Transistors – From Point Contact to Single Electron

D N Bose

This article presents a historical account of the development of electronic devices along with an appreciation for the physics of semiconductors. Starting at the turn of the century with the work of J C Bose on Galena, the article leads us to recent low temperature experiments at IBM to demonstrate single electron transistors.

D N Bose received his Ph D from J J Thomson Physical Laboratory, UK. He works at present in IIT, Kharagpur. His interests are in compound semi-conductor devices and opto-electronics.

Solid State Diode

The first solid state electron device was the metal-semiconductor rectifier discovered by Ferdinand Braun in 1874 even before the discovery of the electron! He found that mercury metal contacts on copper or iron sulphide gave non-linear current-voltage characteristics. In 1904 J C Bose obtained a U.S patent for point contact rectifiers on Galena (PbS). He found that the direction of rectification depended on the metal and also type of crystal. He thus called these positive or negative coherers. It is now known that these were due to *n*-type or *p*-type PbS. He also used these to detect microwave and optical radiation and proposed that the latter could form the basis of solar energy conversion – named Tejometer by him. No wonder Neville Mott said that J C Bose was sixty years ahead of his time!

Picard obtained a patent in 1907 for a point contact rectifier on silicon. These, in the form of tungsten whiskers on Si or PbS, were later widely used for the detection of radio broadcasts. The first large area rectifiers were based on copper-copper oxide (1926) and later evaporated selenium-iron.

The first theory that correctly predicted the direction of rectification was provided by Mott in 1939. This was after the

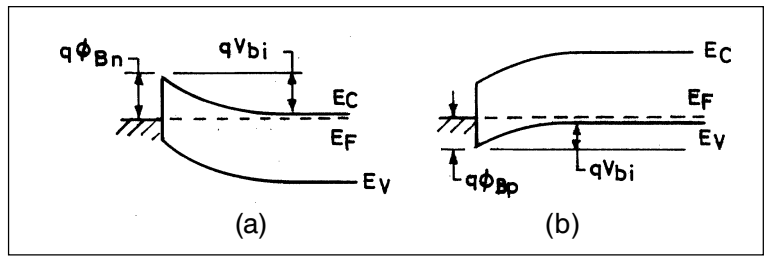


Figure 1. Energy band diagram of metal-semiconductor barrier on (a) n-type and (b) p-type semiconductor.

development of the energy band theory of solids by Wilson. Mott's theory was extended by Schottky and Spenke while Davydov developed his theory independently. All these theories had shortcomings as pointed out by Sah. The correct theory of metal-semiconductor diodes based on thermionic emission of majority carriers over a potential barrier was proposed by Hans Bethe in 1942. This model is illustrated in *Figure 1*.

Abbreviations

FET – Field Effect Transistor
HEMT – High Electron Mobility Transistor
MESFET – Metal Semiconductor FET
MOSFET – Metal Oxide Semiconductor FET
MODFET – Modulation Doped FET
SET – Single Electron Transistor

When an *n*-type semiconductor forms a contact with a metal having a higher work function, electrons are transferred from the semiconductor to the metal, thus forming a potential barrier which reduces the further flow of electrons. In equilibrium, the flow of these electrons is balanced by the flow of electrons from the metal to the semiconductor. These have to overcome the barrier height

$$q\phi_{Bn} = q\,(\phi_M - \chi) \qquad (1)$$

where ϕ_M = metal work function and χ = electron affinity of the semiconductor.

For a *p*-type semiconductor shown in *Figure 1b,* rectification requires a metal with low work function in which case the barrier height is given by

$$q\phi_{Bp} = E_g - q\,(\phi_M - \chi) \qquad (2)$$

This gives the interesting condition that for the same metal on *n*- and *p*-type semiconductors $q\phi_{Bn} + q\phi_{Bp} = E_g$, the band-gap of the semiconductor.

From Bethe's thermionic emission theory, (in analogy with Richardson's theory for emission into vacuum) it is found that the current density at a metal – semiconductor contact is

$$J=J_0[\exp\ (qV/nkT)-1] \qquad (3)$$

where J_0 = reverse saturation current= $A^{**}T^2 \exp(-q\phi_{Bn}/kT)$, A^{**} being the effective Richardson constant, V = applied voltage and n = ideality factor = 1 for ideal contact.

p–n Junction Diode

When one side of a semiconductor e.g. Si is doped with P (donors) to make it *n*-type and the other side with B (acceptors), a *p–n* junction is formed as shown in *Figure 2*. Electrons from the *n*-side diffuse to the *p*-side while holes diffuse in the opposite direction. This leaves positively charged donors on the *n*-side and negatively charged acceptors on the *p*-side unneutralised. Thus a space-charge layer is formed at the junction. In thermal equilibrium, majority carrier motion across the junction is impeded by the potential barrier while minority carriers can flow down hill. The potential barrier adjusts itself such that there is no net current across the junction at V=0.

