Silicon-Molecular Beam Epitaxy

Volume I

Editors

Erich Kasper, Ph.D.
Head
Physical Technology Department
AEG Research Center
Ulm, West Germany

John C. Bean, Ph.D.
Head
Materials Science Department
Physics Research Division
AT&T Bell Laboratories
Murray Hill, New Jersey

CRC Press, Inc.
Boca Raton, Florida

Chapter 8

POSSIBLE DEVICE APPLICATIONS OF SILICON MOLECULAR BEAM EPITAXY

S. Luryi and S. M. Sze

TABLE OF CONTENTS

I. INTRODUCTION

Let us begin by clearly defining our task. The purpose of this work is to discuss the device potential of silicon molecular beam epitaxy (MBE). We will not review the existing MBE literature thoroughly (that is done by other contributors to this book) nor will we announce any startling new device structures (for those one has patent disclosures). As defined by the editors, this is supposed to be a "blue-sky" chapter, which deals with the following question: "what, if anything, is Si-MBE for?" In order not to treat this fundamental question too cavalierly, we shall begin by specifying our guiding principles. A useful solid-state device is one which can be used in electronic applications or can be used to study fundamental physical properties. We shall be mainly concerned with the first category of devices, which means that our assessment of the fruitfulness of the existing directions in Si-MBE research will be guided by the stringent criteria of the future marketplace as we see it today.

Since the invention of the transistor, many advances in semiconductor electronics and physics have been associated with a steady improvement in methods for preparing material structures with precisely controlled composition and dimensions. Since its conception in the early 1970s,[1] MBE has established itself as the most versatile epitaxy technique for growing single-crystal layers of semiconductors, giving the ultimate spatial resolution (on the scale of a nanometer) of the composition and the impurity incorporation in these layers. It has been clearly demonstrated that MBE is gainfully applicable to a number of conventional microwave, logic, and optoelectronic devices. In addition, the extraordinary dimensional control of semiconductor layers afforded by MBE has enabled the fabrication of unconventional device structures, such as modulation-doped layers, superlattices, quantum wells, etc., whose new and remarkable properties result from the confinement of electronic states to narrow layers. The study of such devices has led to discoveries of new and often unanticipated phenomena, the most dramatic of which is undoubtedly the fractional quantum Hall effect.[2]

Most of these advances have been made with the MBE of III-V compounds.[3] It is interesting to note that as this field has entered its later stages of maturity, it faces fierce competition from another technique, metal-organic chemical vapor disposition (MOCVD),[4] which is potentially as powerful and versatile as MBE, but has significant economic advantages. No similar competition as yet is facing Si-MBE: low-temperature MOCVD techniques are unlikely to work in the case of Si, presumably because of the strength of the C-Si chemical bond.

The power of MBE of III-V (and perhaps also II-VI) compound semiconductors for producing new device structures derives in large part from the existence of *lattice-matched* material systems of variable bandgap and refractive index. These systems, notably the GaAs/AlGaAs heterostructures, allow the construction by MBE of virtually arbitrary potential profiles for electrons and holes, including abrupt band discontinuities used to confine carriers in a two-dimensional (2D) state. At present, the most important new electronic device based on carrier confinement at band discontinuities is the so-called modulation-doped field-effect transistor (MODFET).[5,6] Waveguiding of light by the variable refractive index in compound semiconductor systems as well as direct interband transitions, available in these materials, are extensively used for heterostructure lasers,[7] and 2D confinement of carriers in quantum-well structures grown by MBE or MOCVD further improves the laser characteristics.[7,8] A number of other novel electronic and optoelectronic devices available through the "bandgap engineering" of III-V compound semiconductors have been reviewed in Reference 9.

In contrast to the above-mentioned achievements, Si-MBE has had very limited device application. Most of the homoepitaxial Si device proposals (which we shall discuss in Section II) have sought to utilize only the power of MBE to produce abruptly varying dopant profiles. One can think of a number of devices whose performance requires (or is improved by) the

resolution of doped layers on a nanometer scale. However, there is often no particular advantage in having such devices implemented in silicon — in fact, many of them may be expected to perform better in GaAs. As outlined in the beginning of Section I, a stringent test of the potential usefulness of Si-MBE devices is whether they are compatible with very large-scale integration (VLSI) silicon technology and can be expected to complement and enhance the latter. Merely having a novel device structure implemented in Si cannot be recognized as a legitimate goal of MBE.

The heteroepitaxial combination of Si with other semiconductors and hence the "bandgap engineering" are complicated by the lack of suitable lattice-matched partners. A rule of thumb in heteroepitaxy has been that the growth of a good interface requires a lattice mismatch of less than 0.5%. This requirement in the case of silicon is satisfied only by gallium phosphide and aluminum phosphide (GaP and AlP have a perfect lattice match to each other and an acceptable 0.37% mismatch to Si). Investigation of the Si/GaP system by MBE has been reported;[10] its application in the optical pickup "lightpen" is being implemented by Philips Laboratories. A difficulty in creating a GaP/Si heterostructure is that GaP tends to decompose during growth into gallium and phosphorus and then to dope the silicon. Another difficulty is associated with the antiphase disorder due to an atomic ambivalence in the nucleation of epitaxial growth (both Ga and P can bond to a particular site on a Si <100> surface).

These problems do not arise in the case of Ge, which is chemically compatible with Si, but their lattice mismatch is substantial (4%), with Ge having the larger lattice constant. The Ge/Si heterostructure is normally replete with misfit dislocations, which make it difficult to use the electronic properties of the interface. An important advance in this regard is afforded by the strained-layer epitaxy technique[11,12] which allows one to grow thin layers of Ge_xSi_{1-x} alloy on a silicon substrate virtually free of dislocations (the larger alloy lattice compresses to be accommodated by the Si lattice). The maximum thickness of the strained layer (beyond which one has incommensurate growth with the strain relaxed through generating dislocations) depends on the alloy composition x and on the growth conditions. In the low-temperature MBE growth of the strained layer, the formation of dislocations is impeded by a kinetic barrier, so that the actual dislocation-free stained-layer thickness can be an order of magnitude larger than that thermodynamically allowed under equilibrium conditions.[12] The strained layer can be capped by another Si layer which grows without distortion, and then the whole layer-sequence can be repeated, forming what is known as the strained-layer-superlattice (SLS). The SLS of variously designed composition profiles can be grown by MBE virtually free of dislocations, providing silicon technology with the new materials of variable bandgap E_G and refractive index n. Device potential of the Si/Ge SLS is discussed in Section III.

