Appendix B: Transistors

Of course, the transistor is the most important semiconductor device and has enabled essentially all of modern solid-state electronics. However, as a matter of history, electronics began with vacuum tubes. As indicated previously, in 1880 Thomas Edison discovered that a two-element vacuum tube exhibits asymmetric conduction of current, *viz.*, the so-called "Edison effect". Even so, it was not until the early twentieth century that the British physicist, John Ambrose Fleming, discovered that the Edison effect could be used to detect radio waves. Accordingly, Fleming developed and patented a two-element vacuum tube, which came to be known as the "diode". Subsequently, three-element vacuum tubes or "triodes" were developed in the first decade of the twentieth century by Lee de Forest and others. Most importantly, triodes enabled development of the first true electronic amplifiers resulting in great improvement of telephony as well as radio transmitters and receivers. Physically, a vacuum tube operates by biasing two electrodes, conventionally called the "filament" and the "plate". In operation, the filament is heated to a high temperature such that electrons can easily be thermionically emitted into the vacuum. Accordingly, if the plate is positive with respect to the filament, electrons are extracted and current flows. Conversely, if the plate is negative with respect to the filament, current does not flow. This is essentially the Edison effect. Moreover, as asserted above, a two-element tube is a diode; however, if a third electrode, called the "grid" (since it is usually a wire mesh or screen) is installed between the filament and plate, a bias voltage placed on the grid can modulate current flow. It is this "field effect" that allows a triode to operate as an electronic amplifier.

As early as 1925 it was recognized that field effect might also occur within crystalline solids. Indeed, the first patent for a *field effect transistor* (or FET) was filed in Canada by Austrian-Hungarian physicist, Julius Edgar Lilienfeld. Subsequently, in 1934 German physicist, Oskar Heil, also patented a field effect transistor of different design. Even so, no practical devices were ever built or tested. This is most likely a consequence of the lack of high purity semiconductor materials at that time. The development of high quality semiconductor crystals was motivated by the need during the Second World War for fast diodes, which were used in radars as a frequency mixer element in microwave receivers. Vacuum tubes were found to be too slow, but solid-state diodes fabricated from extremely pure germanium were found to be suitable. After the war, John Bardeen and Walter Brattain working under William Shockley at Bell Telephone Laboratories, succeeded in building the first operational "crystal triode", *i.e.*, transistor, in 1947. However, this was not a FET, but rather a *point-contact transistor*. Although commercialized by the Western Electric Company, point-contact transistors were unfortunately found to be too fragile and were soon replaced by the *junction transistor* invented by Shockley in 1948. Subsequently, the *junction field effect transistor* or JFET was also successfully fabricated. (Actually, Bardeen, Brattain, Shockley, and others were trying to fabricate a JFET when the point-contact transistor was discovered since the JFET more closely resembles Lilienfeld's original conception.)

Solid-state electronics was dominated by junction devices until the 1960's, but these subsequently have been supplanted by the *metal-oxide-semiconductor field effect transistor* or MOSFET. The first practical MOSFET was invented and patented in 1959 by Dawon Kahng and Martin M. (John) Atalla, again, at Bell Telephone Laboratories.

As the name suggests, this device combines an MOS capacitor with a *pn*-junction. Accordingly, MOSFET's are both structurally and operationally different from junction transistors. Within this context, devices are made by fabricating an insulating layer on the surface of a semiconductor containing a *pn*-junction which defines the conductive "channel". As for a simple MOS capacitor, a gate electrode is formed on the top of the insulator. Typically, the semiconductor is crystalline silicon and the insulator is thermally oxidized silica. As noted previously, for this material combination the density of localized electron traps at the Si/SiO_2 interface can be quite low. Consequently, well-made silicon MOSFET's are inherently free from trapping and scattering of carriers.

