

## Physics Nobel Prize 2000 goes to Semiconductor Pioneers

*V Venkataraman*

The Nobel Prize in Physics for the year 2000 has been awarded to three scientists who laid the basic foundations for semiconductor devices used in information and communication technology today. According to the official press release issued by the Royal Swedish Academy of Sciences, one half of the prize has been jointly awarded to Zhores I Alferov (A F Ioffe Physico-Technical Institute, St. Petersburg, Russia) and Herbert Kroemer (University of California at Santa Barbara, USA) for “developing semiconductor heterostructures used in high-speed- and opto-electronics” while the other half of the prize goes to Jack S Kilby (Texas Instruments, Dallas, Texas, USA) for “his part in the invention of the integrated circuit”. In awarding this prize, the Nobel committee has rightly recognized the importance of basic research that has made the modern information technology revolution possible.

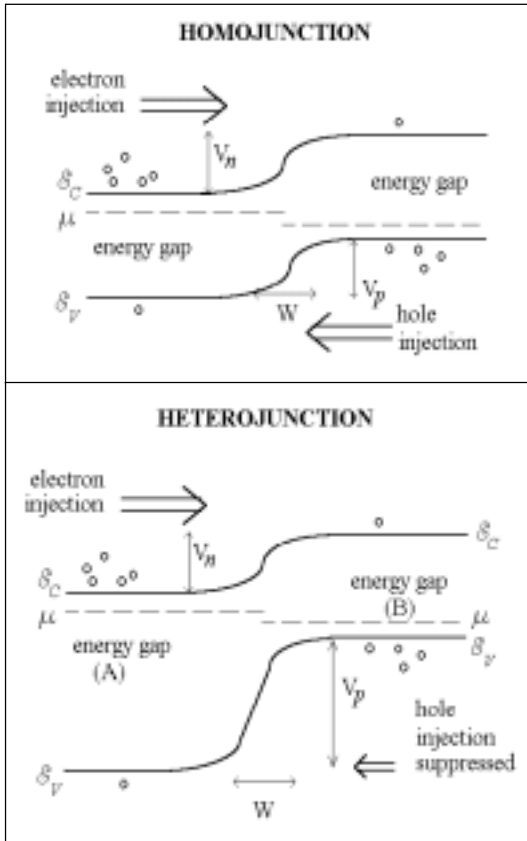
In order to understand the fundamental discoveries of these scientists and put their work in perspective, it is necessary to go back to the very beginning of semiconductor technology i.e. to 1947 when the point-contact transistor was first demonstrated at Bell Laboratories, USA by W Brattain and J Bardeen. Subsequently, in 1948, W Shockley from the

same group filed a patent on the use of *p-n* junctions, instead of point contacts, to inject and collect the carriers. This dramatically improved the performance of the device and the bipolar junction transistor was born. The group was awarded the Physics Nobel Prize in 1956.

The basic *p-n* junction consists of a piece of semiconductor, such as silicon or germanium, with a controlled amount of doping. One part of the material is doped ‘*n*-type’ (for example, by diffusing phosphorous atoms into silicon) while the other part is doped ‘*p*-type’ (for example, by diffusing boron atoms into silicon). These dopant or impurity atoms are ionized inside the semiconductor crystal to provide free charges that can carry current.

The *n*-type dopants provide negatively charged electrons to the conduction band while the *p*-type dopants remove electrons from the valence band to leave behind ‘holes’ which behave as positively charged carriers. Each region is typically doped uniformly with a concentration  $\sim 10^{16} - 10^{19}$  atoms/cc. However the carrier density is substantially reduced near the junction where free electrons and holes diffuse and recombine, leaving behind the ionized dopants. In equilibrium, a potential barrier ( $\sim 1$  eV) along with a substantial electric field builds up in this space-charge double layer or ‘depletion’ region, which typically has a width of  $\sim 1 \mu\text{m}$  near the junction. *Figure 1* shows the energy of the conduction and valence band edges





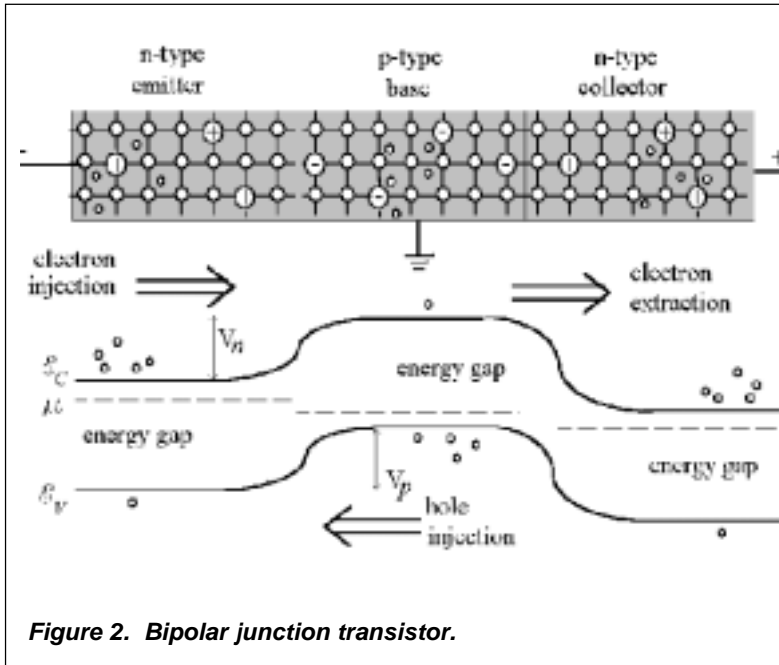
**Figure 1. Energy band diagrams for homo- and hetero- junctions under bias.**

