

Introduction to Transistors

Edited by

Research Lab 207, Science Building

(School of Mathematics and Physics, Changzhou University, Changzhou, China)

A transistor is a semiconductor device used to amplify and switch electronic signals. It is made of a solid piece of semiconductor material, with at least three terminals for connection to an external circuit. A voltage or current applied to one pair of the transistor's terminals changes the current flowing through another pair of terminals. Because the controlled (output) power can be much more than the controlling (input) power, the transistor provides amplification of a signal. Today, some transistors are packaged individually, but many more are found embedded in integrated circuits.



Figure 1a. A replica of the first working transistor.

The transistor is the fundamental building block of modern electronic devices, and is ubiquitous in modern electronic systems. Following its release in the early 1950s the transistor revolutionised the field of electronics, and paved the way for smaller and cheaper radios, calculators, and computers, amongst other things.



Figure 1b. Assorted discrete transistors. Packages in order from top to bottom:
TO-3, TO-126, TO-92, SOT-23

Table of Contents

1. History

2. Importance

3 Simplified operation

 3.1 Transistor as a switch

 3.2 Transistor as an amplifier

4 Comparison with vacuum tubes

 4.1 Advantages

 4.2 Limitations

5 Types

 5.1 Bipolar junction transistor

 5.2 Field-effect transistor

6 Construction

1. History

Physicist Julius Edgar Lilienfeld filed the first patent for a transistor in Canada in 1925, describing a device similar to a Field Effect Transistor or "FET". However, Lilienfeld did not publish any research articles about his devices,[citation needed] nor did his patent cite any examples of devices actually constructed. In 1934, German inventor Oskar Heil patented a similar device.

From 1942 Herbert Mataré experimented with so-called Duodiodes while working on a detector for a Doppler RADAR system. The duodiodes built by him had two separate but very close metal contacts on the semiconductor substrate. He discovered effects that could not be explained by two independently operating diodes and thus formed the basic idea for the later point contact transistor.

In 1947, John Bardeen and Walter Brattain at AT&T's Bell Labs in the United States observed that when electrical contacts were applied to a crystal of germanium, the output power was larger than the input. Solid State Physics Group leader William Shockley saw the potential in this, and over the next few months worked to greatly expand the knowledge of semiconductors. The term transistor was coined by John R. Pierce. According to physicist/historian Robert Arns, legal papers from the Bell Labs patent show that William Shockley and Gerald Pearson had built operational versions from Lilienfeld's patents, yet they never referenced this work in any of their later research papers or historical articles. The name 'transistor' is a portmanteau of the term 'transfer resistor'.

The first silicon transistor was produced by Texas Instruments in 1954. This was the work of Gordon Teal, an expert in growing crystals of high purity, who had previously worked at Bell Labs. The first MOS transistor actually built was by Kahng and Atalla at Bell Labs in 1960.

2. Importance

The transistor is the key active component in practically all modern electronics, and is considered by many to be one of the greatest inventions of the twentieth century. Its importance in today's society rests on its ability to be mass produced using a highly automated process (semiconductor device fabrication) that achieves astonishingly low per-transistor costs.

Although several companies each produce over a billion individually packaged (known as discrete) transistors every year, the vast majority of transistors now produced are in integrated circuits (often shortened to IC, microchips or simply chips), along with diodes, resistors, capacitors and other electronic components, to produce complete electronic circuits. A logic gate consists of up to about twenty transistors whereas an advanced microprocessor, as of 2009, can use as many as 2.3 billion transistors (MOSFETs). "About 60 million transistors were built this year [2002] ... for [each] man, woman, and child on Earth."

The transistor's low cost, flexibility, and reliability have made it a ubiquitous device. Transistorized mechatronic circuits have replaced electromechanical devices in controlling appliances and machinery. It is often easier and cheaper to use a standard microcontroller and write a computer program to carry out a control function than to design an equivalent mechanical control function.

The bipolar junction transistor, or BJT, was the most commonly used transistor in the 1960s and 70s. Even after MOSFETs became widely available, the BJT remained the transistor of choice for many analog circuits such as simple amplifiers because of their greater linearity and ease of manufacture. Desirable properties of MOSFETs, such as their utility in low-power devices, usually in the CMOS configuration, allowed them to capture nearly all market share for digital circuits; more recently MOSFETs have captured most analog and power.