Speaking about heteroepitaxial materials, one usually has in mind the utilization of their unique heterojunction properties: abrupt potential barriers, light wave-guiding, etc. The above-mentioned studies of silicon-based SLS and Si/GaP heterostructures are mainly directed toward the same goal. The goal we shall also stress in this work is quite different in that one does not have to be concerned with utilizing the interfacial properties of a heterojunction formed by silicon and another material. In one view, the Si-MBE research should be strongly oriented toward the possibility of growing high-quality layers of the desired material on a silicon substrate — with intermediate layers if necessary. Leaving a fuller discussion of this topic to Section IV, we should like to emphasize here that in our view it represents the most important direction of research — with the greatest implications for microelectronics of the future. This view is gaining popularity these days, and recently we have observed a real explosion in the number of published reports of a successful growth of device-quality Ge and GaAs layers on silicon wafers.

Topics covered in this chapter are shown schematically in Figure 1. It represents some kind of a genealogical tree extending into the future. In the spirit of our task (the "blue-sky" chapter), we have tried to guess what sort of device offspring one can expect with today's Si-MBE technologies regarded as the parents. Needless to say, these parents are yet often toddlers themselves and their ability to bear fruit will critically depend on their shape when (and if) they reach full maturity. A good example of such an uncertainty is afforded by the technologies of monocrystalline silicon/insulator heterostructures discussed in Section III.B. We can overestimate neither their potential importance (e.g., for three-dimensional integration of memory circuits) and their capability to produce unique and useful device structures nor the hurdles one has to overcome to make these techniques viable. Inasmuch as the critical analysis of these hurdles is not within our scope, we shall have to exercise greater caution when discussing such structures.

In some cases we were not able to discern a viable device offspring — which may be attributed to our lack of imagination. For example, most of the active research on monolithic silicon-silicide-silicon single-crystal structures has had as its main objective the development of a metal-base transistor (MBT). Since, as discussed in Section II.C.3, we do not believe in the potential usefulness of an MBT, our assessment of this direction is rather pessimistic. Of course, the silicide-silicon combinations by MBE may find different and very important uses, and the MBT-like structures can offer an excellent way of studying the material properties.

Finally, it should be emphasized again that any discussion of the device potential of Si-MBE has to be guided by economic considerations. Since MBE is necessarily a slow-growth process, it is and probably will remain a rather expensive proposition. Nevertheless, the cost per square centimeter of a MBE-treated wafer can be expected to go down as the technology matures. An important advance in this regard is the development[13] of an industrial Si-MBE system which offers a substantially enhanced economy due to a simultaneous deposition on several large-area wafers and streamlined throughput. The cost of MBE deposition is also a strong function of the complexity of a structure involved and, ultimately, this should be taken into account (which we shall not attempt to do in this paper) in assessing the potential viability of a device. Moreover, some of the most important applications, such as the heteroepitaxy of III-V compounds on silicon substrates, may not be limited at all by the economics of slow growth. Commercially important structures of this kind will probably be grown in MOCVD reactors vacuum-interlocked with an ultrahigh vacuum (UHV) Si-MBE system. The economic considerations related to this development, as well as our general vision of the future microelectronics which follows from such arguments, will be discussed in Section IV. Our conclusions will be summarized in Section V.

II. HOMOEPITAXIAL DEVICES

This section deals with the "traditional" domain of MBE — devices whose operation requires super-fine resolution of doping and composition profiles and relies on the electrical and optical properties of the resulting homo- or heterojunctions. We begin with silicon-only epitaxial structures and consider possible applications of the abrupt doping profiles available by MBE. Several homoepitaxial structures involving abrupt doping variations are schematically shown in Figure 2. Depending on the carrier transport mechanism, the possible device structures can be classified as bipolar or unipolar. Theory and applications of the bipolar *pn*, *pin*, *pnp*, *npn*, *pnpn*, ... junction devices are known very well[14] and will not be given much space in this work, in particular since the idealized models often used to describe such devices correspond precisely to the degree of perfection which has become available only with MBE. It suffices to state here that we regard improvements of the bipolar technology [especially the junction transistor, but also the impact avalanche transit time (IMPATT)

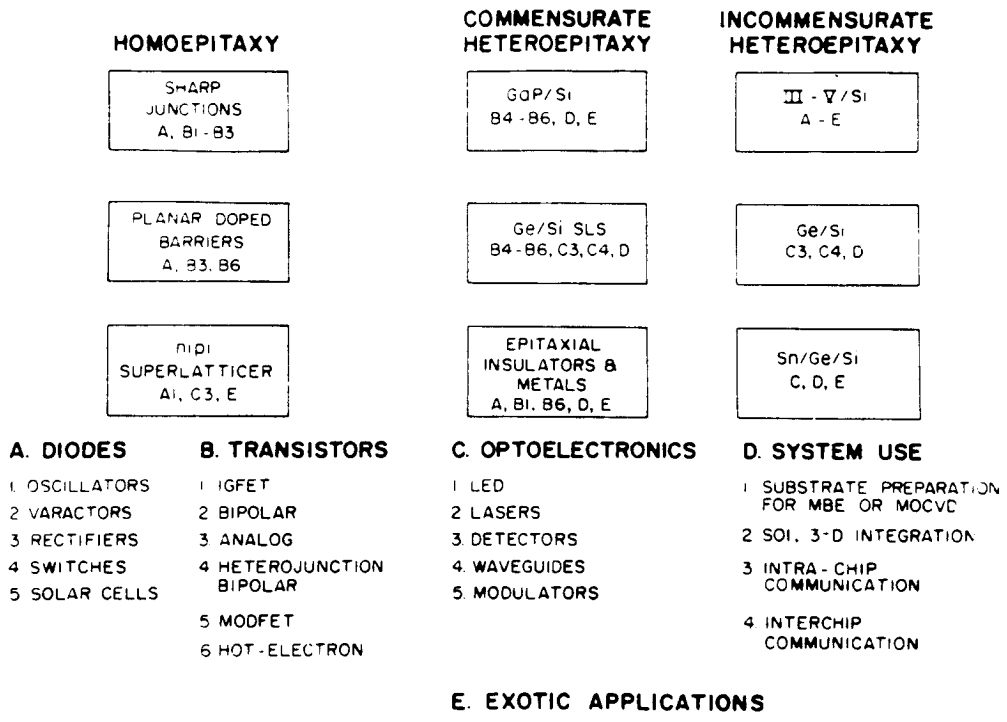

FIGURE 1. Classification of possible device applications of Si-MBE.

diode] as an important goal of Si-MBE research. The same attitude applies to possible improvements of the surface unipolar devices, such as the metal-oxide-semiconductor field-effect transistor (MOSFET), with the precisely engineered doping profiles, controlled by MBE. These topics will be briefly discussed in Section II.A. On the other hand, the theory of bulk unipolar structures (Section II.B), such as the planar-doped barriers, may not be as widely known with its bits and pieces scattered in the literature. These structures by themselves have a limited device potential, but they can serve as elements in more sophisticated three-thermal transistor structures (Section II.C).

