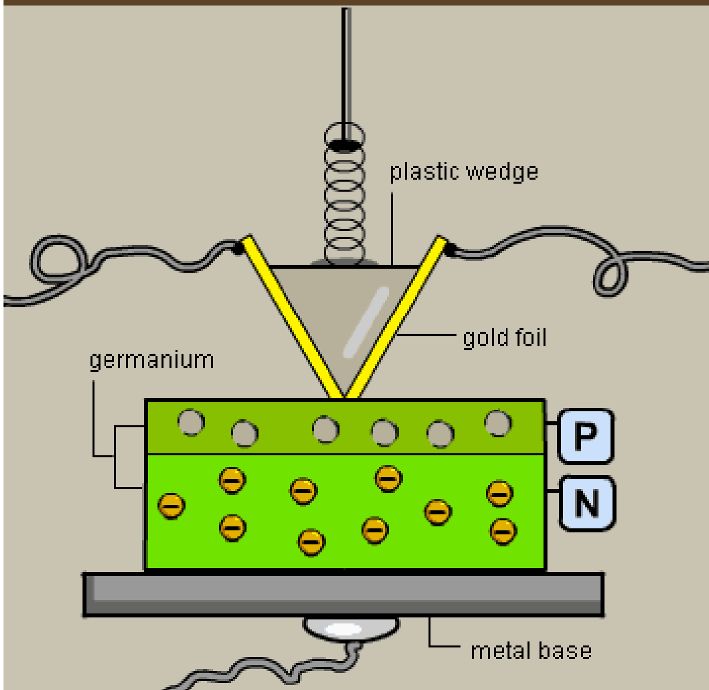
Diode Symbol

The symbol for the diode that is being used when you draw electrical diagrams shows its rectifying property. The symbol is an arrow that shows in which direction the current is allowed to pass, i.e. the forward biased direction of the diode. Today the diode is still being used as a detector in electrical circuits such as in the television or radio receiver. The diode is also being used as a converter of AC (Alternating Current) to DC (Direct Current) in power supply units.



The First Transistor
In the late 40s three American scientists named William Shockley, John Bardeen and Walter Brattain at Bell Labs, announced the creation of the first transistor. The name transistor is a combination of the words transfer and resistor - a transfer resistor - a transistor. When it was announced the name was explained; "because it is a resistor or semiconductor device which can amplify electrical signals as they are transferred through it from input to output terminals." This, the very first transistor was called a point-contact transistor. Shockley, Bardeen and Brattain received the Nobel Prize in Physics 1956 "for their researches on semiconductors and their discovery of the transistor effect."