3. Simplified operation

The essential usefulness of a transistor comes from its ability to use a small signal applied between one pair of its terminals to control a much larger signal at another pair of terminals. This property is called gain. A transistor can control its output in proportion to the input signal; that is, it can act as an amplifier. Alternatively, the transistor can be used to turn current on or off in a circuit as an electrically controlled switch, where the amount of current is determined by other circuit elements.

The two types of transistors have slight differences in how they are used in a circuit. A bipolar transistor has terminals labeled base, collector, and emitter. A small current at the base terminal (that is, flowing from the base to the emitter) can control or switch a much larger current between the collector and emitter terminals. For a field-effect transistor, the terminals are labeled gate, source, and drain, and a voltage at the gate can control a current between source and drain.

The image to the right represents a typical bipolar transistor in a circuit. Charge will flow between emitter and collector terminals depending on the current in the base. Since internally the base and emitter connections behave like a semiconductor diode, a voltage drop develops between base and emitter while the base current exists. The amount of this voltage depends on the material the transistor is made from, and is referred to as V_{BE} .

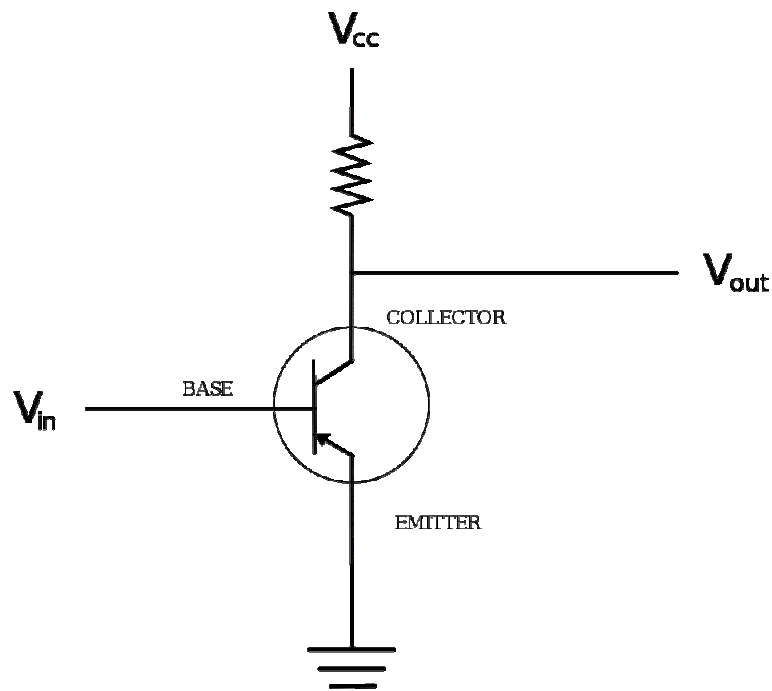


Figure 2. Simple circuit to show the labels of a bipolar transistor.

Transistor as a switch

Transistors are commonly used as electronic switches, for both high power applications including switched-mode power supplies and low power applications such as logic gates.

In a grounded-emitter transistor circuit, such as the light-switch circuit shown, as the base voltage rises the base and collector current rise exponentially, and the collector voltage drops because of the collector load resistor. The relevant equations:

$$V_{RC} = I_{CE} \times RC, \text{ the voltage across the load (the lamp with resistance } RC)$$

$$V_{RC} + V_{CE} = V_{CC}, \text{ the supply voltage shown as } 6V$$

If V_{CE} could fall to 0 (perfect closed switch) then I_c could go no higher than V_{CC} / RC , even with higher base voltage and current. The transistor is then said to be saturated. Hence, values of input voltage can be chosen such that the output is either completely off or completely on. The transistor is acting as a switch, and this type of operation is common in digital circuits where only "on" and "off" values are relevant.