A. Improvement of Conventional Si Devices by MBE

By "conventional" devices we mean those which have already managed to find commercial applications. All such devices have been extensively studied and most of them are well understood by now.[14] With this understanding, it is quite easy to see where MBE can make an improvement. Nevertheless, there has been relatively little Si-MBE activity in that direction and for a good reason. The reason, of course, lies in economic considerations. Indeed, the main advantage of silicon technology lies in the economics of VLSI. Once a device leaves that proven terrain, it no longer competes against similar silicon devices fabricated without the help of MBE, but rather against various devices in all possible materials that may be capable of performing the same function. A particular Si device can win in this competition only in two situations: (1) the device is compatible with a VLSI technology and can enhance the latter's performance and (2) because of some intrinsic material advantage a discrete silicon device can be expected to outperform the competitors, even if the latter are manufactured with the same degree of sophistication, including MBE.

1. Transistors

There may be room here for realizing Situation 1. The performance of a bipolar junction transistor can undoubtedly benefit from MBE. One advantage is the possibility of employing

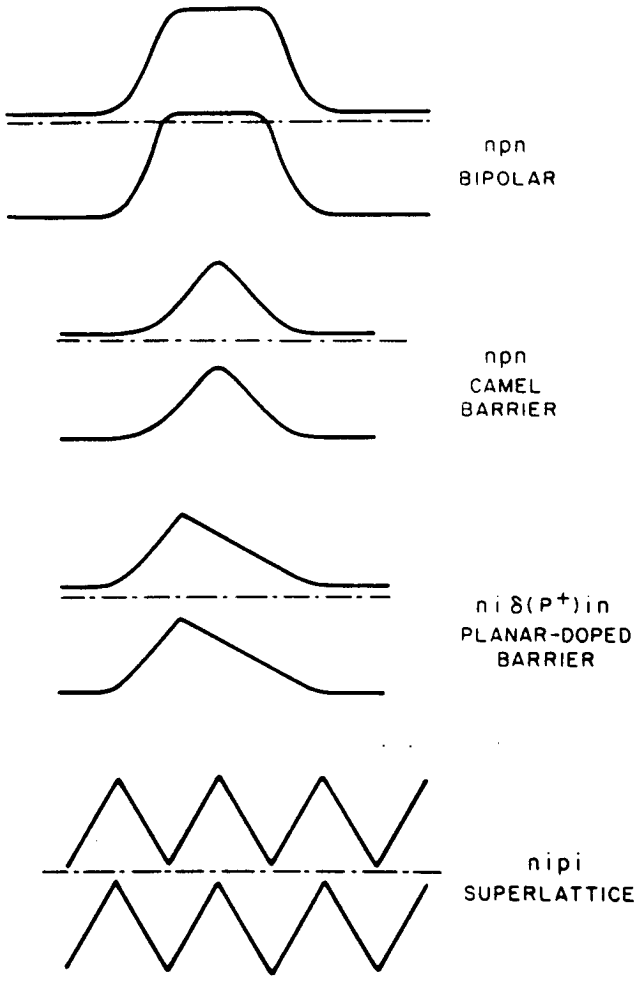

FIGURE 2. Homoepitaxial structures with abrupt doping variations.

uncompensated *pn* junctions.[15] By precisely controlling the doping profile, one can maximize the current gain and increase the speed. Recently, Kasper and Wörner reported[16] a bipolar frequency-divider circuit by Si-MBE — with clock frequencies of up to 2.8 GHz. Notably, this work relied entirely (apart from the use of MBE) on a commercial processing line -- thus demonstrating the compatibility of Si-MBE with the standard bipolar fabrication technology.

The MBE may also provide an ideal way for implementing the Moll-Ross drift transistor (Reference 14, page 138). If the base doping N decreases exponentially from the emitter junction and toward the collector junction, then the built-in field, $E_B = (kT/q)\, d(\ln N)/dx$, enhances the minority-carrier drift through the base, leading to shorter delays. As a discrete device, a Si drift transistor by MBE should be judged in comparison against heterojunction bipolar transistors (Reference 14, page 182) which can give both the higher gain by totally suppressing minority injection into the emitter and, potentially, the higher speed by incorporating Kroemer's idea of assisting the minority-carrier transport with a built-in quasielectric field due to a graded chemical composition of the base[17] (Figure 3). As far as large-scale circuit applications are concerned, the use of bipolar transistors is, generally, impeded by the thermal instability characteristic of a current-driven device, which does not favor a

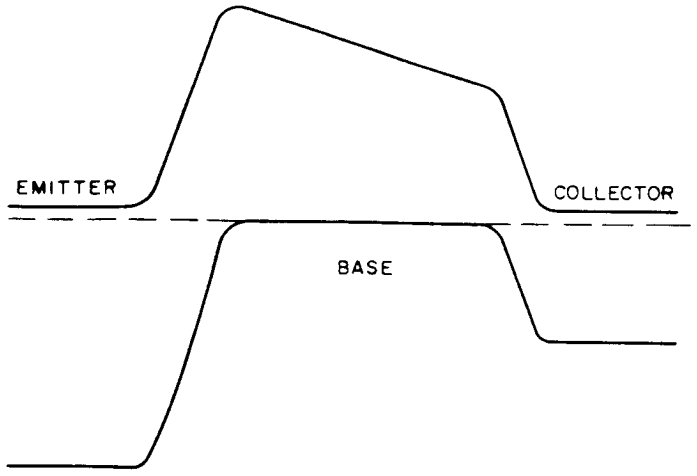

FIGURE 3. Schematic band diagram at zero bias of a heterojunction bipolar transistor with compositionally graded base and wide-gap emitter.[17]

parallel circuit connection of such devices, as compared to field effect transistors (FETs). This instability consists of the following. If in the course of a circuit operation one of the parallel devices develops a higher current, then the local temperature in its vicinity is enhanced due to the Joule heating, and consequently the current in this device increases still further. This is probably the largest single drawback of the bipolar concept, and MBE can hardly remedy this situation.

However, the future large digital systems may well combine the low-power CMOS circuits (allowing a large degree of parallel connections) with the higher current-drive capabilities (and possibly higher speeds) of bipolar transistors. Bipolar elements are particularly desirable for irregular of nonrecursive VLSI circuits, with variable input and output loading conditions, and for driving high fan-in loads, both internal and external to the chip.[18] Si-MBE may have a role in this development. Of course, the issue of compatibility is very acute here; for instance, sharp doping profiles provided by MBE will not stand against some of the standard VLSI processing steps, such as the high-temperature oxidation. Nevertheless, we believe this issue can and will be resolved.

Let us now discuss a different question: what role can MBE play in the ongoing perfection of the silicon MOSFET itself? Several possibilities should be examined.

Creation of a precisely controlled doping profile — This may be used for adjusting the threshold voltage as well as for reducing the unwelcome short-channel effects, such as the punch-through between source and drain. In present-day technology, these goals are achieved by ion implantation (Reference 14, pages 456 to 469). The power of MBE in this regard has not been demonstrated, in part because the thermal redistribution of impurities in subsequent processing would wash-out and destroy an ideal doping profile achieved during the epitaxial growth. It is clear that the nonuniform-doping and buried-channel device applications will depend on the development of advanced processing techniques, such as the low-temperature oxidation. Even then, it remains to be seen whether the MBE advantage will be sufficiently important to justify its use.