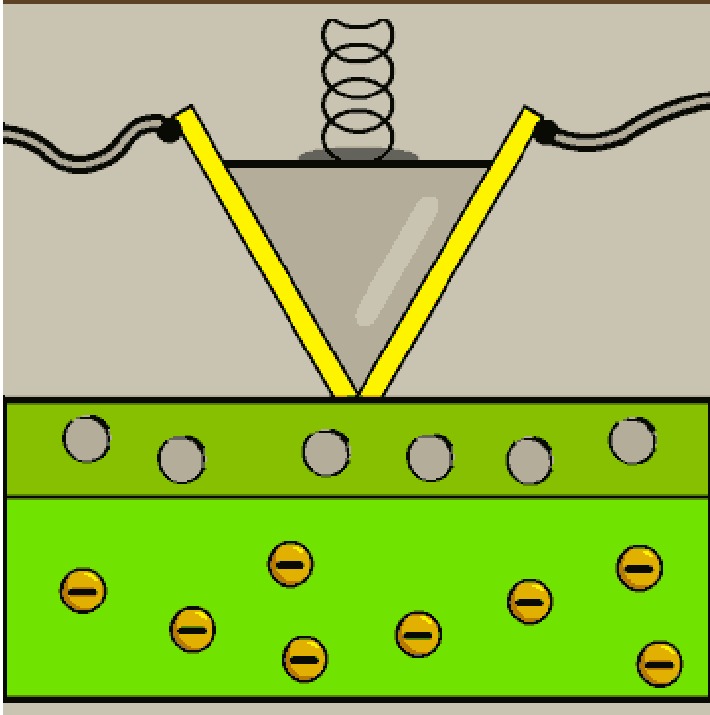
THE TRANSISTOR 19:39



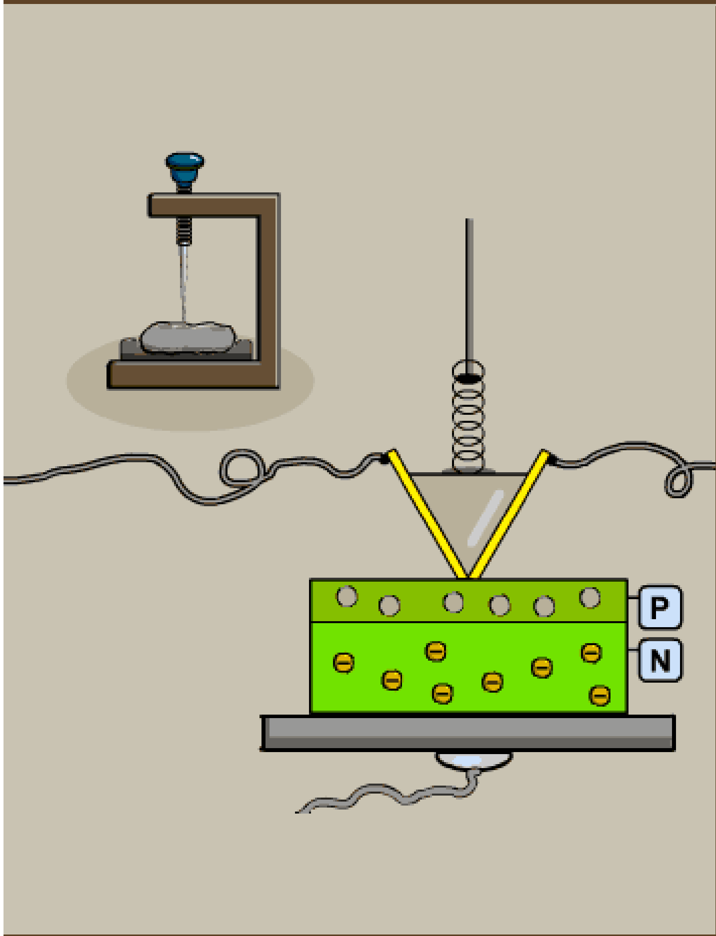
The Point-Contact Transistor

The first point-contact transistor consisted of a germanium crystal placed on a metal plate. The germanium was of n-type with a so-called p-type inversion layer. Pressing down on this crystal was a plastic wedge, which was kept in place by a modified paper clip. On this wedge a thin gold foil, cut in two halves with a razor blade, had been cemented. The slit was very important to be able to make this point-contact device to work.

THE TRANSISTOR 20:39



In the beginning, the scientists tried to place two ordinary metal contacts on the surface of the germanium crystal. The problem was to bring these contacts sufficiently close to each other, to obtain the transistor effect. The contacts needed to be within a few hundredths of a millimetre, about the thickness of a paper, and this was very hard to achieve with conventional metal wire contacts. With the use of the two point-contacts created by the slit in the gold foil, the contacts could be placed within a distance of 0.05 mm of each other.

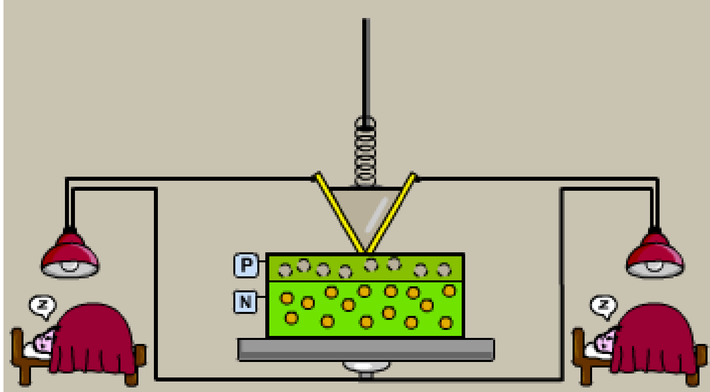


Point-Contact and Crystal Detector

The construction of the point-contact transistor resembles the crystal detector in the crystal radio. You could think of the point-contact transistor as a crystal detector with two whiskers instead of one - the two gold contacts touching the crystal.

The transistor consists of three electrical terminals, one on each gold contact and one on the metal base. The diode, of which the crystal detector was an example, on the other hand has only two terminals. You could therefore refer to the transistor as a triode. As you might already know tri comes from the latin and greek word for three.