plotted as a function of position. When a positive voltage is applied to the  $p$ -type region with respect to the  $n$ -type, carriers are forced across the potential barrier and they recombine to give rise to a substantial forward current. On the other hand, if the voltage is reversed, the current is limited to a small value by the supply of electrons in the  $p$ -region and holes in the  $n$ -region. We thus obtain the familiar diode-like rectifying characteristics.

When two  $p$ - $n$  junctions are brought close together, they interact to form the bipolar

junction transistor (*Figure 2*). As shown first by Shockley, the forward biased 'emitter-base' junction can now supply electrons to the reverse-biased 'collector-base' junction. The collector current is determined by the electron injection from the emitter and substantially independent of the collector voltage itself. This is the basic transistor action which can be used to build circuits that switch, oscillate or amplify electronic signals. Note that the current through the collector-base junction is very large, although it is reverse biased, due to the proximity of the forward-biased emitter-base junction. If the two junctions are separated by more than the electron diffusion length in the base (typically a few microns), then all injected electrons will recombine with holes inside the base and no collector current results. To obtain sufficient coupling, the base width is typically a micron or smaller. Ignoring recombination, the base current is determined by the hole current injected across the forward-biased emitter-base junction which, in turn, is proportional to the base doping. For sufficient gain, (i.e. ratio of collector to base current), the base doping has to be lower than the emitter doping in a normal bipolar transistor. As we shall see later, this limits the speed of the transistor.

Soon after the invention of the transistor, several companies in the early 1950's started production of silicon and germanium discrete devices consisting of the appropriately doped semiconductor pieces suitably packaged and wire bonded to three terminals.



These devices immediately replaced the bulky, power-hungry vacuum tubes in many low power applications. As designers grew ambitious, circuits with hundreds of transistors, which were impossible with vacuum tubes due to the size and power constraints, were being envisaged and fabricated. This led to the 'interconnect' problem: How does one reliably connect many hundreds of transistors and other passive elements to form a useful circuit? In 1958 while Jack Kilby was working for Texas Instruments on this problem, he came up with a simple but profound idea: fabricate the entire circuit on a single piece of semiconductor! It was well known that doped semiconductors could be used as resistors and  $p-n$  junctions behaved like simple parallel-plate capacitors due to the depletion region sandwiched between the

two bulk doped regions. Kilby realized that both active elements such as transistors and passive elements such as resistors and capacitors could be fabricated simultaneously on a piece of semiconductor and connected using metal films. As proof of this concept, he built an RC oscillator on a piece of germanium using chemical etching masked with black wax to fabricate the different elements

and gold wire bonding to interconnect them. When he connected a DC power supply to this small grey piece of solid, a sine wave appeared on the oscilloscope! That was the first working integrated circuit (IC) and it must have been quite an amazing sight in the summer of 1958. A similar concept was proposed independently by Robert Noyce of Fairchild Semiconductor who, with J A Hoerni, also advanced the idea of using silicon as the basic semiconductor, silicon-dioxide for masking the diffusion and etching steps and aluminium films for interconnects. The latter procedure, called the planar process, is more widely used in IC technology today rather than Kilby's original mesa etching process.

As IC technology advanced with shrinking transistor geometries and higher packing



densities, it soon became apparent that the basic transistor structure had some serious limitations. One of the major problems was the requirement that the base doping has to be lower than the emitter doping for high gain. This leads to a large resistance for the narrow base region, and, in combination with the emitter-base capacitance, ultimately limits the switching speed of the transistor.