Improvement of the material quality — Ultra-high vacuum cleaning of the substrate and *in situ* oxidation can produce automatically smooth interfaces virtually free of hydrocarbon contamination. This may lead to a substantially improved carrier mobility in the inversion layer. It should be noted that this line of development does not necessarily involve slow growth processes and thus it may be subject to different economic considerations than the Si-MBE in general.

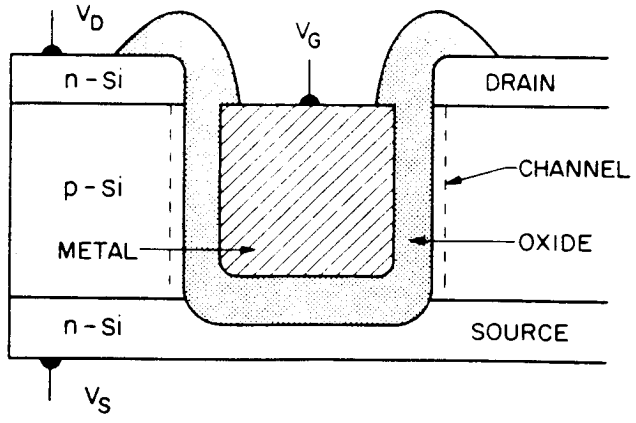

FIGURE 4. A sidewall MOSFET by MBE. Instead of the vertical channel, one can also have a V-groove channel on a slanted surface obtained by anisotropic etching. However, in that case, the effective gate length would have to be substantially larger than the source-to-drain distance, resulting in a large parasitic capacitance.

Radically new FET structures — There is, of course, no way of predicting the form this development may take. One important direction may be associated with growth on patterned surfaces.[19] An interesting though untested possibility is that of a "sidewall" MOSFET, illustrated in Figure 4. In such a hypothetical device, the source-drain separation would be determined during MBE growth — on a much finer scale than that achievable with any lithographic process. This device is conceptually similar to the VMOS or UMOST structures considered for high-power and VLSI applications (Reference 14, pages 494 and 495) prior to the advent of MBE.

2. Diodes

An important group of conventional devices which can certainly benefit from MBE are the transit-time diodes, such as the IMPATT and the BARITT (Reference 14, Chapter 10). Of the two, the IMPATT (Read) diode is by far more important, being today one of the most valuable sources of microwave power (the BARITT diode also has a limited use in low-power applications requiring reduced noise). Performance of transit-time devices, especially at millimeter and submillimeter wave frequencies, depends on the precise control of layer thicknesses and of the doping (Figure 5), which naturally calls for the use of MBE. Indeed, one of the earliest applications of MBE was the fabrication of a GaAs IMPATT diode.[20] There have also been reports of Si-MBE IMPATT diodes[21] operating in the W band (75 to 110 GHz).

For super-high-frequency IMPATT applications, silicon may be preferred over GaAs.[14] One reason is its shorter intrinsic avalanche response time, τ_i (the time by which the ionization rate lags the electric field), which limits the use of GaAs to below 100 GHz. For Si, τ_i is less than 10^{-13} sec. The other reason is the nearly twice larger high-field drift velocity in Si, from which one can expect a higher power output at a given frequency in the millimeter range. On the other hand, GaAs IMPATTs are preferable below 100 GHz because of their larger efficiency and lower noise. Higher efficiency of GaAs IMPATTs results from the fact that the drift velocity in GaAs saturates at lower electric fields, which allows a larger margin for voltage swings. As far as noise is concerned, a disadvantage of silicon lies in the large disparity between the impact ionization rates for electron and holes, $\alpha_n >> \alpha_p$, which requires a larger number of ionization events to achieve the same amount of multiplication.*

* It is interesting to note that the same physical property which makes Si-IMPATT noisier is beneficial when one considers the noise performance of avalanche photodiodes (Section III.A.1).

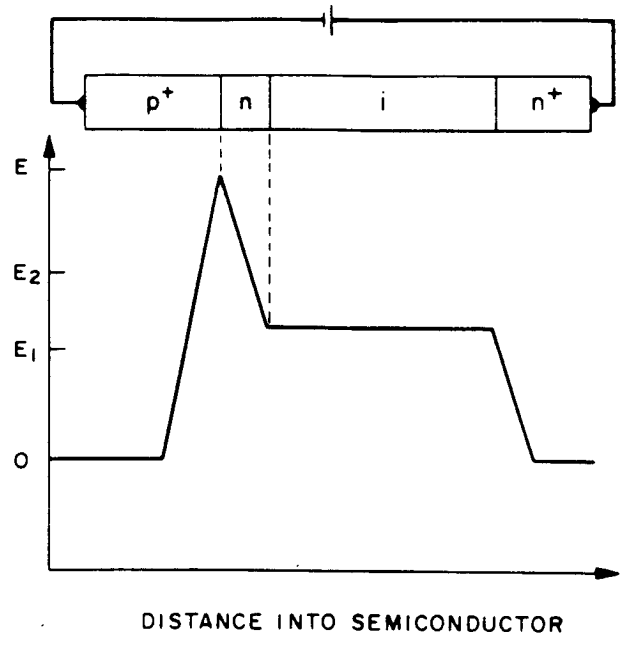

FIGURE 5. Schematic diagram of the electric field profile in a Read diode. Threshold field for impact ionization breakdown is denoted by E_2 and the field beyond which the electron drift velocity can be considered saturated by E_1.

It appears to us that fabrication of a discrete BARITT device would hardly ever justify the use of MBE, even though it alone is capable of producing the optimum (planar-doped) device structure.[22] Moreover, we see no reason why Si should be preferred in such an application. However, a situation may be envisaged in which a low-noise local BARITT oscillator is called for in a large custom VLSI circuit in the fabrication of which Si-MBE is employed for a different purpose. A similar situation may arise for other types of conventional diode structures improved by MBE, such as varactors, switching diodes, etc. (Reference 23, pages 87 to 101).

B. Theory of Bulk Unipolar Structures with Abrupt Doping Profiles

These structures often involve junctions between a heavily doped and an undoped silicon layer. In such junctions the effects of carrier diffusion into the undoped layer are important and one cannot use the familiar depletion approximation which works so well for understanding properties of bipolar junctions.