THE TRANSISTOR 22:39

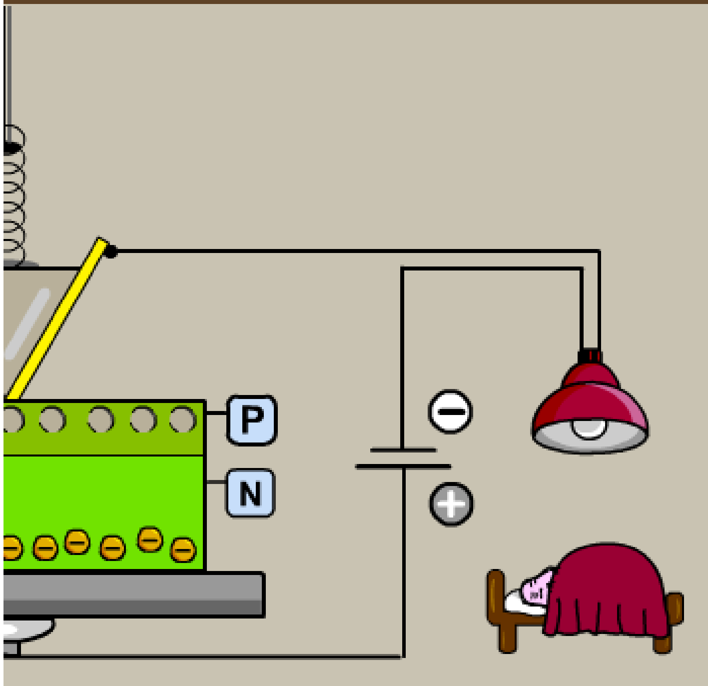


Point-Contact Circuit

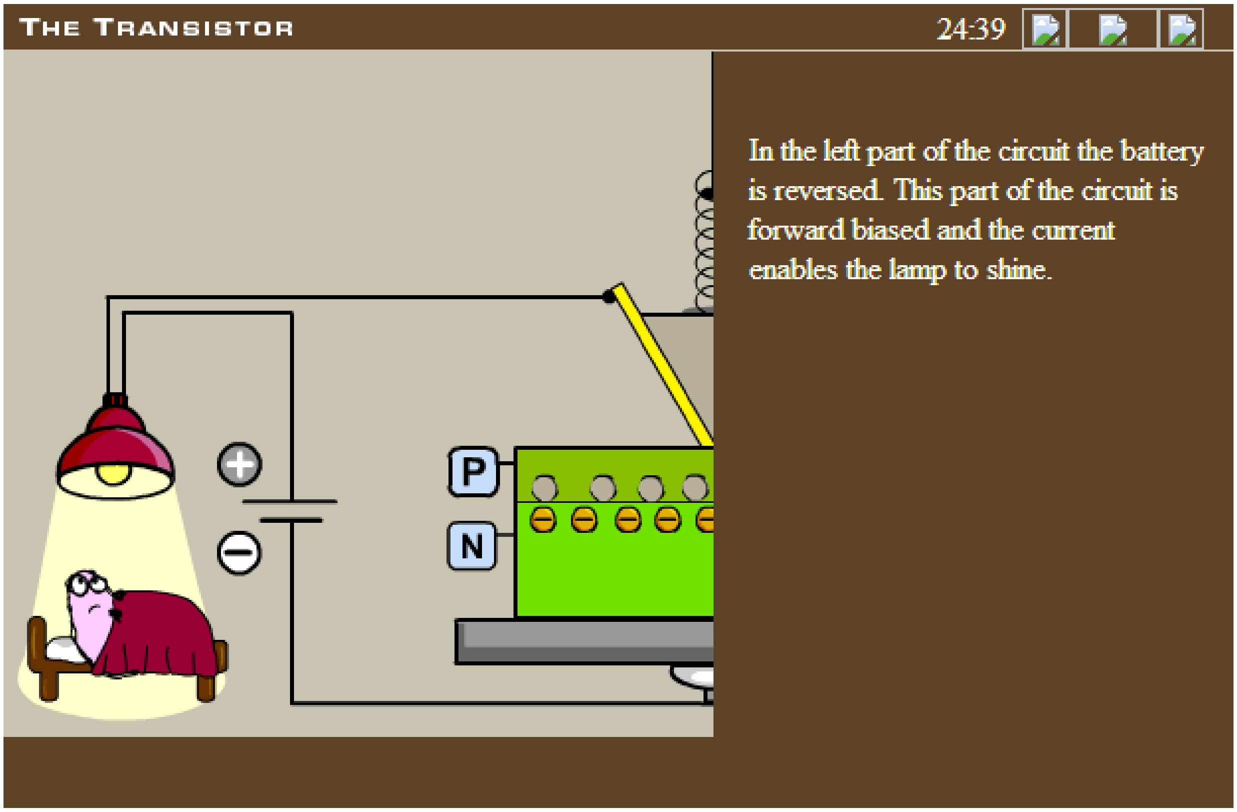
If we introduce the transistor in a circuit, such as the one in the picture, you can see that all three terminals are connected. We can think of the left and the right half of the circuit as two separate parts, although they are physically connected to each other in reality. Each of these halves has characteristics similar to the diode that we described previously.

THE TRANSISTOR

23:39

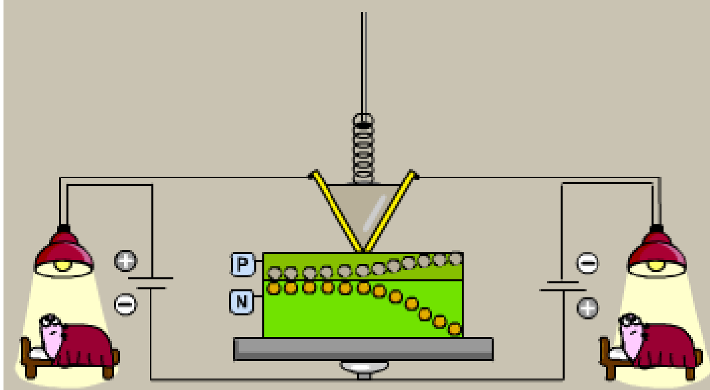


The right part consists of a germanium crystal, one single point-contact made of gold, a lamp and a battery. The battery is placed with the positive pole versus the n-type area of the germanium crystal. The negative pole is placed versus the p-type area and as in the example of the reversed biased diode no current is present in the circuit. The lamp does not shine.