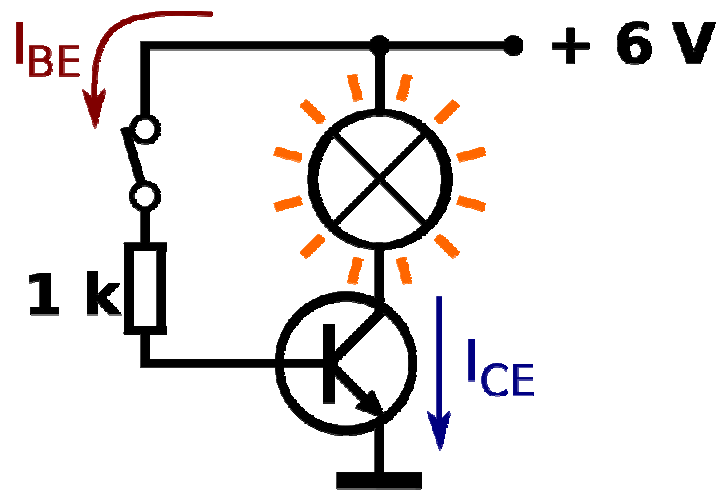


Figure 3. BJT used as an electronic switch, in grounded-emitter configuration.

Transistor as an amplifier

The common-emitter amplifier is designed so that a small change in voltage in (V_{in}) changes the small current through the base of the transistor and the transistor's current amplification combined with the properties of the circuit mean that small swings in V_{in} produce large changes in V_{out} .

Various configurations of single transistor amplifier are possible, with some providing current gain, some voltage gain, and some both.

From mobile phones to televisions, vast numbers of products include amplifiers for sound reproduction, radio transmission, and signal processing. The first discrete transistor audio amplifiers barely supplied a few hundred milliwatts, but power and audio fidelity gradually increased as better transistors became available and amplifier architecture evolved. Modern transistor audio amplifiers of up to a few hundred watts are common and relatively inexpensive.

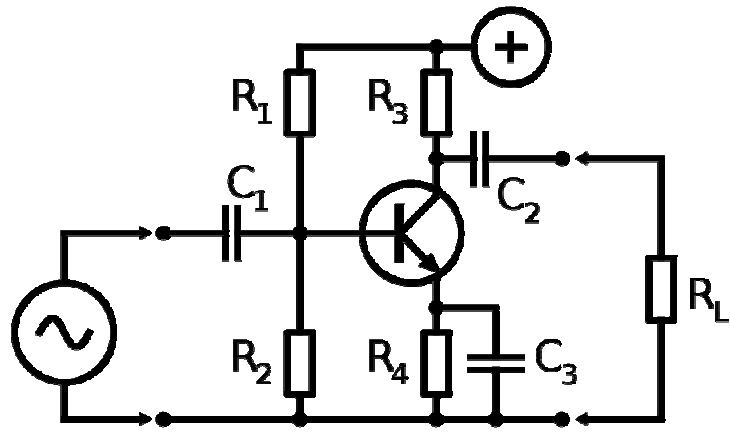


Figure 4. Amplifier circuit, standard common-emitter configuration.

4. Comparison with vacuum tubes

Prior to the development of transistors, vacuum (electron) tubes (or in the UK "thermionic valves" or just "valves") were the main active components in electronic equipment.

Advantages

The key advantages that have allowed transistors to replace their vacuum tube predecessors in most applications are:

1. Small size and minimal weight, allowing the development of miniaturized electronic devices.
2. Highly automated manufacturing processes, resulting in low per-unit cost.
3. Lower possible operating voltages, making transistors suitable for small, battery-powered applications.
4. No warm-up period for cathode heaters required after power application.
5. Lower power dissipation and generally greater energy efficiency.
6. Higher reliability and greater physical ruggedness.
7. Extremely long life. Some transistorized devices have been in service for more than 50 years.
8. Complementary devices available, facilitating the design of complementary-symmetry circuits, something not possible with vacuum tubes.
9. Insensitivity to mechanical shock and vibration, thus avoiding the problem of microphonics in audio applications.

Limitations

1. Silicon transistors do not operate at voltages higher than about 1,000 volts (SiC devices can be operated as high as 3,000 volts). In contrast, electron tubes have been developed that can be operated at tens of thousands of volts.
2. High power, high frequency operation, such as that used in over-the-air television broadcasting, is better achieved in electron tubes due to improved electron mobility in a vacuum.
3. Silicon transistors are much more vulnerable than electron tubes to an electromagnetic

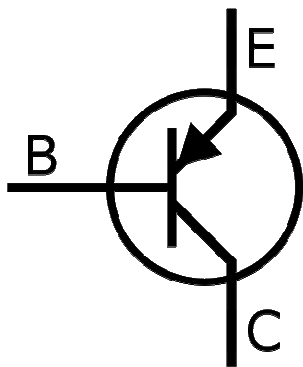
pulse generated by a high-altitude nuclear explosion.