1. Symmetric n-i-n Structure

To illustrate these effects, let us consider the case of a symmetric *nin* junction (Figure 6). In this case, the potential barrier $\phi(x)$ is due to the mobile charge diffusing into the intrinsic *i* layer from the *n* layers doped to the level N_D. The shape of this barrier, as well as the field and the charge distributions in the junction, can be rather neatly expressed in a closed analytic form by solving the Poisson and the drift-diffusion equations. Introduce the dimensionless variables: coordinate $\xi \equiv x/x_0$; electric field, $\epsilon \equiv E/E_0$; concentration, $\nu = n/N_0$, and current, $j = J/J_0$, where

$$N_0 \equiv N_D \tag{1a}$$

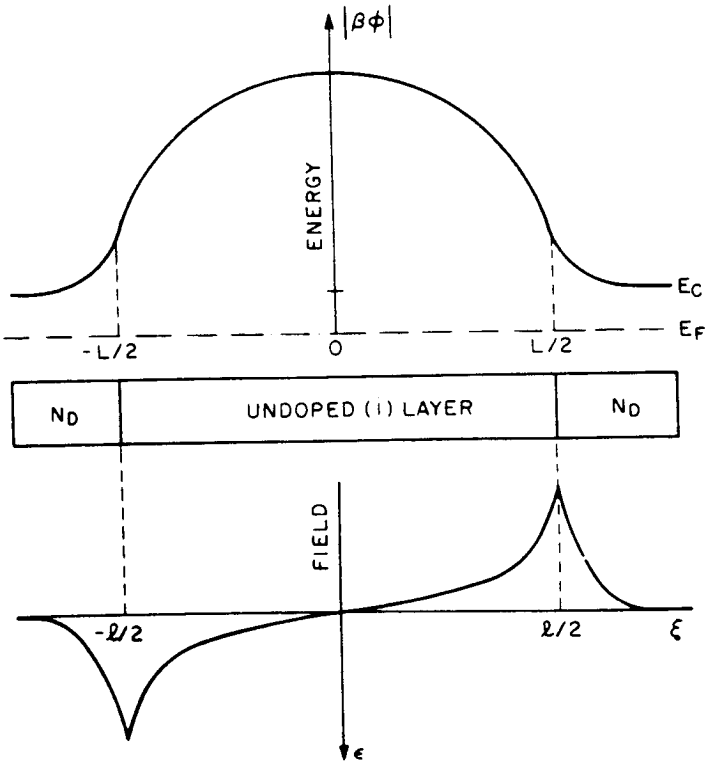

FIGURE 6. Conduction-band diagram and electric field profile in a thin symmetric *nin* junction in equilibrium. Solution below is exact provided the intrinsic carrier concentration in the *i* layer can be neglected, i.e., if $e\phi << E_G/2$.

$$x_0 \equiv L_D = (\kappa kT/q^2N_D)^{1/2} \tag{1b}$$

$$E_0 \equiv kT/qL_D = (kTN_D/\kappa)^{1/2} \tag{1c}$$

$$J_0 \equiv q\,N_D\mu E_0 = q\,N_D D/L_D = \kappa\beta\mu E_0^3 \tag{1d}$$

L_D being the Debye length in the doped layers, κ being the dielectric permittivity, and μ and D, respectively, being the mobility and the diffusion constant in the *i* layer, related by the Einstein formula:

$$qD = \mu kT \tag{2}$$

We set the origin $x = 0$ in the middle of the *i* layer of thickness L (in dimensionless units $\ell \equiv L/L_D$). The Poisson equation in these units is of the form

$$\frac{d\epsilon}{d\xi} = -\nu, \quad |\xi| \leq \ell/2 \tag{3a}$$

$$\frac{\partial^2(\beta\phi)}{d\xi^2} = e^{\beta\phi} - 1, \quad |\xi| \geq \ell/2 \tag{3b}$$

where $\beta \equiv q/kT$ and $\epsilon = -\partial(\beta\phi)/\partial\xi$. The drift-diffusion equation in the i layer with the help of Equations 2 and 3a can be put in the form:

$$\frac{\partial}{\partial\xi}\left(\frac{1}{2}\epsilon^2 + \frac{\partial\epsilon}{\partial\xi}\right) = j = 0 \tag{4}$$

The solution of Equation 4, satisfying $\epsilon(0) = 0$ and the physical condition $\nu > 0$, is given by

$$\epsilon = -2\gamma \tan(\gamma\xi), \quad |\gamma\xi| < \pi/2 \tag{5}$$

with a real γ. Integrating Equation 5, we find the potential everywhere in the i region:

$$\beta\phi = \beta\phi(0) - \ln\cos^2(\gamma\xi), \quad |\xi| \leq \ell/2 \tag{6}$$

where

$$e^{\beta\phi(0)} = \nu(0) = 2\gamma^2 \tag{7}$$

On the other hand, outside the i layer the potential is found by integrating Equation 3b, namely,

$$e^{\beta\phi} - \beta\phi - 1 = \epsilon^2/2, \quad |\xi| \geq \ell/2 \tag{8}$$

Matching Equations 6 and 8 at $|\xi| = \ell/2$ and using Equation 7, we find

$$\cos^2(\gamma\ell/2) = 2\gamma^2 \exp(1 - 2\gamma^2) \tag{9}$$

This equation determines $\gamma(\ell)$ and thus completely solves the problem. The proper choice of the solution is provided by the condition in Equation 5, namely, $|\gamma\ell| < \pi$. It is instructive to plot $\nu(0)$ vs. ℓ, Figure 7. In the limit $\ell >> 1$ (i.e., $L >> L_D$ and $n(0) << N_D$) one has $|\gamma\ell| \to \pi$ and hence $\nu(0) \approx 2\pi^2/\ell^2$ or, equivalently,

$$n(0) \approx \frac{2\pi^2\kappa kT}{q^2L^2}, \quad n(0) \ll N_D \tag{10}$$

Note that in this limit $n(0)$ is independent of N_D. Of course, the above solution is valid only if $n_i << n(0)$, where n_i (T) is the intrinsic carrier concentration in silicon (approximately $10^{10}\,cm^{-3}$ at 300 K). Similar results have been obtained[24] by a somewhat different mathematics.

When an external bias is applied to an *nin* structure, the above equilibrium considerations do not apply. At very low currents one can, qualitatively, view the current as a case of thermionic emmision over the barrier in the middle of the i region. Lowering of this barrier will be at first proportional to the applied voltage V, and hence the current $J \propto \exp(\beta V/2)$ (where the factor of 2 describes the ideality factor of a symmetric diode). However, as the applied bias increases, the top of the barrier ("virtual cathode") moves toward the negative electrode and the ideality factor rapidly degrades (compare the discussion of planar-doped barriers in the next section). Eventually, the virtual cathode settles near the edge of the i region, and the current becomes space-charge limited — the initial exponential characteristic being replaced by a power law, analogous to Child's law of vacuum electronics. This transition to space-charge limited emission was recently considered by Grinzberg and Luryi[25]