THE TRANSISTOR
24:39
In the left part of the circuit the battery is reversed. This part of the circuit is forward biased and the current enables the lamp to shine.
P
N

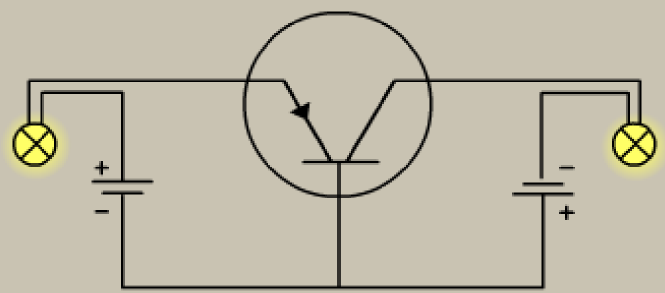
The Transistor

25:39



If we bring these two parts together, as they really are in the point-contact transistor circuit, both lamps will shine. This is because the left, forward biased, part of the circuit introduces enough holes to be able to activate the right part as well. So with the transistor you are able to control and influence one part of the circuit, the right part, with the aid of the other, the left part.

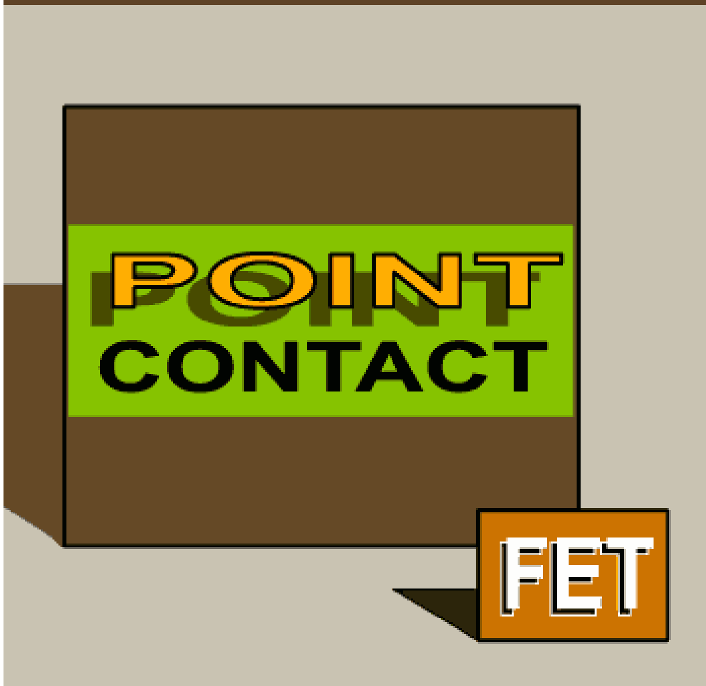
The three electrodes of the transistor have different names. Connected to the metal base is the contact simply called the base-electrode. The gold contact where the current enters (left) is called the emitter-electrode and the other gold contact (right) is called the collector-electrode.



Point-Contact Symbol

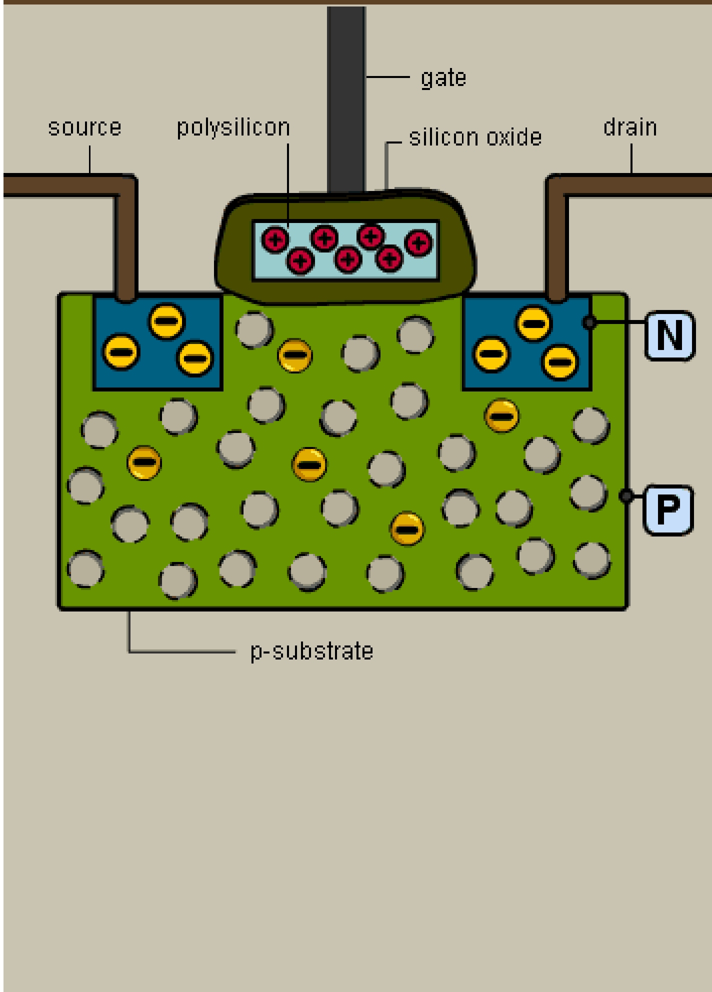
Just as the diode, there is a schematic representation of the point-contact transistor. As you might notice there is a high resemblance between the schematic symbol and the physical transistor. The emitter is always represented with an arrow.