When the *p*-side is made positive (forward bias), the barrier height decreases, majority carriers diffuse across the junction

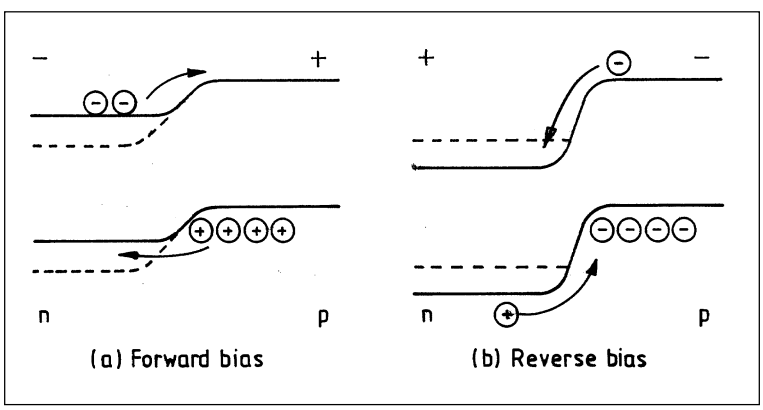


Figure 2. Energy band diagram of p-n junction under (a) forward bias and (b) reverse bias.

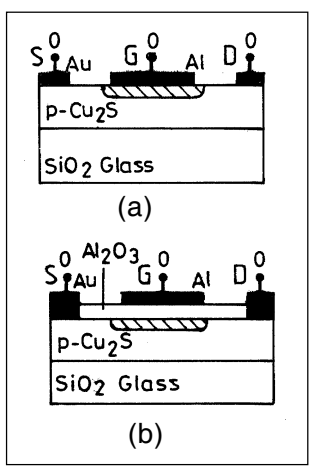
and so the current increases rapidly. When the p-side is made negative (reverse bias), the barrier height increases and so the majority carrier current is reduced. At large negative bias it is mainly the minority carriers that contribute to the reverse saturation current. The I-V characteristic of such a *p-n* junction was derived by Shockley and is the same as (3) with

$$J_0 = q[(D_n n_p / L_p) + (D_p p_n / L_n)] \quad (4)$$

where n_p and p_n are the minority carrier concentrations, D_n and D_p are the diffusion constants and L_n and L_p are the corresponding diffusion lengths.

The potential across the junction obviously depends on the doping concentrations on either side i.e. on the difference between the Fermi levels before junction formation. The *p–n* junction is a basic unit of most semiconductor devices. Light absorbed near a *p–n* junction creates electrons and holes, which because of the built-in field are separated, electrons going to the *n*-side and holes to the *p*-side. Thus an external voltage i.e. photovoltage is developed which forms the basis of solar cells.

Figure 3. (a) MESFET and (b) MOSFET structures patented by Lilienfeld in 1926 and 1930 respectively.



Field Effect Transistor

The first patent on a transistor was taken by Lilienfeld in 1926. The device is shown in *Figure 3a*. It consists of a thin film of the semiconductor Cu_2S deposited on glass. Two gold electrodes are placed on either side of the film, like the cathode and anode of vacuum tubes. The role of the control grid is played by an Al electrode, reducing the current when a depletion layer extends into the Cu_2S film. This structure is now known as a Metal Semiconductor Field Effect Transistor (MESFET).

The next structure that Lilienfeld invented in 1930 had a thin insulating layer of Al_2O_3 placed between the Al and Cu_2S

as shown in *Figure 3b*. This is the first Metal Oxide Semiconductor Field Effect Transistor (MOSFET).

Pearson and Shockley tried the same experiment in 1948 using a metal gate with air as dielectric on the semiconductor germanium. When a voltage was placed on the gate electrode creating a transverse electric field (hence the name Field Effect Transistor) a change in the current occurred between the source and the drain contacts. However this change was 1/10th of the expected value and decayed with time in seconds! Thus the MOSFET had a premature birth and had to be rediscovered in 1960 in the form known today by Kahng and Atalla.