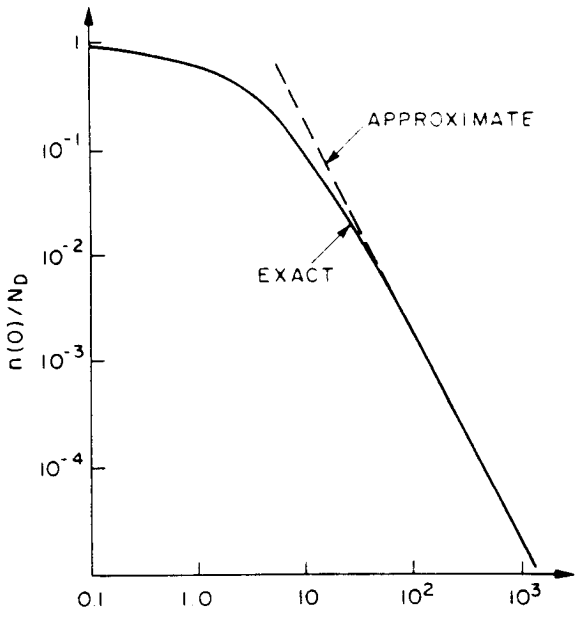

FIGURE 7. Dependence of the equilibrium carrier concentration in the middle of the undoped *i* layer, due to electron diffusion from the adjacent *n*-doped layers, on the *i* layer thickness L and *n* layer doping N_D. Solid line shows the exact solutions given by Equations 9 and 7, where $\nu \equiv n/N_D$ and $\ell \equiv L/L_D$, and dashed line corresponds to the approximation of Equation 10.

in a model which assumes a constant field-independent mobility. For higher fields, when the electron drift velocity saturates, that model cannot be expected to be a good approximation. Rigorous analytic description of the current-voltage characteristics in *nin* diodes is of substantial academic interest, but still lacking in the literature.

2. Planar-Doped Barriers

The planar-doped-barrier (PDB) rectifying structure (Figure 8) was first demonstrated by Malik et al.[26] in GaAs MBE-grown samples. It represents an extension (the limiting case) of the earlier work by Shannon[27] on camel-barrier diodes. The first Si-PDB diode was also demonstrated by MBE.[28] We begin by reviewing the theory of rectification and charge injection in this important structure.[29,30]

A PDB [*n*-*i*-δ(p^+)-*i*-*n*] structure represents a nearly intrinsic (*i*) layer of thickness L sandwiched between two *n*-type layers of low resistivity. In the process of MBE growth a p^+-doped layer of ("infinitesimal") thickness $\delta << L$ and doping N_A is built into the *i* region. Acceptors in the p^+ layer are completely ionized forming a negative charge sheet of surface density $\Sigma = qN_A\delta$ which gives rise to a triangular potential barrier with shoulders L_1 and L_2 and the height Φ approximately given by

$$\Phi \approx -\frac{\Sigma L_1 L_2}{\kappa L} \tag{11}$$

The exact shape of the potential barrier, including the effects of the mobile charge diffusion into the *i* layers, can be found in the manner similar to that described above for the *nin* case. The appropriate solution of the nonlinear drift-diffusion Equation 4 is now of the form

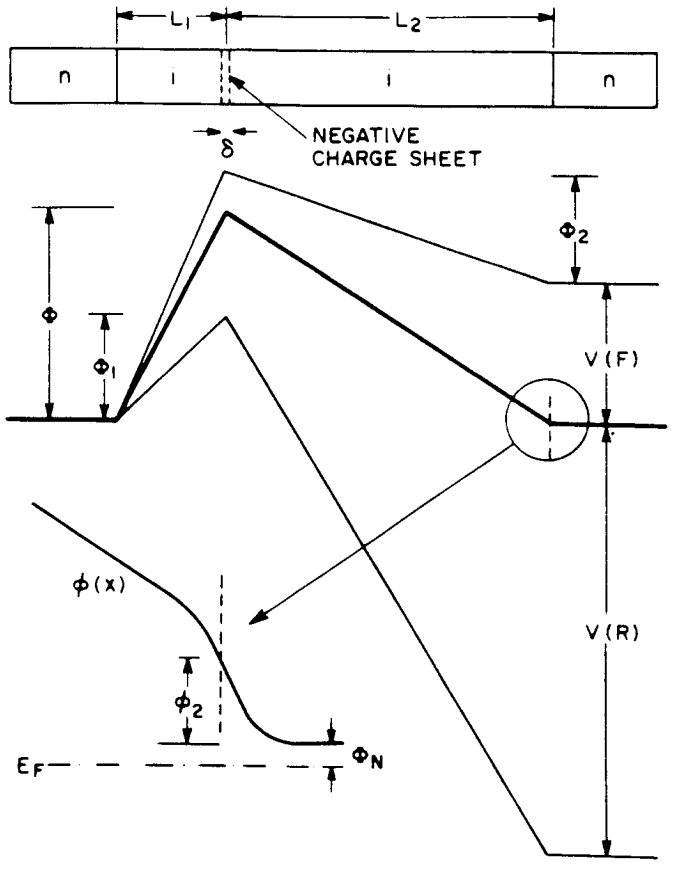

FIGURE 8. Schematic illustration of the planar-doped triangular barrier. In the example drawn, the barrier shoulders are related as 3:1. Evidently, one has to apply approximately three times larger voltage in the reverse (R) than in the forward (F) direction to obtain the same barrier height ($\Phi_1 = \Phi_2$) for the thermionic emission, hence the rectification. Insert in the lower left corner shows the barrier shape including the effects of carrier diffusion into the *i* layer.

$$\epsilon = \begin{cases} 2\gamma_1 \coth[\gamma_1(\xi + \xi_1)], & 0 \leq \xi \leq \ell_1 \\ -2\gamma_2 \coth[\gamma_2(\ell - \xi + \xi_2)], & \ell_1 \leq \xi \leq \ell = \ell_1 + \ell_2 \end{cases} \tag{12}$$

where $\ell_i = L_i/L_D$ and ξ_i, γ_i, $i = 1, 2$ are constants to be determined. At $\xi = 0$ and $\xi = \ell$, one has from Equations 8 and 12:

$$e^{\beta\Phi_i} - \beta\phi_i - 1 = 2\gamma_i^2 \coth^2(\gamma_i\xi_i), \quad i = 1, 2 \tag{13a}$$

Differentiating Equation 12 and using Equation 3a one also has

$$\nu_i = e^{\beta\Phi_i} = 2\gamma_i^2/\sinh^2(\gamma_i\xi_i), \quad i = 1, 2 \tag{13b}$$

where ν_1 and ν_2 are the dimensionless electron concentrations at the boundaries of the *i* layer. From Equations 13a and 13b, we find that

$$\beta\phi_i = -1 - 2\gamma_i^2, \quad i = 1, 2 \tag{14}$$

$$\sinh^2(\gamma_i\xi_i) = 2\gamma_i^2 \exp(1 + 2\gamma_i^2), \quad i = 1, 2 \tag{15}$$