THE TRANSISTOR 27:39



Today's Transistors

The point-contact transistor was however rather big, compared with today's transistors, and difficult to produce in large scale. The most used transistor of today is therefore of a diffent kind. It is called a field-effect transistor or FET for short. There are several kinds of FETs today and the one we are going to look at is called metal-oxide semiconductor field-effect transistor (MOSFET). The basis for the function of the different FETs is however the same.

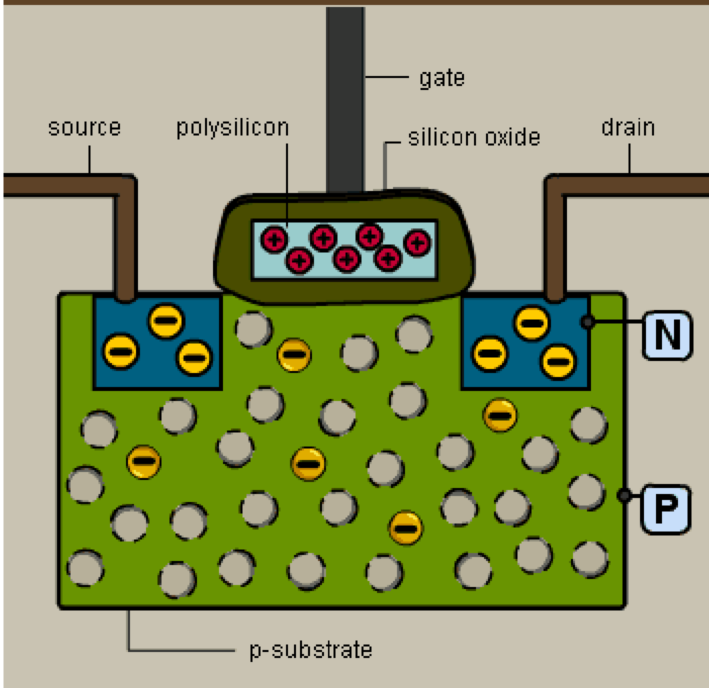


The MOSFET

The MOSFET is constructed in a different way compared to the point-contact transistor. First of all, the semiconducting material has been replaced. Instead of germanium, silicon is being used. Silicon is less expensive and has physical properties better suited for today's transistors. The MOSFET is made up of silicon layers with different kinds of doping. The electrodes in this type of transistor are called source, drain and gate. The current flowing from source to drain is controlled with the charge of the gate, which is constructed of a highly conductive material, often polysilicon. To function properly this gate-electrode needs to be insulated from the rest of the transistor. This insulation is made possible by a layer of silicon oxide that surrounds the gate.