To estimate the density of surface charges involved when a transverse electric field is applied, let us take a semiconductor with a resistivity ρ = 2.4 ohm-cm as forming one electrode. If the mobility of the carriers μ=2600 cm^2/V.s, then since $\rho=1/\sigma$ and $\sigma=ne\mu$, the concentration n=1.0$\times$10^{15}/cm^3.

The thickness T of the surface space charge layer in which the electrons are confined can be 100 nm. Then the surface concentration of electrons is $n \times T = 1.0 \times 10^{10}$/cm^3 (*Figure 4*).

Now if a transverse electric field of 30,000 V/cm is applied

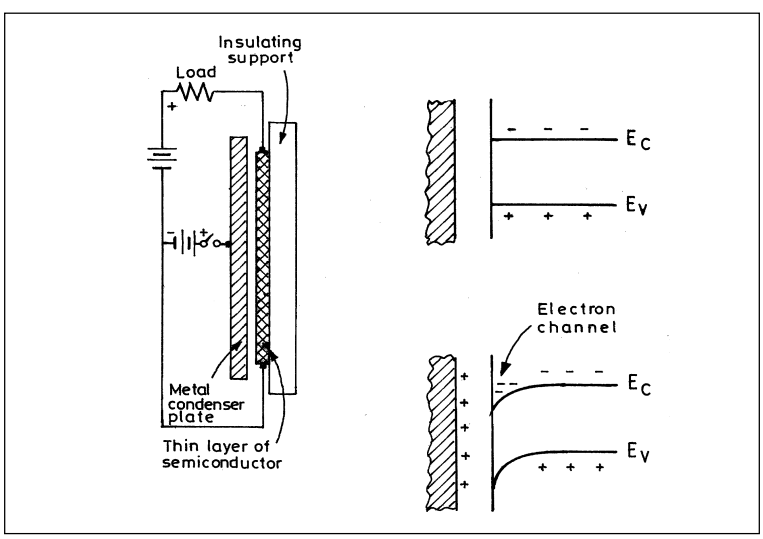


Figure 4. Diagram of induced surface charge in a semiconductor due to transverse external field.

The gate voltage required to turn the MOSFET *on* is called the *threshold voltage* and depends on the oxide thickness, metal work function, semiconductor doping concentration etc.

across a dielectric of relative permittivity ε_r=2, the charge induced is found to be 3.3 $\times 10^{10}/cm^2$, three times the background charge present. Thus the surface conductivity of the semiconductor should increase by a factor of 4. Experimentally Shockley found that only 10% increase in conductivity occurred. This remained a puzzle at that time.

Surface States

Walter Brattain and John Bardeen at Bell Telephone Laboratories were looking at rectification at metal point contacts made on semiconductors – the purest substance available being poly-crystalline germanium. The barrier heights were measured with a series of metals with different work functions. Surprisingly the measured barrier heights were found to depend only slightly on the metal work function. Bardeen came to the conclusion that the surface of the semiconductor was playing a major role and acting as a reservoir of charge and neutralising the differences between metal work functions.

It was realised that at the surface there was an abrupt termination of the periodic potential and there were unsatisfied or 'dangling bonds' which could accept electrons in localised energy states called *surface states*. This explained the failure of Shockley's Field Effect Transistor – the charge induced by the external field was being trapped in the surface states in times of the order of seconds and hence the small transient change in conductance.

MOSFET

The key to the development of the first successful MOSFET was the growth of SiO_2 as a dielectric insulator on Si. It was found that such thermally grown amorphous oxides gave very low density of states ($<10^{10}/cm^2/eV$) at the Si–SiO_2 interface. This permitted the modulation of charge by a

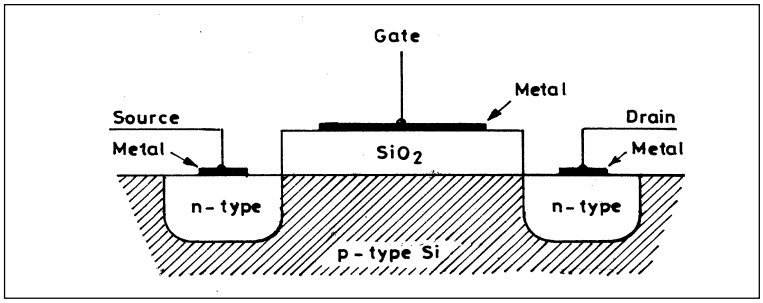


Figure 5. Structure of Si-SiO_2 MOSFET.

metal gate electrode on the surface. *Figure 5* shows the structure of a planar MOSFET as used in an integrated circuit. When the drain is made positive with respect to the source electrode, electrons can flow from source to drain through a narrow surface channel only when a sufficient positive gate voltage is applied. Since the bulk of the semiconductor is *p*-type, this requires an *n*-type channel or an inversion layer to form on the surface. The gate voltage required to turn the MOSFET *on* is called the *threshold voltage* and depends on the oxide thickness, metal work function, semiconductor doping concentration etc.