Integrating Equation 12, we obtain the potential everywhere in the i layer relative to $\phi = 0$ in the bulk of the doped layers:

$$\beta\phi(\xi) = \beta\phi_1 - \ln\left\{\frac{\sinh[\gamma_1(\xi + \xi_1)]}{\sinh(\gamma_1\xi_1)}\right\}^2, \quad 0 \leq \xi \leq \ell_1 \tag{16a}$$

$$\beta\phi(\xi) = \beta\phi_2 - \ln\left\{\frac{\sinh[\gamma_2(\ell - \xi + \xi_2)]}{\sinh(\gamma_2\xi_2)}\right\}^2, \quad \ell_1 \leq \xi \leq \ell \tag{16b}$$

From the continuity of ϕ, Equations 16a and 16b at $\xi = \ell_1$, and using Equations 14 and 15, we have

$$e^{\beta\Phi} = \frac{2\gamma_1^2}{\sinh^2[\gamma_1(\xi_1 + \ell_1)]} = \frac{2\gamma_2^2}{\sinh^2[\gamma_2(\xi_2 + \ell_2)]} \tag{17}$$

On the other hand, the discontinuity in the electric field at $x = L_1$ is given by Gauss's law, $E_1 - E_2 = \Sigma/\kappa$, where $E_1 = E(L_1 - \delta/2)$ and $E_2 = E(L_1 + \delta/2)$, whence

$$2\gamma_1 \coth[\gamma_1(\ell_1 + \xi_1)] + 2\gamma_2 \coth[\gamma_2(\ell_2 + \xi_2)] = \frac{\Sigma}{(\kappa kTN_D)^{1/2}} \tag{18}$$

Equations 15, 17, and 18 constitute a system of four equations in four unknowns ξ_i, γ_i, which can be solved for any given PDB parameters.

The barrier height Φ is then given exactly by Equation 17. The general expression $\Phi = \Phi(\Sigma, L_1, L_2, N_D)$ is too cumbersome to be presented here. However, this expression simplifies considerably if one notes that in all practical PDB in equilibrium one has $|\epsilon_i\ell_i| >> 1$, or, equivalently, $E_iL_i >> kT/q$. Under this condition $2\gamma_i = |\epsilon_i|$, and Equation 17 reduces to

$$\begin{aligned}\beta\Phi &= \beta\Phi_i \\ &= -|\epsilon_i|\ell_i + \ln\{2\epsilon_i^2([1 + (\epsilon_i^2/2)\exp(1 + \epsilon_i^2/2)]^{1/2} \\ &\quad - [(\epsilon_i^2/2)\exp(1 + \epsilon_i^2/2)]^{1/2})^2\}\end{aligned} \tag{19}$$

In the stronger limit $|\epsilon_i| >> 1$, or $|E_i| >> 63$ kV/cm (for $N_D = 10^{18}$ cm^{-3} in Si at room temperature), Equation 19 simplifies further, namely

$$\beta\Phi = -|\epsilon_i|\ell_i - \epsilon_i^2/2 - 1, \quad i = 1, 2 \tag{20}$$

Solving Equation 20 together with the Gauss law written in the form

$$|\epsilon_1| + |\epsilon_2| = \sigma \approx \Sigma/(\kappa kTN_D)^{1/2} \tag{18a}$$

where σ is the dimensionless charge-sheet density ($\sigma = 2.4$ at room temperature for $\Sigma/q = 10^{12}$cm^{-2} and $N_D = 10^{18}$cm^{-3}), we find

$$\beta\Phi = -\frac{\sigma\ell_1\ell_2}{\ell} - \frac{\sigma(\ell_1 - \ell_2)^2}{2\ell(\ell + \sigma)} - \frac{\sigma^2}{8}\left[1 + \left(\frac{\ell_1 - \ell_2}{\ell_1 + \sigma}\right)^2\right] - \frac{2}{\ell} \tag{21a}$$

Thus for a symmetric PDB ($\ell_1 = \ell_2$), we have

$$|\Phi| = \frac{\Sigma L_1 L_2}{\kappa L} + \frac{\Sigma^2}{8\kappa q N_D} + \frac{2kTL_D}{qL} \tag{21b}$$

We see that corrections to Equation 11 become appreciable for large Σ and/or small N_D, i.e., precisely in the limit when Equations 21a and 21b are strictly valid. It should be noted that even though the strong inequality $|\epsilon_i| >> 1$ is rarely fulfilled in practice, Equations 21a and 21b provide a good approximation already when $|\epsilon_i| \geq 2$, i.e., $E_i^2 \geq 4kTN_D/\kappa$.

Next, we consider a PDB under applied bias. It is possible to solve Equation 4 for $J \neq 0$ on each barrier shoulder (the solution is simple if one notes that j is usually a small parameter, since J_0 defined by Equation 1d is typically greater than 10^7 A/cm^2) and then obtain the current-voltage (I-V) characteristic by matching the solutions for $n(x)$ at $x = L_1$. However, such a solution would be almost meaningless, as it neglects the effects of carrier acceleration and heating (e.g., the velocity saturation) on the downhill slope of the barrier.

A better approach is to use the thermionic emission theory (Reference 14, page 255). It amounts to assuming that the quasi-Fermi level is constant on the uphill slope of the barrier all the way to $x = L_1$. In the low-current limit (discussed below), the I-V characteristic of a PDB can be calculated by the following procedure: Assume ϵ_1 and ϵ_2 subject to Gauss's law (Equation 18a). Calculate Φ_1 and Φ_2 in terms of the ratio ϵ_1/ϵ_2. Set $\Phi_1 - \Phi_2 = V$, and

$$J = qN_D v_T e^{\beta\Phi_i} = A^* T^2 e^{\beta(\Phi_i + \Phi_N)} \tag{22a}$$

where Φ_i is the injecting (uphill) barrier, $\beta\Phi_N = \ln(N_D/N_C) < 0$ is the separation between the conduction band and the Fermi level in the (nondegenerately doped) emitter, A* is the effective Richardson constant, and

$$v_T = (kT/2\pi m^*)^{1/2} \tag{23}$$

is the mean thermal velocity of carriers in a given direction. Repeating the calculation (which can be done with a pocket calculator) for different ratios ϵ_1/ϵ_2, one determines the entire low-current portion of the I-V characteristic.