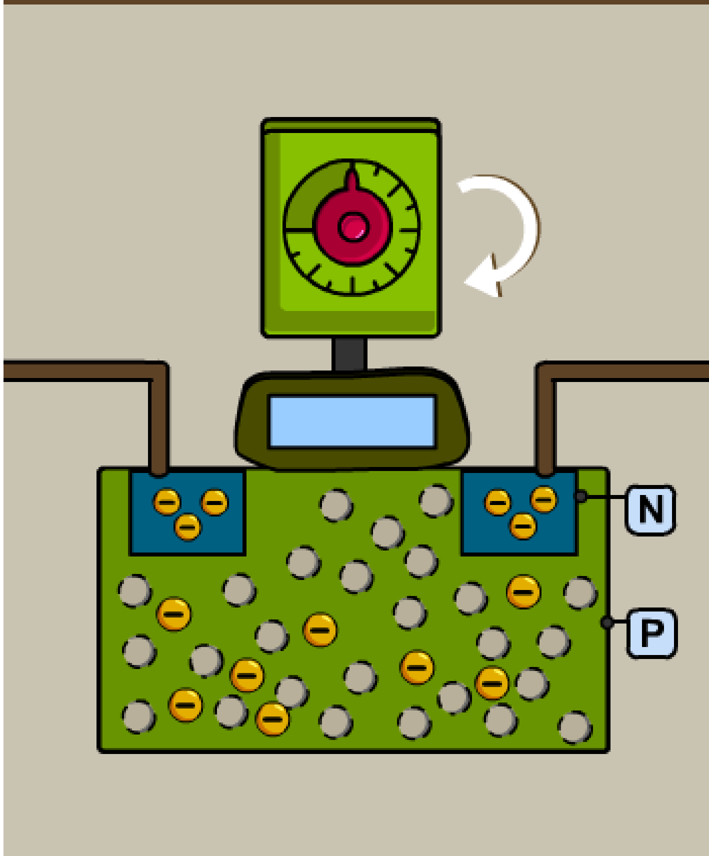
THE TRANSISTOR

29:39



The type of MOSFET we are looking at is called a n-channel transistor. If we change the doping of the source and drain from n to p, and the p-type substrate to n-type we have a p-channel transistor. This transistor functions in the same way, but with the polarities changed.

THE TRANSISTOR 30:39

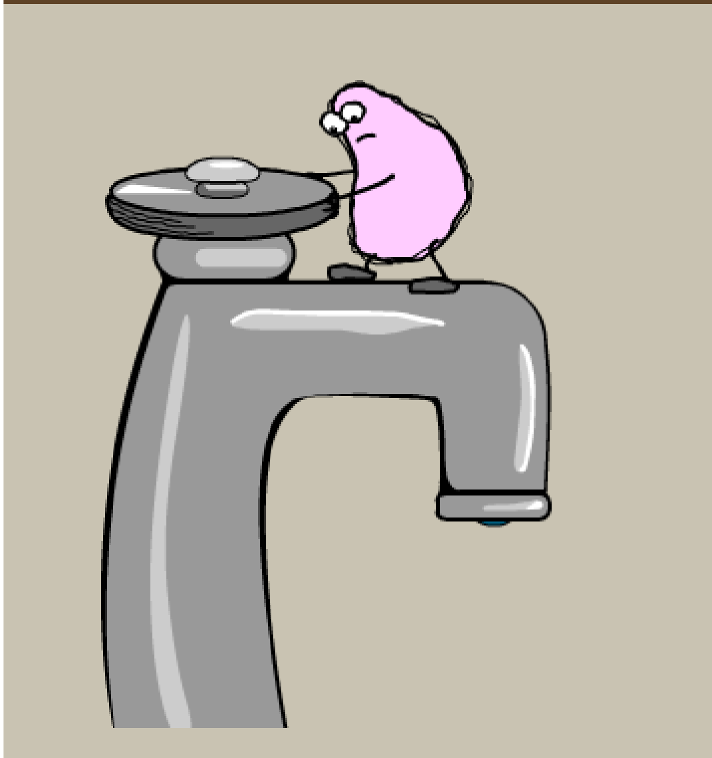


Gate Control

When the voltage (charge) of the gate is changed, more or less electrons are attracted from the p-type substrate area. Although there are a lot of holes in the p-substrate there are still some electrons left in this semiconductor area. With the input signal to the gate you can increase the current from source to drain (output signal) as well as decrease it. If there are a lot of positive charges on the gate then it will attract a lot of electrons and the current flowing from source to drain will increase.

You can try to change the charge of the gate by modulating the knob and see what happens.

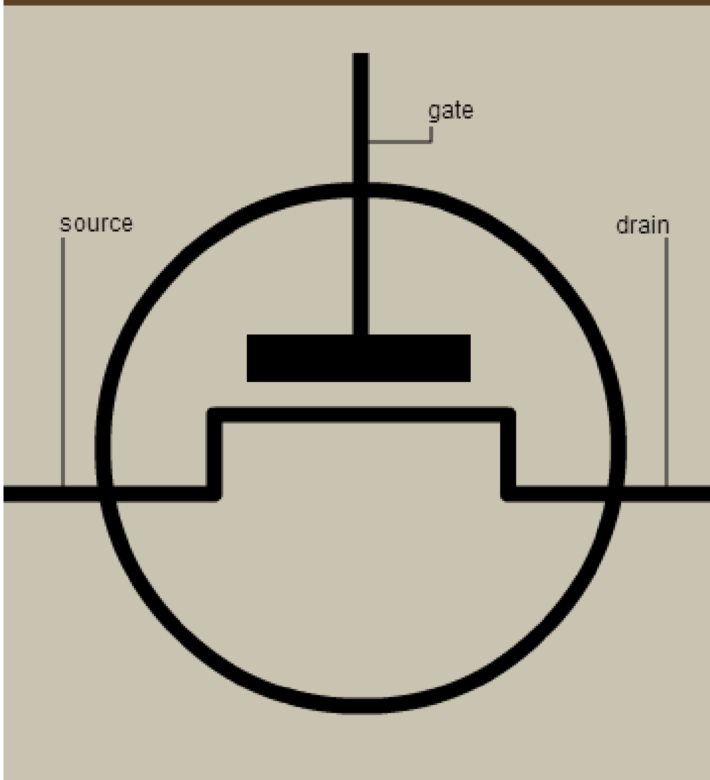
THE TRANSISTOR 31:39



Water Tap

You can think of the operation of any transistor as an ordinary faucet. The water enters the faucet in the pipeline from the water distributor, which would correspond to the source in the MOSFET. The water then leaves the faucet into the sink, this would be the drain in the MOSFET. The water tap controls the amount, flow, of water. In the MOSFET the gate operates as this controller. With a small force you can control the water flow with the water tap, just as you can control the current flowing from the source to the drain, with a small change of the charge of the gate.