The figure shows a *n*-channel MOS (NMOS), while a complementary device fabricated on a *n*-type substrate is a PMOS. These are the basic elements of switches and memories in ICs. They have advantages over the junction transistors of smaller size, like higher device density/area, fewer fabrication steps and lower power consumption. A combination of an NMOS and PMOS constitute a CMOS (Complementary MOS) which can act as a flip-flop, consuming power only during switching, since at a given gate voltage only one of the MOS devices is on. These devices are used extensively in watches and other portable consumer applications.

Point Contact Transistor

The point contact transistor is mainly of historical interest as it was soon surpassed by the Junction Transistor. While examining metal point contacts on a semiconductor, Brattain

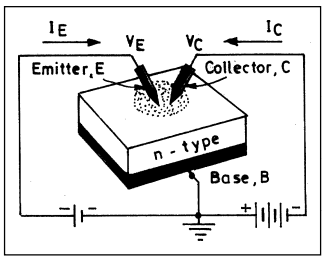


Figure 6. Point contact transistor on n-type semiconductor.

brought another contact in close proximity to the first one as shown in *Figure 6*. He observed a large change in the current flowing through the second contact when it was brought very close to the first one, with the first contact biased positive while the second one was biased negative. The polarities indicated are for an *n*-type semiconductor.

Bardeen realised that the key phenomenon was the injection of minority carriers i.e. holes by the injecting point contact (*emitter*) thus creating a modulation of the conductivity of the semiconductor. A small fraction of these holes recombined with the majority carriers, electrons, giving rise to a *base* current. The majority of the holes however diffused to the second metal contact, the *collector*. Modulation of the small base current caused a large change in the collector current, thus giving rise to current gain $\beta = \Delta I_C / \Delta I_B$. The first amplification by a semiconductor transistor action was discovered in December 1947, 50 years after the discovery of the electron.

Junction Transistor

Shockley developed the junction transistor in 1948 shortly after this. It overcame some of the problems associated with point contacts which were mechanically unreliable and also depended on the properties of surfaces. In its simplest form shown in *Figure 7* the junction transistor consists of two *p-n* junctions in a single crystal semiconductor separated by a distance much less than the diffusion length of the carriers. For a *n-p-n*

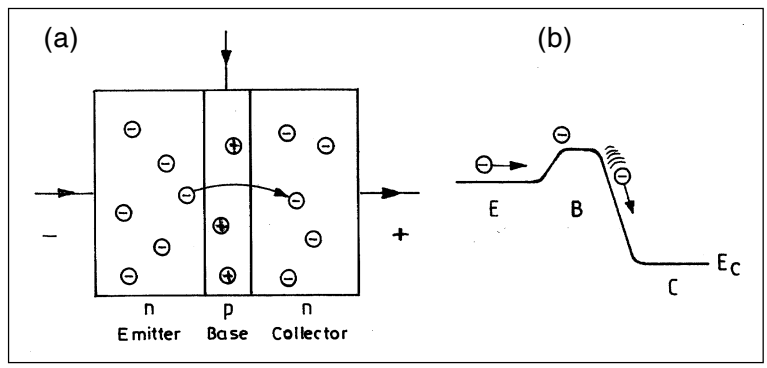


Figure 7. Junction Transistor structure and band diagram.

device as shown, application of a forward bias between emitter and base (B +ve) results in injection of electrons into the *p*-type base. Here the majority carriers are holes which can recombine with the injected electrons. Since the base-collector junction is reverse biased (C +ve) the electrons that diffuse across the base to the B-C junction are swept into the collector i.e. they fall to a lower potential.

The motion of the electrons through the base is due to diffusion. This is because of the large concentration gradient dn/dx across the base.