Bipolar Transistors

Both point-contact and junction transistors are *bipolar*, which means that electrons and holes are majority carriers in different parts of the device; hence, current flowing through the device is carried by both electrons and holes. Physically, a bipolar junction transistor or BJT consists of two "back-to-back" *pn*-junctions that are in close proximity such that depletion regions interact. Accordingly, there are two kinds of bipolar transistors, *viz.*, *npn* and *pnp*. These designations specify the arrangement of the interacting junctions. A schematic representation of an *npn* BJT is shown in the following figure:

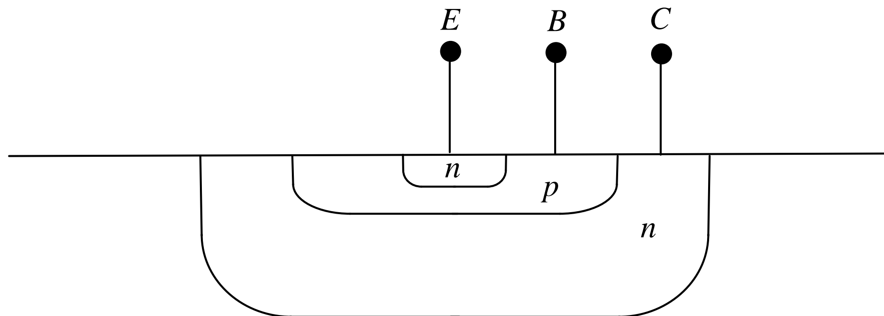

Fig. 71: Schematic of an *npn* BJT

Of course, a *pnp* is identical except that donor and acceptor doping is inverted. Here, *E*, *B*, and *C* denote *emitter*, *base*, and *collector* connections. (These designations descend from the original point-contact transistor for which in particular the "base" was a slab of semiconductor, *viz.*, germanium, to which "emitter" and "collector" point contacts were made.) Concomitantly, the two junctions are called "base-emitter" and "base-collector" junctions. In normal operation, for an *npn* transistor the base-emitter junction is forward biased (*i.e.*, base is positive with respect to the emitter) and the base-collector junction is reverse biased (*i.e.*, base is negative with respect to the emitter). A BJT is a current controlled device and in normal operation is governed by the equations:

$$I_E = I_0\left(e^{qV_{BE}/kT} - 1\right)$$

$$I_C = \alpha I_E = \left(\frac{\beta}{\beta+1}\right) I_E$$

$$I_B = (1-\alpha) I_E = \left(\frac{1}{\beta+1}\right) I_E$$

Here, α and β are identified as "common base" and "common emitter" current gains, respectively. Clearly, these two parameters are not independent and, as such, are determined by the physical characteristics of the transistor structure. The value of α is usually close to unity, *e.g.*, 0.980 to 0.998, which implies that β has a value between 49 and 499. Within this context, it is clear that I_E is determined by the diode equation for which V_{BE} is to be interpreted as the potential difference across the base-emitter junction. Typically, this is just the diffusion potential and, hence, is 0.5 to 0.7 volts in normal operation. (Obviously, I_0 remains defined as reverse saturation current just as in the case of a simple diode.) For completeness, if both junctions are forward biased, the transistor is said to be "saturated" and if both junctions are reversed biased the transistor is said to be "cut-off". In saturation, current flowing through the device is large and essentially independent of base current or V_{BE}. Conversely, in cut-off only very small leakage currents flow. It is possible to operate the device in inverted mode in which the base-collector junction is forward biased and the base-emitter junction is reverse biased. In principle if the device is exactly symmetric normal and inverted operation would have identical characteristics; however, as is evident from the figure, BJT's are generally asymmetric and inverted operation exhibits inferior characteristics.