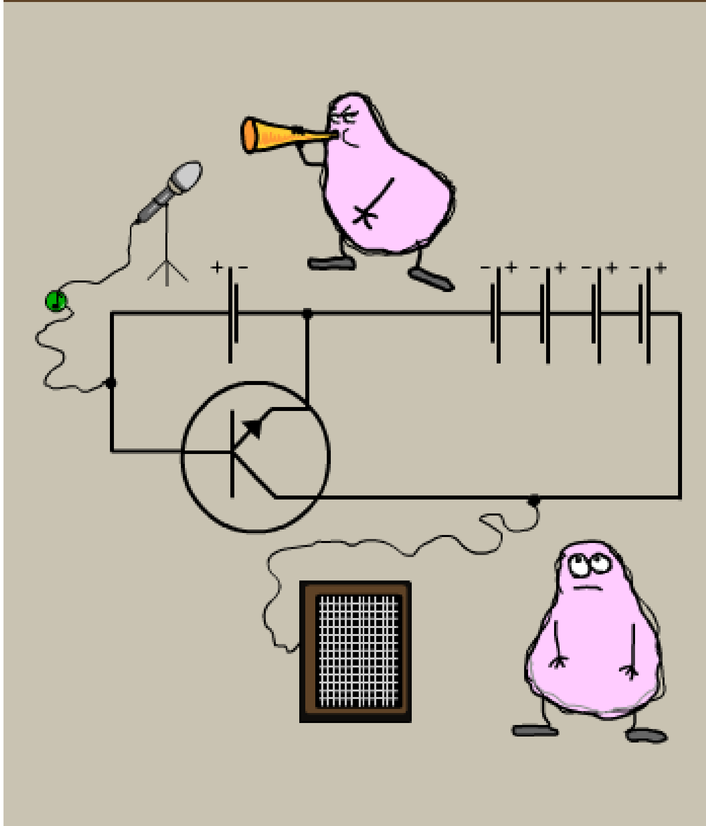
THE TRANSISTOR 32:39



MOSFET Symbol

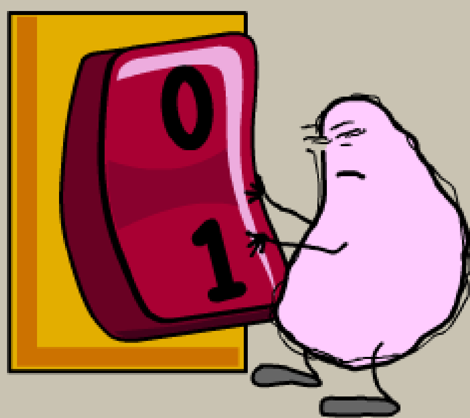
There is also a symbol for this type of transistors. This symbol resembles the one for the point-contact transistor but they are not identical. The names of the electrodes also differ. In the FET the electrodes are called source-gate-drain and as you might remember they are called emitter-base-collector in the point-contact transistor.

Both the point-contact transistor and the MOSFET can be used for the same purposes. The two main uses of any transistor are as a modulator/amplifier and as a switch.



Amplification

In the case of amplification the transistor is used to amplify a signal. One example of such a signal can be a sound. To the left is an example of a circuit that is used to amplify sound signals. The sound entering the microphone is converted to an electrical signal that is amplified in the transistor. This amplified sound signal then travels through the circuit until it reaches the loudspeaker. This speaker converts the electrical sound signal back into a sound. The sound leaving the speaker is the same as the sound that entered the microphone, only much louder. This is called amplification – the sound is being amplified.

