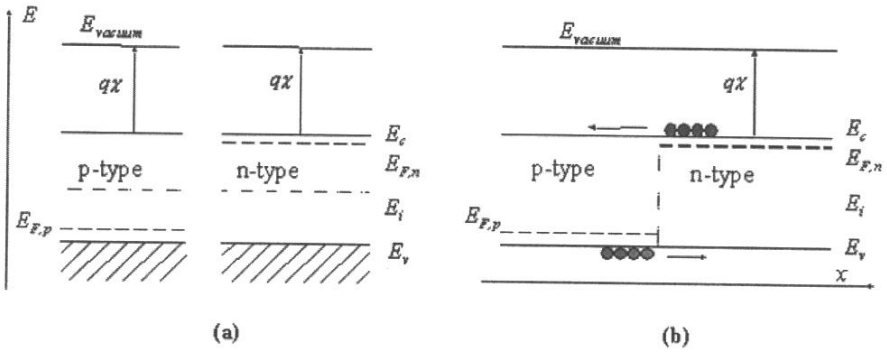
Lecture 4: Transistor physics

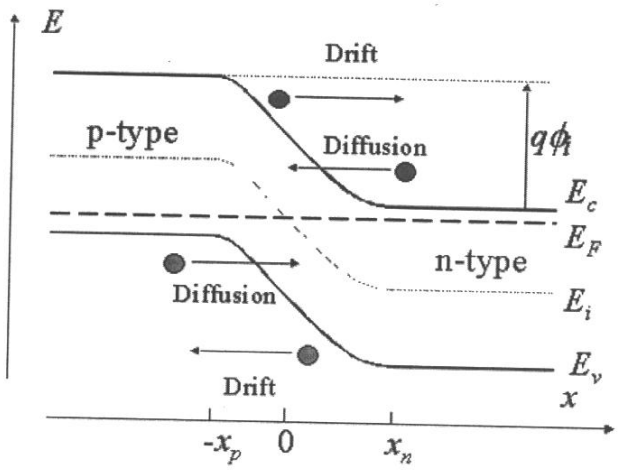
p-n junction: consists of two semiconductor regions with opposite doping type as shown in the figure. The region on the left is *p*-type with an acceptor density N_a, while the region on the right is *n*-type with a donor density N_d. Dopants are assumed to be shallow, so that the electron (hole) density in the *n*-type (*p*-type) region is approximately equal to the donor (acceptor) density. The junction is biased with a voltage V_a as shown in the figure. The junction is forward-biased if a positive voltage is applied to the *p*-doped region and reversed-biased if a negative voltage is applied to the *p*-doped region. The contact to the *p*-type region is also called the anode, while the contact to the n-type region is called the cathode, in reference to the anions or positive carriers and cations or negative carriers in each of these regions.



Energy band diagram of a p-n junction (a) before and (b) after merging the n-type and p-type regions.

To reach thermal equilibrium, electrons/holes close to the metallurgical junction diffuse across the junction into the p-type/n-type region where hardly any electrons/holes are present. This process leaves the ionized donors (acceptors) behind, creating a region around the junction, which is depleted of mobile carriers. We call this region the depletion region, extending from $x = -x_p$ to $x = x_n$. The charge due to the ionized donors and acceptors causes an electric field, which in turn causes a drift of carriers in the opposite direction. The diffusion of carriers continues until the drift current balances the diffusion current, thereby reaching thermal equilibrium as indicated by a constant Fermi energy. This situation is shown in the next figure. While in thermal equilibrium no external voltage is applied between the n-type and p-type material, there is an internal potential, ϕ_i, which is caused by the workfunction difference between the n-type and p-type semiconductors, called the built-in potential.

Lecture 4: Transistors (cont.)



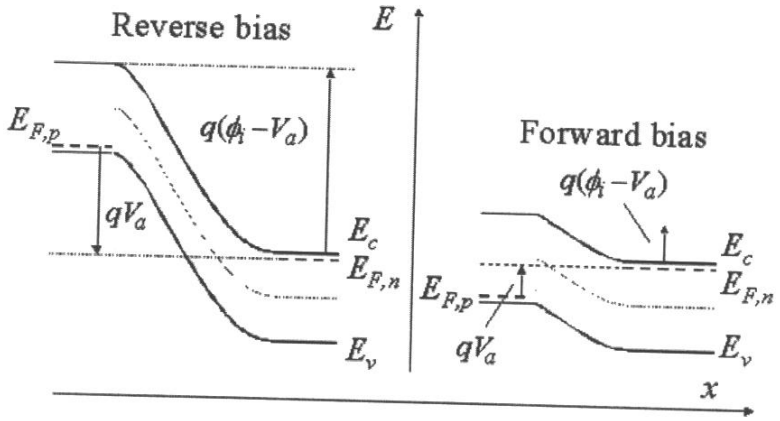
p-n thermal equilibrium at the Fermi energy