5. Types

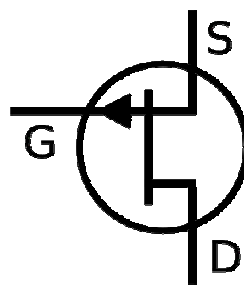
Transistors are categorized by

1. Semiconductor material: germanium, silicon, gallium arsenide, silicon carbide, etc.
2. Structure: BJT, JFET, IGFET (MOSFET), IGBT, "other types"
3. Polarity: NPN, PNP (BJTs); N-channel, P-channel (FETs)
4. Maximum power rating: low, medium, high
5. Maximum operating frequency: low, medium, high, radio frequency (RF), microwave
(The maximum effective frequency of a transistor is denoted by the term f_T , an abbreviation for "frequency of transition". The frequency of transition is the frequency at which the transistor yields unity gain).
6. Application: switch, general purpose, audio, high voltage, super-beta, matched pair
7. Physical packaging: through hole metal, through hole plastic, surface mount, ball grid array, power modules
8. Amplification factor h_{fe} (transistor beta)

Thus, a particular transistor may be described as silicon, surface mount, BJT, NPN, low power, high frequency switch.



PNP



P-channel

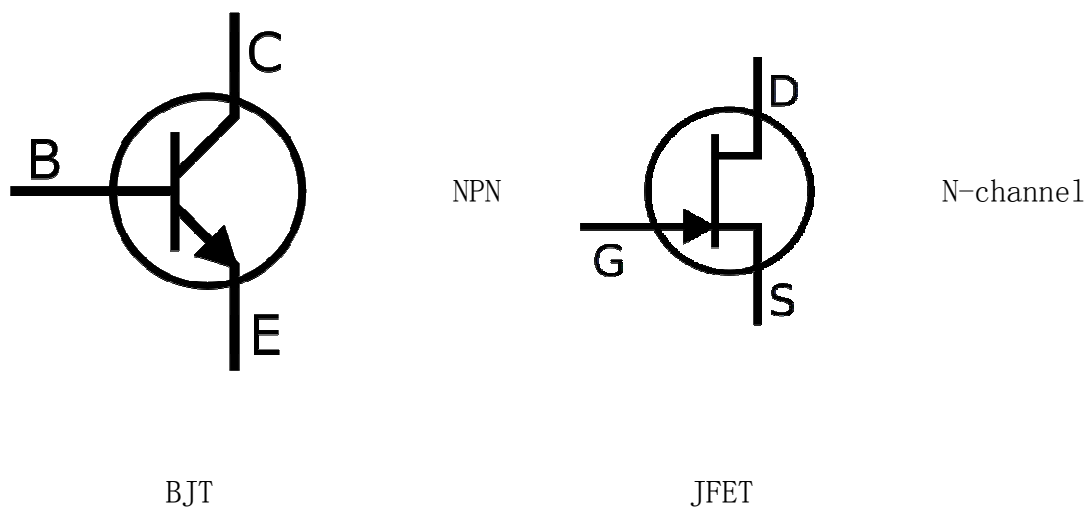


Figure 5. BJT and JFET symbols

Bipolar junction transistor

Bipolar transistors are so named because they conduct by using both majority and minority carriers. The bipolar junction transistor (BJT), the first type of transistor to be mass-produced, is a combination of two junction diodes, and is formed of either a thin layer of p-type semiconductor sandwiched between two n-type semiconductors (an n-p-n transistor), or a thin layer of n-type semiconductor sandwiched between two p-type semiconductors (a p-n-p transistor). This construction produces two p-n junctions: a base-emitter junction and a base-collector junction, separated by a thin region of semiconductor known as the base region (two junction diodes wired together without sharing an intervening semiconducting region will not make a transistor).

The BJT has three terminals, corresponding to the three layers of semiconductor - **an emitter, a base, and a collector**. It is useful in amplifiers because the currents at the emitter and collector are controllable by a relatively small base current." In an NPN transistor operating in the active region, the emitter-base junction is forward biased (electrons and holes recombine at the junction), and electrons are injected into the base region. Because the base is narrow, most of these electrons will diffuse into the reverse-biased (electrons and holes are formed at, and move away from the junction) base-collector junction and be swept into the

collector; perhaps one-hundredth of the electrons will recombine in the base, which is the dominant mechanism in the base current. By controlling the number of electrons that can leave the base, the number of electrons entering the collector can be controlled. Collector current is approximately β (common-emitter current gain) times the base current. It is typically greater than 100 for small-signal transistors but can be smaller in transistors designed for high-power applications.