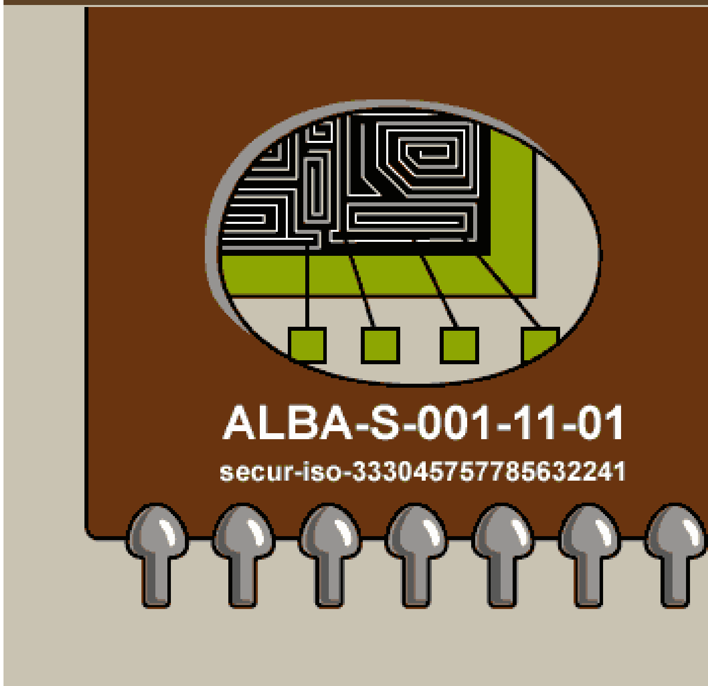


Switching

In the case of switching the transistor is used as an electronic switch. The transistor can turn an electrical circuit on or off, just as in the case with the point-contact transistor and the two lamps. With the left part you could turn on the right lamp.

The difference between a transistor and a mechanical switch is that the transistor doesn't contain any moving parts and therefore doesn't break quite as easily. The transistor is also operated by an electrical current and not by a mechanical force, such as the pushing of a finger. This enables the transistor to switch between on/off much faster than a mechanical one.

THE TRANSISTOR 35:39



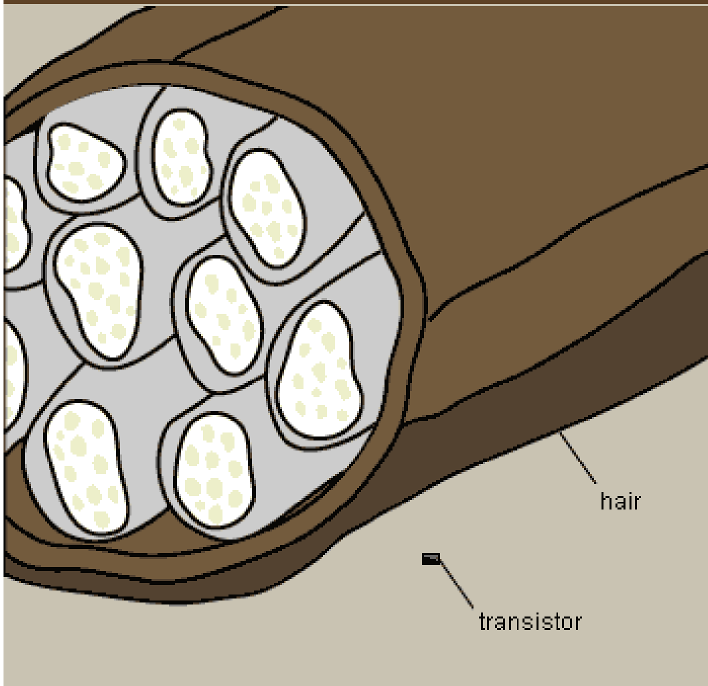
Computers

The property of the transistor, being able to switch between two different states (on-off) is very important for a computer's function. In a computer the transistor can be made to switch between two binary states called 0 and 1. The transistor is used by the computer to do calculations, etc. In today's complex computers there are several thousands, even millions of transistors. In a computer it is not present as a single isolated item, instead it is part of something that is called an integrated circuit.

THE TRANSISTOR 36:39

The Integrated Circuit

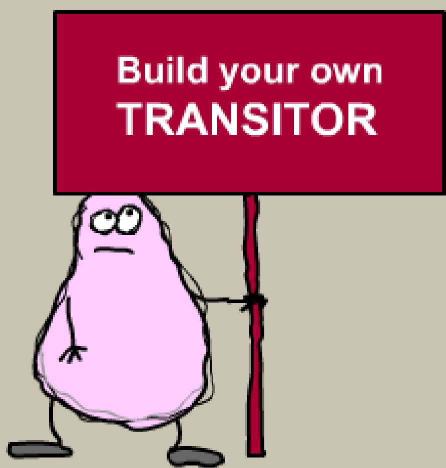
An integrated circuit is made of a single piece of semiconductor material and contains, apart from transistors, other electronical components. The inventors of the integrated circuit were Jack Kilby and Robert Noyce who died in 1990. Jack Kilby was awarded the Nobel Prize in Physics 2000.



Small Transistors

In a computer chip there can exist as many as a million transistors. In some of the most advanced chips even several hundreds of millions of transistors are present. To be able to fit that many transistors into something as small as a computer chip, they need to be extremely small. In fact, today's transistors are many, many times smaller then even a single human hair.

THE TRANSISTOR 38:39



Build the First Transistor - The Point-Contact Transistor

Now that you have learned about the function of the transistor you can try to build your own. The transistor you are going to build is the first transistor ever made - the point-contact transistor.

Good luck!