The built-in potential in a semiconductor equals the potential across the depletion region in thermal equilibrium. Since thermal equilibrium implies that the Fermi energy is constant throughout the p-n diode, the built-in potential (times q) equals the difference in the Fermi energies, E_{Fn} and E_{Fp}.

Now consider a p-n diode with an applied bias voltage, V_a. A forward bias corresponds to applying a positive voltage to the anode (p-type region) relative to the cathode (n-type region). A reverse bias corresponds to a negative voltage applied to the cathode. Both bias modes are illustrated in the next figure. The applied voltage is proportional to the difference between the Fermi energy in the n-type and p-type quasi-neutral regions. As a negative voltage is applied, the potential across the semiconductor increases and so does the depletion layer width. As a positive voltage is applied, the potential across the semiconductor decreases and with it the depletion layer width. The total potential across the semiconductor equals the built-in potential minus the applied voltage, or

$$\phi = \phi_i - V_a$$

Lecture 4: Transistors (cont.)

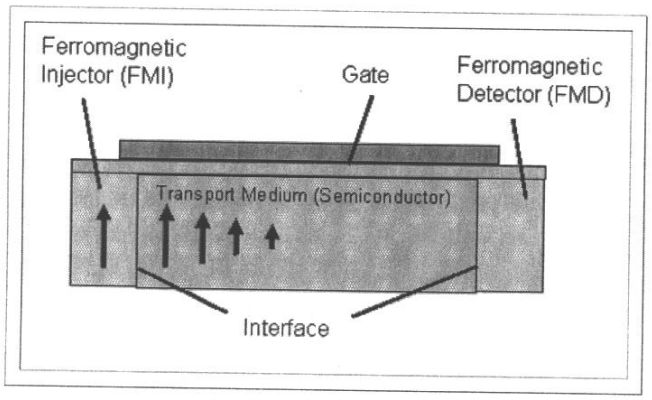


bipolar junction transistor: first solid-state amplifier element and started the solid-state electronics revolution. Bardeen, Brattain and Shockley at the Bell Laboratories invented it in 1948 as part of a post-war effort to replace vacuum tubes with solid-state devices. Solid-state rectifiers were already in use at the time and were preferred over vacuum diodes because of their smaller size, lower weight and higher reliability. A solid-state replacement for a vacuum triode was expected to yield similar advantages. The work at Bell Laboratories was highly successful and culminated in Bardeen, Brattain and Shockley receiving the Nobel Prize in 1956. Since then, the technology has progressed rapidly. The development of a planar process yielded the first circuits on a chip and for a decade, Bipolar transistor operational amplifiers and digital TTL circuits are the workhorses of any circuit designer. Bipolar transistors remain important devices for ultra-high-speed discrete logic circuits such as emitter coupled logic (ECL), power-switching applications and in microwave power amplifiers.

A bipolar junction transistor consists of two back-to-back p-n junctions, who share a thin common region with width, w_B. Contacts are made to all three regions, the two outer regions called the emitter and collector and the middle region called the base. ~~The structure of an NPN bipolar transistor is shown in the next figure (a).~~ The device is called "bipolar" since its operation involves both types of mobile carriers, electrons and holes which travel in opposite directions. Notice that the region barrer between the emitter and base is forward biased while the region barrier between the base and collector is reversed biased. When the collector fills with electrons, an increased current, and therefore load is produced on the base/collector resistor thereby producing conduction for a switch, or increased power for an amplifier.

Lecture 4: Transistors (cont.)

MRAM - Magnetoresistive RAM that depends on spintronics (the study of devices based on the electron spin polarity) by using spin transistors (see figure below). In the semiconducting region the electronic spins can be manipulated by the application of an electric field. The bit 1 and 0 states are realized when the injected spins are coherently rotated to be parallel or antiparallel to the magnetic moment of the device. The magnetic tunnel junctions are where the normal metal layer is replaced by a thin insulator. Such devices have demonstrated fast speed, high density, low power consumption, nonvolatility, and radiation hardness. They are promising replacements for the semiconductor RAM currently in use.