It is important to remember that the motion of the electrons through the base is due to *diffusion*. This is because of the large concentration gradient dn/dx across the base. Thus the electron current density is

$$J_n = +q.\ D_n\ \mathrm{d}n/\mathrm{d}x \qquad (5)$$

where the +ve sign is due to the negative charge of the electron.

From the Nernst–Einstein relation it is known that the diffusion coefficient D_n is related to the mobility by

$$D_n = (kT/q)\ \mu_n. \qquad (6)$$

For electrons in Si, $\mu = 1600\,\mathrm{cm}^2/\mathrm{V.s}$ and since $kT/q = 0.026\,\mathrm{eV}$ at $T = 300\,\mathrm{K}$, $D_n = 40\ \mathrm{cm}^2/\mathrm{s}$. For a base width of 1 μm = 10^{-4} cm, $n_E = 10^{17}/\mathrm{cm}^3$ and $n_B = 10^5/\mathrm{cm}^3$, $\mathrm{d}n/\mathrm{d}x = 10^{21}/\mathrm{cm}^4$.

Thus $J_n = 1.6 \times 10^{-19} \times 40 \times 10^{21} = 6.4 \times 10^3\ \mathrm{A/cm}^2$! This is due to an applied forward bias of only 0.6 V.

What is the requirement for the base width? Some of the electrons recombine with holes in the base thus giving rise to a small base current. However the majority of the electrons are able to get through to the B-C junction. An important parameter is the

$$\text{diffusion length } L_n = (D_n\ .\ \tau_n)^{1/2} \qquad (7)$$

Box 1. Gain

The common base gain α_0 of a Junction Transistor depends on :
(a) Emitter Injection Efficiency $\gamma = I_n/(I_n + I_p)$ where $I_n = E–B$ electron current and $I_p = B – E$ hole current,
(b) Base Transport Factor $\alpha_T = I_{nB}/(I_{nB}+I_{pB})$ and
(c) Collector Mutiplication Factor $M \rightarrow 1$ for $V_{CB} < 20$ V.
Common Emitter Current Gain $\alpha_0 = \gamma . \alpha_T . M = I_C/I_E = 0.950–0.998$.
Common Base Current Gain $= I_C/(I_E – I_C) = 19–499$.

Power gain occurs in a junction transistor because the $E–B$ junction being forward biased presents a low impedance whereas the $B–C$ junction being reverse biased has a high impedance. Thus the same current flowing through input and output circuits gives a Power Gain $= I_C^2 R_L/I_E^2 R_S$ which can be large even if $I_C = I_E$. Here R_L = Load Resistance and R_S = Source Resistance and Power Gain $= R_L/R_S$. The input ac signal is amplified by drawing the additional power from the dc supply.

where τ_n = minority carrier life-time. If $\tau_n = 10^{-5}$s, $L_n = 2 \times 10^{-2}$cm = 200 μm. Since the concentration of electrons in the base decreases as $n(x) = n_E \exp(-x/L_n)$, most of the electrons reach the B-C junction without recombination if $x << L_n$.

The earliest transistors had base widths of 30 –100 μm. The modern trend towards thinner base widths is to improve the frequency response i.e. get higher speed. For diffusive motion the transit time through an uniformly doped base is

$$\tau_t = W_B^2/2D_n \qquad (8)$$

where W_B = base width. Thus for $W_B = 1$ μm, $\tau_t = 1.25 \times 10^{-10}$s and cut-off frequency $f_c = 1/2\pi\tau_t = 1.25 \times 10^9$ Hz. Since electrons have higher mobility than holes $\mu_n >> \mu_p$, *n-p-n* devices are preferred to *p-n-p* for higher speed.

High Electron Mobility Transistor (HEMT)

The time taken by an electron to travel by *drift* between the

Semiconductor	E_g(300K) (eV)	μ_n (300K) ($\times 10^3$cm^2/V.s)	μ_n (77K) ($\times 10^3$cm^2/V.s)	v_p ($\times 10^7$cm/s)	v_{sat} ($\times 10^7$ cm/s)
Si	1.12	1.45	20	-	1.2
GaAs	1.42	8.5	200	2.0	0.6
InP	1.34	5.37	130	2.5	0.8
$In_{0.53}Ga_{0.47}As$	0.75	13.8	70	3.0	0.6

Table 1.

source and the drain of a MOSFET under low electric fields ($E < 10^3$ V/cm) is given by $t_t = L/v = L/\mu_n$. E and since $E = V/l$,

$$t_t = L^2/\mu_n.V \qquad (9)$$

where μ_n = electron mobility. For $L = 1\mu m$, μ_n=1150 cm^2/V.s and V=0.1 V, $t_t = 6.66 \times 10^{-11}$s.