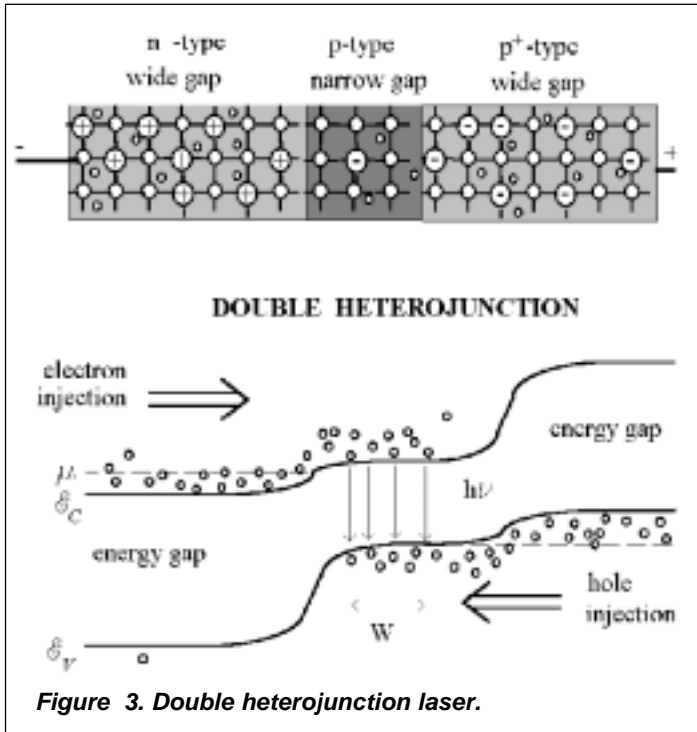
From the energy band diagram for a homojunction (*Figure 1*), it can be seen that the potential barrier for electrons and holes is identical at the junction ( $V_n = V_p$ ). Is there a way to increase the barrier for holes without affecting the electrons? In 1957, Herbert Kroemer proposed the use of a wide energy gap semiconductor as the emitter. In such a 'heterostructure', the energy band diagram shows that (*Figure 1*) the potential barrier for holes is increased ( $V_p > V_n$ ), thereby reducing the hole injection current substantially for a given doping. Although the idea of a wide bandgap emitter was suggested by Shockley in his original patent, it was Kroemer who presented the first comprehensive analysis of the 'heterojunction' bipolar transistor (HBT) including the effects of graded energy gaps. Kroemer showed that, in the HBT, the base doping can be increased beyond the emitter doping without degrading the gain. The high frequency performance of the transistor dramatically improves due to the reduced base resistance.

Another desirable effect of the lower base doping is an improvement in the output impedance through the suppression of the base-

width modulation or 'early' effect. Today HBTs dominate the high frequency market such as satellite communications, mobile telephony and other wireless applications.

The development of heterojunction devices was delayed since Kroemer's 1957 proposal due to difficulties in fabricating a junction between two different semiconductors. Kroemer however pursued his basic idea of 'using energy gap variations to control separately and independently the motion of electrons and holes' and applied it to the semiconductor laser. It was known, at that time, that heavily doped *p-n* junctions made out of certain semiconductors such as gallium arsenide emit light when forward biased. The basic mechanism involves the recombination of injected electrons and holes at the junction, with the bandgap energy converted into photons. In 1962, R N Hall showed lasing from a *p-n* junction structure. However the threshold currents were too large and the device operated only at low temperatures. From an analysis of the structure, it was clear that rapid diffusion of carriers away from the junction limited the carrier densities, and hence the optical gain. In 1963, the idea of a double heterojunction (DH) laser was independently suggested by Kroemer and Alferov. The heterojunctions confine the electrons and holes close to the junction and lasing action should be possible with lower threshold currents (*Figure 3*). It took another six years for a practical demonstration of this concept when, in 1969, Alferov and co-workers grew the first aluminium gal-





lithium arsenide (AlGaAs) and gallium arsenide (GaAs) heterojunction by a process called liquid phase epitaxy.

AlGaAs, a ternary alloy semiconductor, has a wider energy gap than its closely related cousin, GaAs. More importantly, the lattice constants of the two materials are closely matched, permitting the growth of high quality, defect free heterojunctions. A year later, in 1970, Alferov's group in Russia demonstrated the first DH laser operating continuously at room temperature. Subsequently, within a few months, M B Panish and his colleagues at Bell Laboratories, USA, also obtained room temperature lasing in a heterojunction structure. An important advantage of this structure is the larger refrac-

tive index of GaAs compared to the outer AlGaAs layers. This leads to a waveguide effect which confines the light to the junction and results in higher optical gain. The DH laser was an instant success and, with advances in growth techniques such as molecular beam epitaxy and chemical vapour deposition, the HBT and many other heterojunction devices have become a practical reality.

In today's Internet age, personal computers, multimedia devices, fibre optic networks, and wireless communications have become part of everyday life. In-

side all these gadgets, you will find tiny integrated circuits, semiconductor lasers and heterojunction transistors silently and reliably operating for many years continuously. For this, we are indebted to Kilby, Kroemer and Alferov and their pioneering achievements nearly forty years ago.

### Suggested Reading

- [1] D N Bose, *Transistors – From Point Contact to Single Electron*, *Resonance*, Vol.2, No.12, 1997.
- [2] Amit Roy, *Discovery of Transistor Effect that changed the Communication World*, *Resonance*, Vol.3, No.9, 1998.
- [3] Official website of the Nobel Foundation, <http://www.nobel.se/>.

V Venkataraman, Department of Physics, Indian Institute of Science, Bangalore 560012, India.