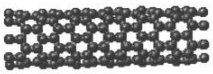


Schematic diagram of a spin transistor in MRAM.

The MRAM bit cell is a magnetic tunnel junction, which consists of two thin magnetic layers separated by an ultrathin layer (1 nm) of oxidized Al. The tunneling process itself is an example of a macroscopic quantum effect. The magnetoresistance is a result of s-d shell electron exchange which polarizes conduction electrons to be parallel with the layer magnetizations. The uniformity of the tunnel barrier is critical for well-defined resistance values and to minimize tunneling hotspots. The bit shape less than 1 angstrom thick and is typically elliptical, with a magnetic free layer of thickness 4-6 nm. The bit shape defines a shape anisotropy along with a switching field. The state of the bit (0 or 1) is programmed using current pulses that are nanoseconds in duration. The pulses are sent down conductive lines that are surrounded by a thin permeable magnetic film to enhance the generated field. A major challenge to error free programming is to minimize bit switching that is thermally activated. One way to do this during fabrication is to minimize this effect by using multilayer magnetic films to increase the energy barrier to magnetization reversal.

Lecture 4: Transistors (cont.)

NRAM – RAM made up of transistors formed by carbon molecules rolled up into "nanotubes" with a diameter around a billionth of a meter and only a few hundred nanometers long which effectively allows them to be insulating, semiconducting, or conducting material depending on the gyration of the carbon atoms within the tube. The tubes can function as a nanoscale storage device even after power is removed because the tubes will maintain their electric charge, making them a nonvolatile device. The huge advantage however over standard SRAM is in density and speed: nanotubes can be switched from on to off in half a nanosecond, and billions of tubes can exist in a single square millimeter. Comparable silicon-based memory cells require 50 nanoseconds to operate and are currently 100 times the size of a single nanotube. Note: present Pentium IV chips have 100 million transistors; NRAM allows the possibility for a new nano-microchip of the same size to have up to 100 billion bit transistors.



Nanotube depiction in NRAM.

Nanotubes themselves consist of a cylindrical array of carbon atoms whose diameter is only about 1 nanometre. The company Nantero now has memory chips consisting of billions of nanotubes, each a few hundred nanometres long, suspended from a silicon wafer. Another wafer sits about 100 nanometres below the first. Because the nanotubes that Nantero uses conduct electricity, a small electric charge at one point on the second wafer will draw several dozen nanotubes towards it. Once they are there, they stay there. That is because they are bound by Van der Waals forces—intermolecular bonds that do not depend on external power for their maintenance. (exercise – why?) An additional application of current, however, will release the nanotubes. This means that a group of a few dozen nanotubes can act as a memory element, storing a single bit (either a one or a zero) of the binary code that computers use to operate. If the connection between the wafers is live at a particular point, the bit represented is a one. If not, it is a zero. Because the wafers are so close together, those data can move rapidly from place to place. Nantero's new memory can read or write a bit in as little as half a nanosecond. The best RAM chips, by contrast, need ten nanoseconds to perform a similar operation. Nantero plans to market these memories within a year; the only complication is that existing fabrication techniques (like with MRAM) rely on standard semiconductor technology. To correct the orientation of tubes aligning in the wrong direction, for instance, Nantero uses an electron beam to zap the tubes that are not pointing in the right direction. That leaves only those that are hanging down towards the opposite wafer.

Lecture 4: Transistors (cont.)

FeRAM – It is a kind of nonvolatile RAM that can be used to replaced SRAM and DRAM (like MRAM or NRAM). Each cell comprises 2 MOSFETS with 2 ferroelectic capacitors (ferroelectrics have spontaneous electric polarization which can be controlled by an external electric field). The electrode of the ferroelectric capacitor corresponds to the collector of a PNP bipolar transistor, but the polarization is maintained without power such that bit information remains in storage. The capacitors RC constand is 200 ns, with a 11 x 7 um^2 cell. FeRAM is making its way into electrical appliances, digital cameras, and audio devices and optical storage (Kbit, Mbit, and Gbit storage needed respectively). Its advantage of course is that read and write access are faster than nonvolatile flash memory (100 ns vs 1 msec) while having the same cell size as other types of standard RAM. Its disadvantage is that it is 10 times slower than conventional SRAM, 2 times slower than conventional DRAM, and still uses a large cell (compared to say MRAM or NRAM).