Under high electric fields ($E>10^4$ V/cm) i.e. for short gate lengths L, the velocity of the electrons saturate at a constant value v_{sat} depending upon the material. In this case

$$t_t = L/v_{sat} \ . \qquad (10)$$

For $L = 1\mu$m and $v_{sat} = 1.2 \times 10^7$ cm/s, $t_t = 8.25 \times 10^{-12}$ s.

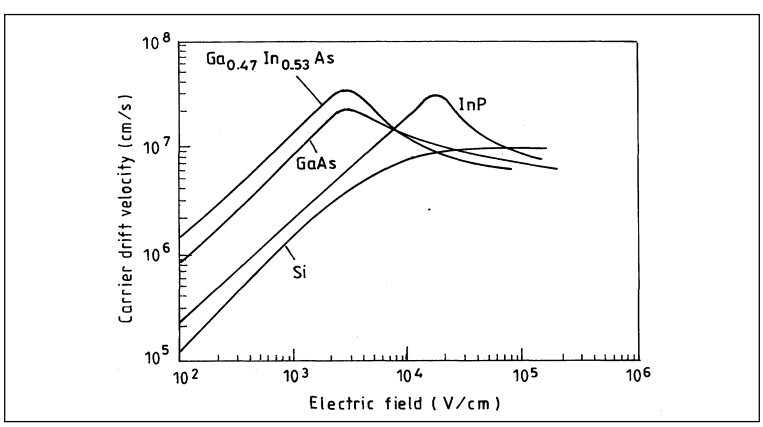


Figure 8. Velocity-field characteristic of semiconductors.

Box 2

From *Table I* it is found that the electron mobility μ_n increases considerably with decrease in temperature. This is because of reduced thermal scattering of the carriers by phonons. It was shown by Shockley that for acoustic phonon scattering, the temperature dependence of mobility is given by $\mu_L \sim AT^{-x}$ where $x = 1.5–2.3$.

At low temperatures electron mobility is controlled by ionised impurity concentration N_I such that $\mu_I \sim B.T^{+y}/N_I$ where y=2.3 for electrons in Si. Since $N_I = N_D^+ + N_A^-$, where N_D^+ = concentration of ionised donors and N_A^- = concentration of ionised acceptors, decrease in their concentration would result in higher mobility. However these donors and acceptors are necessary to provide the carriers. At a given temperature, if the scattering mechanisms are considered independent, the effective mobility μ_{eff} is given by Matthiessen's rule

$$1/\mu_{eff} = 1/\mu_I + 1/\mu_L.$$

Thus for high speed devices high values of μ_n and v_{sat} are desirable. The variation of velocity v with electric field E is shown in *Figure 8*. Values for some semiconductors are given in *Table 1*.

An ingenious way of overcoming this problem of impurity scattering called modulation doping was devised by Gossard at Bell Labs. A heterojunction, typically of GaAs-AlGaAs with the donors placed in the AlGaAs layer from which the electrons are transferred to the undoped GaAs layer is used here. This is because the conduction band in GaAs is at a lower energy than in AlGaAs. The electrons form a thin sheet near the interface due to the band bending at the hetero-junction thus forming a two dimensional (2D) electron gas. The electrons being physically separated by the barrier are not subject to ionised impurity scattering and can have very high mobility. Devices based on such structures are thus called High Electron Mobility Transistors (HEMT) or Modulation Doped FETs (MODFET). *Figure 9* shows how the mobility of electrons in such 2 dimensional layers at GaAs/AlGaAs heterojunctions have increased to 2×10^6 cm^2/V.s at low temperatures from 1982 to 1992.

The fabrication of such structures calls for advanced epitaxial growth techniques such as Molecular Beam Epitaxy (MBE) or Organometallic Vapour Phase Epitaxy (OMVPE) such that the junctions are atomically abrupt and the epitaxial layer purity is extremely high.

The structure of a HEMT or MODFET is shown in *Figure 10a*. A thin undoped AlGaAs spacer layer is grown between the doped AlGaAs and the undoped GaAs to prevent traces of the donors entering the GaAs. Electrons can flow between the n^+ source and the n^+ drain electrodes as in a MOSFET, being controlled by voltage applied to the gate electrode *(Figure 10b)*. When this is positive, electrons in GaAs are attracted towards the interface forming a sheet of electrons like a 2 dimensional electron gas. These electrons show a very high mobility since there are no scattering impurity centres in the GaAs. These GaAs/AlGaAs HEMT devices have operated at frequencies upto 20 GHz.