Unlike the FET, the BJT is a low-input-impedance device. Also, as the base-emitter voltage (V_{be}) is increased the base-emitter current and hence the collector-emitter current (I_{ce}) increase exponentially according to the Shockley diode model and the Ebers-Moll model. Because of this exponential relationship, the BJT has a higher transconductance than the FET.

Bipolar transistors can be made to conduct by exposure to light, since absorption of photons in the base region generates a photocurrent that acts as a base current; the collector current is approximately β times the photocurrent. Devices designed for this purpose have a transparent window in the package and are called phototransistors.

Field-effect transistor

The field-effect transistor (FET), sometimes called a unipolar transistor, uses either electrons (in N-channel FET) or holes (in P-channel FET) for conduction. The four terminals of the FET are named source, gate, drain, and body (substrate). On most FETs, the body is connected to the source inside the package, and this will be assumed for the following description.

In FETs, the drain-to-source current flows via a conducting channel that connects the source region to the drain region. The conductivity is varied by the electric field that is produced when a voltage is applied between the gate and source terminals; hence the current flowing between the drain and source is controlled by the voltage applied between the gate and source. As the gate-source voltage (V_{gs}) is increased, the drain-source current (I_{ds})

increases exponentially for V_{gs} below threshold, and then at a roughly quadratic rate ($I_D \propto (V_{gs} - V_T)^2$) (where V_T is the threshold voltage at which drain current begins) in the "space-charge-limited" region above threshold. A quadratic behavior is not observed in modern devices, for example, at the 65 nm technology node. For low noise at narrow bandwidth the higher input resistance of the FET is advantageous.

FETs are divided into two families: junction FET (JFET) and insulated gate FET (IGFET). The IGFET is more commonly known as a metal–oxide–semiconductor FET (MOSFET), reflecting its original construction from layers of metal (the gate), oxide (the insulation), and semiconductor. Unlike IGFETs, the JFET gate forms a PN diode with the channel which lies between the source and drain. Functionally, this makes the N-channel JFET the solid state equivalent of the vacuum tube triode which, similarly, forms a diode between its grid and cathode. Also, both devices operate in the depletion mode, they both have a high input impedance, and they both conduct current under the control of an input voltage.

Metal–semiconductor FETs (MESFETs) are JFETs in which the reverse biased PN junction is replaced by a metal–semiconductor Schottky-junction. These, and the HEMTs (high electron mobility transistors, or HFETs), in which a two-dimensional electron gas with very high carrier mobility is used for charge transport, are especially suitable for use at very high frequencies (microwave frequencies; several GHz).

Unlike bipolar transistors, FETs do not inherently amplify a photocurrent. Nevertheless, there are ways to use them, especially JFETs, as light-sensitive devices, by exploiting the photocurrents in channel–gate or channel–body junctions.

FETs are further divided into depletion-mode and enhancement-mode types, depending on whether the channel is turned on or off with zero gate-to-source voltage. For enhancement mode, the channel is off at zero bias, and a gate potential can "enhance" the conduction. For depletion mode, the channel is on at zero bias, and a gate potential (of the opposite polarity) can "deplete" the channel, reducing conduction. For either mode, a more positive gate

voltage corresponds to a higher current for N-channel devices and a lower current for P-channel devices. Nearly all JFETs are depletion-mode as the diode junctions would forward bias and conduct if they were enhancement mode devices; most IGFETs are enhancement-mode types.

6. Construction

The first BJTs were made from germanium (Ge). Silicon (Si) types currently predominate but certain advanced microwave and high performance versions now employ the compound semiconductor material gallium arsenide (GaAs) and the semiconductor alloy silicon germanium (SiGe). Single element semiconductor material (Ge and Si) is described as elemental.