Unipolar Transistors

In contrast to bipolar transistors all kinds of FET's are *unipolar*, which means that current is carried through the device either by electrons, *viz.*, *n*-channel, or holes, *viz.*, *p*-channel. In this regard, as asserted previously FET's resemble a vacuum tube triode, which are also unipolar devices, *i.e.*, current is carried by electrons. A schematic of a *p*-channel JFET is shown below:

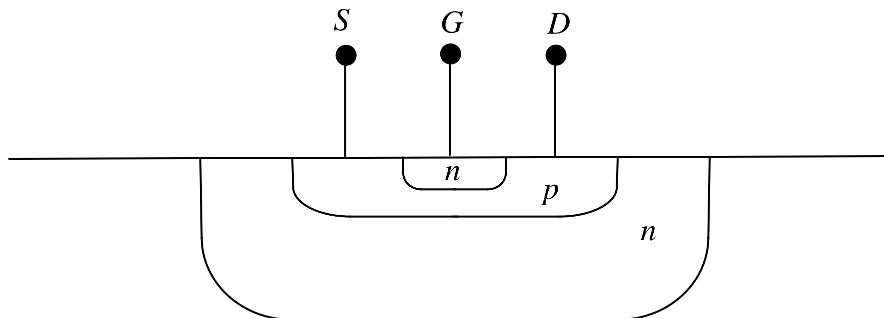

Fig. 72: Schematic of a *p*-channel JFET

Clearly, with the exception of external connections this figure is essentially identical to previous figure of an *npn* transistor and, thus, also consists of back-to-back junctions; however, operation is quite different. (In practice, the deep *n*-type region is reverse biased and, as such, serves to confine carriers, *viz.*, holes, to the channel.) Of course, in analogy to a BJT an *n*-channel JFET corresponds to inversion of donor and acceptor doping. As a matter of convention, *S*, *G*, and *D* denote *source*, *gate*, and *drain* connections. Concomitantly, if the gate is relatively unbiased and if the source is more positive than the drain, current can be expected to flow through the channel between source and drain. However, if the gate is biased positively with respect to the channel, *i.e.*, the gate-channel junction is reverse biased, then the depletion region expands and reduces effective conductivity of the channel. Indeed, if the channel is sufficiently thin, the depletion region can extend all the way across the channel, thus, effectively "pinching" it off so that very little current flows. In this case, the JFET is said to be "off". In general, this kind of operation is characteristic of a "depletion mode" FET.

Clearly, if the gate-channel junction is forward biased, an undesirably large current flows through the gate into the channel. Therefore, JFET's are generally not operated in this mode. However, if the gate is insulated as in a MOSFET, operation in "enhancement mode" is possible. Such a device is illustrated as follows:

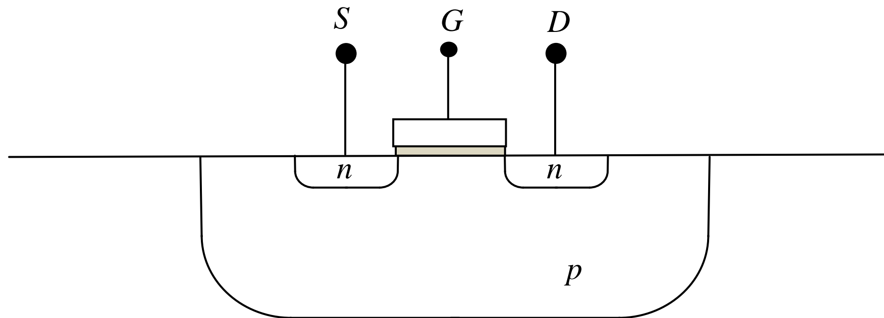

Fig. 73: Schematic of an enhancement mode *n*-channel MOSFET

Here, the light gray region is an insulator, typically formed of thermal oxide, although for very small devices other types of dielectrics may be used. Obviously, as asserted previously, a MOSFET is the combination of an MOS capacitor and *pn*-junctions. In any case, if the surface is accumulated or depleted under the MOS structure, very little current, *i.e.*, only leakage current, can flow between the source and drain. However, as the surface becomes inverted the surface becomes *n*-type and substantial current can flow. Indeed, in strong inversion the device becomes saturated in analogy to a BJT. Physically, source-drain current, I_D is determined by the bias voltage applied to the gate. Therefore, a FET is a voltage controlled device. In the case that the channel is accumulated or depleted, I_D is approximated by the expression:

$$I_D \cong I_{D0} e^{q(V_{GS}-V_t)/nkT}$$

Here, V_{GS} is the potential difference between gate and source connections and V_t is *threshold voltage*. In practice, V_t is determined by device structure and broadly corresponds to the potential associated with minimum capacitance of the MOS structure. Concomitantly, the ideality factor, n, corresponds to the formula:

$$n = 1 + \frac{C_s}{C_{ox}}$$

Of course, C_{ox} are C_s are oxide and depletion capacitances per unit area defined just as for a simple MOS capacitor. Naturally, the preceding expression is characteristic of cut-off and corresponds to $V_{GS} < V_t$. Accordingly, in normal operation, *i.e.*, $V_{GS} > V_t$ and $V_{DS} < (V_{GS} - V_t)$, I_D can be represented as follows:

$$I_D = \mu C_{ox} \frac{w}{l}\left((V_{GS} - V_t)V_{DS} - \frac{V_{DS}^2}{2}\right)$$

As might be expected, V_{DS} is the potential difference between drain and source connections, μ is effective carrier mobility, and w and l are surface width and length dimensions of the channel. Naturally, in saturation, *i.e.*, $V_{GS} > V_t$ and $V_{DS} \geq (V_{GS} - V_t)$, I_D becomes independent of V_{DS}, hence:

$$I_D = \mu C_{ox} \frac{w}{l}\left(\frac{(V_{GS} - V_t)^2}{2}\right)$$

This expression does not account for anomalies and is applicable only to ideal devices for which l is much greater than w.

For completeness, a depletion mode MOSFET is illustrated by the following figure, thus:

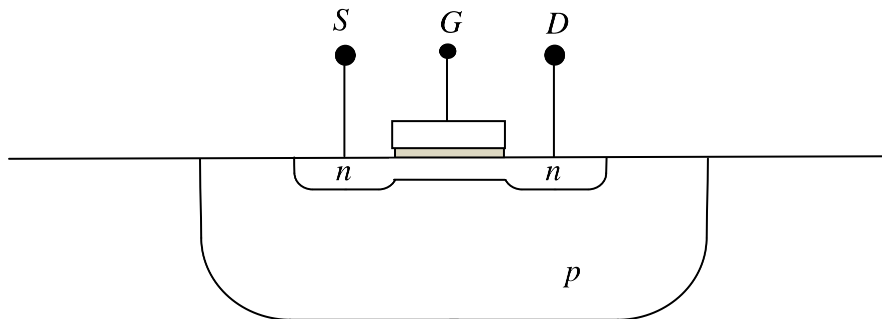

Fig. 74: Schematic of a depletion mode *n*-channel MOSFET

Clearly, operation of this device can be expected to be similar to an *n*-channel JFET, that is to say, if the gate is sufficiently negative with respect to the channel, the semiconductor surface is depleted and the channel becomes “pinched off”.

State-of-the-Art Devices

In conclusion, transistor structure has progressed far beyond the simple ideal structures described previously. The current state-of-the-art is embodied by devices with channel lengths of 32 to 22 nm or less. The former is still of planar design, but includes modifications such as a stressed channel and high-k gate dielectric. The reason for using stressed materials is to increase carrier mobility. Indeed, it has long been known that stress causes "splitting" of the band structure, which is characterized by sub-bands in which carriers have lower effective mass and, thus, higher mobility. As a practical matter, stress can be induced by epitaxial deposition of SiGe alloy on a pure silicon substrate. In this case, the SiGe alloy has a slightly larger lattice parameter than pure silicon, which results in biaxial stress in the deposited layer. In addition, stress may be modified and controlled by deposition of intrinsically stressed dielectric layers over the channel. As noted elsewhere, high-k gate dielectric is needed to reduce gate to channel leakage current which is inherent in ordinary thermal silica.

Reduction of channel length to 22 nm or less requires even more radical modification of transistor structure. In this case, a three dimensional "FinFET" structure has been introduced. Such a structure is characterized by a thin "blade" or "fin" of silicon, which is almost completely isolated from the substrate. The gate can then "wrap around" the fin on three sides, a so-called "tri-gate" structure, thus, resulting in more effective control of the channel.