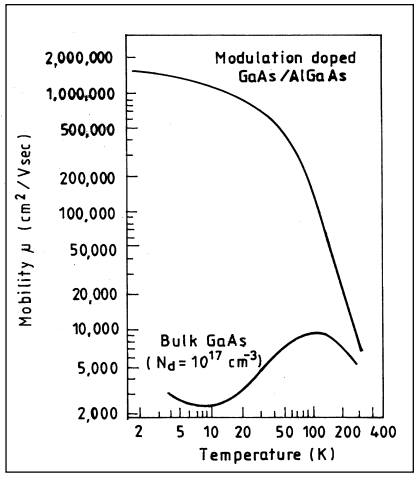
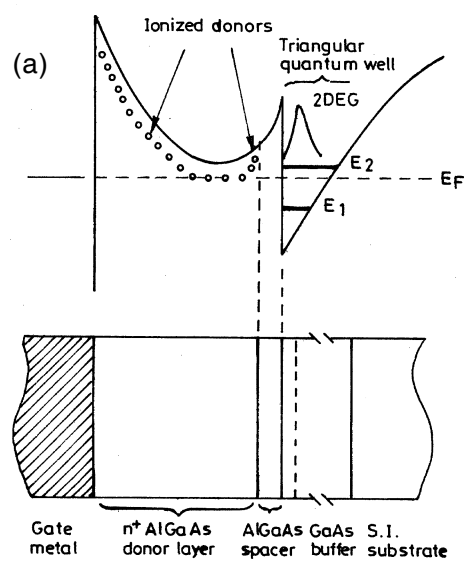


Figure 9 (right). Mobility of 2D electrons vs temperature.



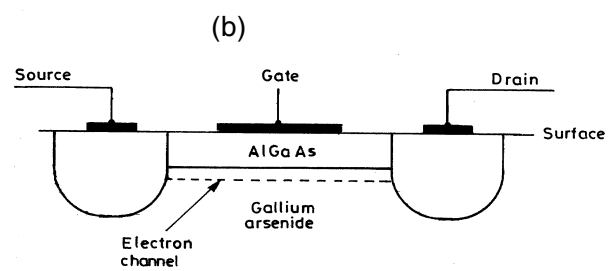


Figure 10. (a) Doping profile of MODFET (b) Structure of MODFET.

Suggested Reading) for a more detailed description.

Suggested Reading

- **William Shockley. *Electrons and Holes in Semiconductors*. D Van Nostrand, 1950.**
- **S M Sze. *Physics of Semiconductor Devices*. John Wiley, 1981.**
- **C T Sah. *Fundamentals of Solid-State Electronics*. World Scientific, 1991.**
- **M A Kastner. Single Electron Transistor. *Reviews of Modern Physics*. Vol. 64. No. 3. pp. 849–858, 1992.**
- **Richard Turton. *The Quantum Dot*. Oxford University Press, 1995.**

Table 1 shows that $In_{1-x}Ga_xAs$ has a higher mobility and saturation velocity than GaAs. This ternary compound for x=0.47 also has the same lattice parameter as InP and can be grown on it epitaxially without any strain. Energy band diagram similar to GaAs/AlGaAs is applicable to this heterojunction. HEMT devices made from InGaAs/InP have shown the highest speed operating at frequencies upto 350 GHz.

Single Electron Transistor

The gate length of a MOSFET has continously been reduced from 15 μm to 0.25 μm. This reduces the transit time of electrons from source to drain and hence increases the speed of the devices. The lower limit is determined by the lithographic process. At present UV sources with wavelength between 180 nm–230 nm are used to define the geometry and devices with L=0.25μm are currently state-of-the-art as in the case of pentium microprocessors. With electron-beam, X-ray or ion beam lithography even finer geometries are possible.

Consider the case where the gate length L as well as the gate width W have been reduced to 7 nm each and the oxide thickness is 200 A with ε_r=3.9. The capacitance of such a device $=0.86 \times 10^{-19}$ F. If 1 V is applied across this, the charge stored $Q=C.V=8.58\times 10^{-19}$ C=5 electrons. Thus it is seen that variation of the applied voltage can cause the storage or release of individual electrons. Such a single electron memory has been reported recently.