Rough parameters for the most common semiconductor materials used to make transistors are given in the table below; it must be noted that these parameters will vary with increase in temperature, electric field, impurity level, strain, and sundry other factors:

The junction forward voltage is the voltage applied to the emitter-base junction of a BJT in order to make the base conduct a specified current. The current increases exponentially as the junction forward voltage is increased. The values given in the table are typical for a current of 1 mA (the same values apply to semiconductor diodes). The lower the junction forward voltage the better, as this means that less power is required to "drive" the transistor. The junction forward voltage for a given current decreases with increase in temperature. For a typical silicon junction the change is $-2.1 \text{ mV}/^\circ\text{C}$.

The density of mobile carriers in the channel of a MOSFET is a function of the electric field forming the channel and of various other phenomena such as the impurity level in the channel. Some impurities, called dopants, are introduced deliberately in making a MOSFET, to control the MOSFET electrical behavior.

The electron mobility and hole mobility columns show the average speed that electrons and holes diffuse through the semiconductor material with an electric field of 1 volt per meter applied across the material. In general, the higher the electron mobility the speedier the transistor. The table indicates that Ge is a better material than Si in this respect. However, Ge has four major shortcomings compared to silicon and gallium arsenide:

1. Its maximum temperature is limited;

2. it has relatively high leakage current;
3. it cannot withstand high voltages;
4. it is less suitable for fabricating integrated circuits.

Because the electron mobility is higher than the hole mobility for all semiconductor materials, a given bipolar NPN transistor tends to be swifter than an equivalent PNP transistor type. GaAs has the highest electron mobility of the three semiconductors. It is for this reason that GaAs is used in high frequency applications. A relatively recent FET development, the high electron mobility transistor (HEMT), has a heterostructure (junction between different semiconductor materials) of aluminium gallium arsenide (AlGaAs)-gallium arsenide (GaAs) which has twice the electron mobility of a GaAs-metal barrier junction. Because of their high speed and low noise, HEMTs are used in satellite receivers working at frequencies around 12 GHz.

Max. junction temperature values represent a cross section taken from various manufacturers' data sheets. This temperature should not be exceeded or the transistor may be damaged.

Al-Si junction refers to the high-speed (aluminum-silicon) semiconductor-metal barrier diode, commonly known as a Schottky diode. This is included in the table because some silicon power IGFETs have a parasitic reverse Schottky diode formed between the source and drain as part of the fabrication process. This diode can be a nuisance, but sometimes it is used in the circuit.

Packaging

Transistors come in many different packages (semiconductor packages) (see images). The two main categories are through-hole (or leaded), and surface-mount, also known as surface mount device (SMD). The ball grid array (BGA) is the latest surface mount package (currently only for large transistor arrays). It has solder "balls" on the underside in place of leads. Because they are smaller and have shorter interconnections, SMDs have better high frequency characteristics but lower power rating.

Transistor packages are made of glass, metal, ceramic, or plastic. The package often dictates the power rating and frequency characteristics. Power transistors have larger packages that can be clamped to heat sinks for enhanced cooling. Additionally, most power transistors have the collector or drain physically connected to the metal can/metal plate. At the other extreme, some surface-mount microwave transistors are as small as grains of sand.

Often a given transistor type is available in sundry packages. Transistor packages are mainly standardized, but the assignment of a transistor's functions to the terminals is not: other transistor types can assign other functions to the package's terminals. Even for the same transistor type the terminal assignment can vary (normally indicated by a suffix letter to the part number, q.e. BC212L and BC212K).

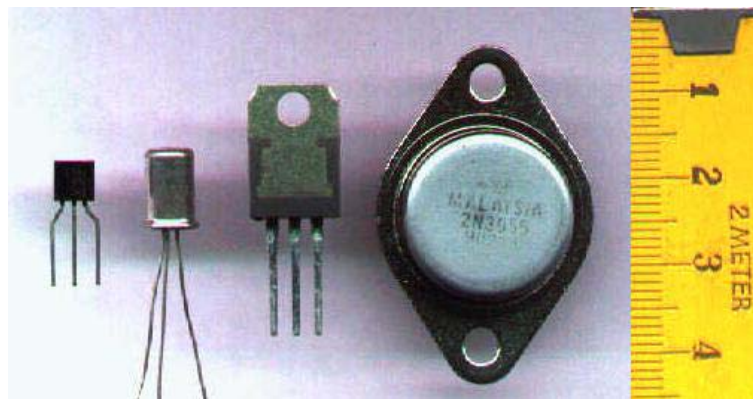


Figure 6. Through-hole transistors (tape measure marked in centimetres)