At these dimensions the discreteness of charge leads to discreteness of energy. The energy of a particle having charge Q and capacitance C is $E=Q^2/2C$. Adding an electron to such a particle takes an energy $q^2/2C$. The charge on the particle thus repels other electrons and prevents them from being collected. This phenomenon is known as *Coulomb blockade*.

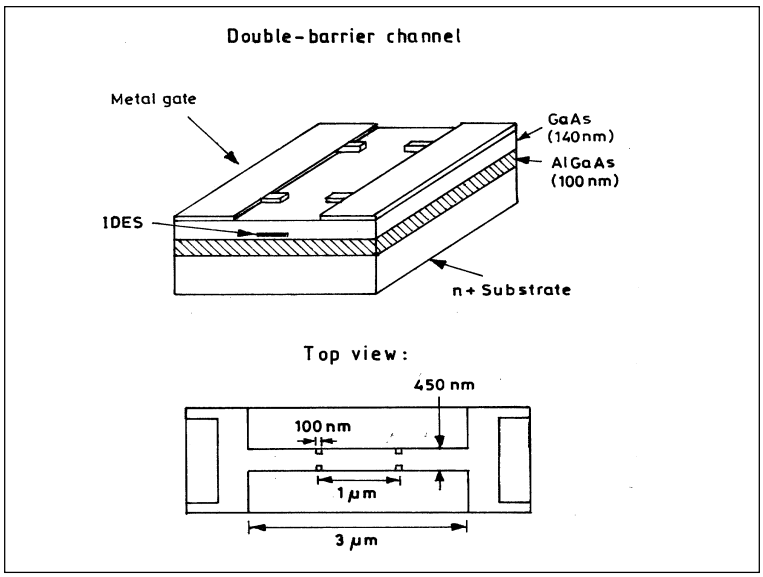


Figure 11. Structure of Single Electron Transistor (SET).

One of the first devices exhibiting transport of single electrons was fabricated by Meirav and others at IBM, USA. The structure shown in *Figure 11* consists of HEMT - like 2-D electron gas layer. On the surface a narrow metal gate is deposited and biased negative such that electrons are repelled from it. The gate geometry, defined by electron-beam lithography, has two narrow constrictions, the distance between which can be controlled. In the examples shown the values of L_0 are 0.6μm and 0.8μm. The electrons are thus confined between the narrow channels to produce a narrow 1-D electron gas.

When the gate voltage is varied the conductance *G* shows sharp

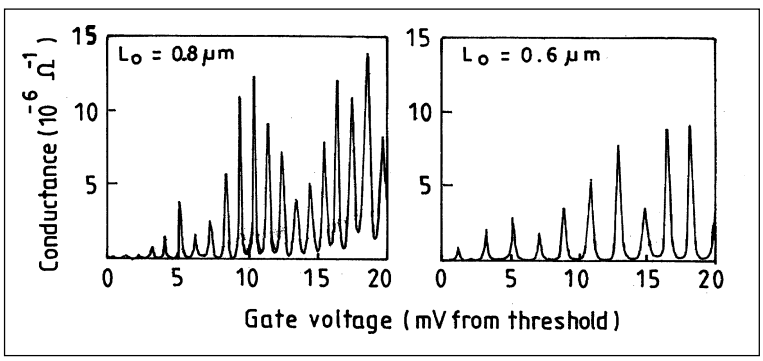


Figure 12. Conductance G vs gate voltage V_g for SET.

oscillations, the period being determined by L_0. Each period corresponds to the transfer of the same number of electrons to the region between the constrictions. For a shorter value of L_0, the capacitance is smaller and so the voltage to add one electron is larger as shown in *Figure 12*. The period of oscillations is simply q/C. To prove that a single electron is involved, the absolute value of capacitance must be found. The variation of potential between the constrictions has been calculated and the capacitance thus computed. From this it was shown that the variation of gate voltage indeed controls the addition of *one* electron. The transistor thus turns on and off again every time one electron is added. These experiments must be done at T=100 – 400 mK so that thermal fluctuations are reduced (i.e. kT is very small). Thus these devices are unlikely to find practical applications right now but they enhance our basic understanding of physics at mesoscopic levels. They may also help define an absolute quantum standard for current measurement. Readers are referred to the review by Kastner (see

Address for Correspondence
D N Bose
Semiconductor Division
Materials Science Centre
Indian Institute of Technology
Kharagpur 721 302, India