This was first done in Reference 29 and a good agreement with the experimental results[26] for two different GaAs PDBs was found. Figure 9 (solid line) shows the calculated I-V characteristic for an exemplary silicon PDB structure. In the limit $E_i^2 \geq 4kTN_D/\kappa$ one can use Equation 20 and calculate I(V) in the closed form:

$$J = A^* T^2 e^{\beta(\Phi_0 - \kappa V^2/2qL^2N_D)}\left[e^{\beta V(\tilde{L}_1/L)} - e^{-\beta V(\tilde{L}_2/L)}\right] \tag{22b}$$

where Φ_0 is the sum of Φ given by Equation 21a and Φ_N, and $\tilde{L}_i$ are the effective barrier shoulder lengths,

$$\tilde{L}_i = L_i + \frac{\Sigma}{qN_D}\left(1 - \frac{2L_i}{L}\right) \tag{24}$$

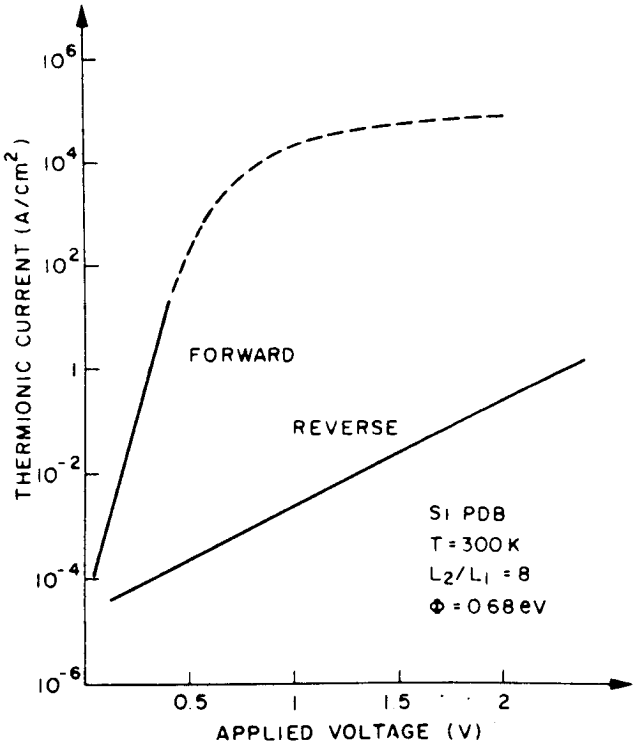

FIGURE 9. Calculated I-V characteristics for a Si PDB diode with the following parameters: $L_1 = 250$ Å, $L_2 = 2000$ Å, $N_D = 10^{16}$ cm^{-3}, $\Sigma/e = 2 \times 10^{12}$ cm^{-2}.

We see that at low biases the PDB diode characteristics are approximately exponential with the ideality factors $n_i \equiv L/\tilde{L}_i > 1$; note from Equation 24 that $n_1^{-1} + n_2^{-1} = 1$. A strongly asymmetric diode ($L_1 >> L_2$) exhibits rectification, with Φ_1 being the injecting and Φ_2 being the blocking barrier. Most of the usefulness of planar-doped barriers derives from this property.

At high applied biases, the exponential I-V characteristic of a PDB begins to saturate and eventually it is replaced by a linear law (see the dashed line in Figure 9). This happens[29,30] because of (1) slowing down of the effective diffusion velocity on the uphill slope, and (2) screening of the applied field by the mobile charge dynamically stored (i.e., stored while in transit) on the downhill slope. For potential applications of the PDB concept, it is most important to understand at what current levels this saturation occurs. Consider limitation 1 first.

As the applied bias increases, the injecting potential (Φ_1, for concreteness) flattens out. In the limit $1/\ell_1 << \epsilon_1 << 1$, Equation 19 is still valid and gives $\beta\Phi_1 \approx -\epsilon_1\ell_1 + \ln(2\epsilon_1^2)$. If we formally substitute this expression in Equation 22, we obtain

$$J = 2\kappa\beta v_T E_1^2 e^{-\beta E_1 L_1} \tag{25}$$

Equation 25 contains a preexponential factor which decreases with bias. It formally predicts $J \to 0$ in the limit of a flat uphill slope $E_i = 0$, while it is physically clear that without a barrier we must have a space-charge limited current flow. The curve J(E) goes through a maximum at $\beta E_1 L_1 = 2$. The reason for this unphysical behavior is twofold. On the one

hand, Equation 19 loses its validity when $\beta E_1 L_1 \leq 2$, which may be interpreted as a restriction on the effective diffusion velocity μE_1 on the uphill slope which must exceed a minimum value $2D/L_1$,

$$\mu E_1 \gg D/L_1 \tag{26a}$$

(when Equation 26a is violated, the potential can still be determined exactly by Equations 15 to 18). On the other hand, the thermionic theory itself (Equation 22) applies only if the electron mean free path exceeds the distance in which the barrier falls by kT from its maximum value.[31] The latter condition can be cast in the form

$$\mu E_1 \gg v_T \tag{26b}$$

Restrictions (Equations 26a and 26b) are roughly equivalent. Using Equation 25, we can thus write limitation 1 in the form of a restriction on the current density:

$$J \ll \frac{\kappa kT}{qL_1}\frac{v_T}{L_1} \equiv J_{C1} \tag{27}$$

Next, we consider limitation 2. Electrons emitted over the peak of the barrier drift downhill with a saturation velocity v_S. The density ρ of the injected charge is proportional to the current density, $\rho = J/v_S$. This charge screens the applied field by contributing a space-charge potential $\Delta\phi = JL_2^2/2\kappa v_S$ which must be subtracted from V when one evaluates Φ_1. If without this effect the characteristic would be similar to Equation 22, i.e., approximately $J \propto \exp(qV/n_1kT)$, where n_1 is an ideality factor of the injecting barrier, then due to the screening it becomes $J \propto \exp[(q/n_1kT)(V - JL_2^2/2\kappa v_S)]$. The screening effect can thus be described by an effective ideality factor $\bar{n}_1(J)$, namely,[30]

$$\bar{n}_1(J) = n_1\left(1 + \frac{qJ}{n_1kT}\frac{L_2^2}{2\kappa v_S}\right) \tag{28}$$

whence we can express limitation 2 also as a restriction on the current density:

$$J \ll \frac{\kappa kT}{qL_2}\frac{v_S}{L_2} \equiv J_{C2} \tag{29}$$

When the conditions in Equations 27 and 29 are fulfilled, we have a thermionic current and an efficient charge injection over the triangular barrier. Above J_{C1} or J_{C2} one deals with a space-charge limited current. At room temperature, $v_T \sim v_S \sim 10^7$ cm/sec and for $L_i \sim 10^{-5}$ cm ($i = 1$ or 2) one has $J_{Ci} \sim 3$ kA/cm^2. At lower T limitation 1 becomes progressively more restrictive since $J_{C1} \propto T^{3/2}$.

Finally, let us emphasize that in the entire discussion in this section we have neglected all possible *bipolar* effects, such as an accumulation of mobile holes under the triangular barrier. The minority carriers can be injected electrically, as in a bipolar transistor, or optically as in a photodiode. Habib and Board[32] developed a two-carrier model which represents an extension of the theory discussed in this section and includes the effects of minority-carrier storage on the barrier height and the I-V characteristics of a PDB diode. These effects may be important for certain applications discussed by the authors.[32]

C. Potential Effect Transistors

By the physical principle involved, most of the three-terminal semiconductor devices can be classified into either of the two groups: potential-effect and field-effect transistors. In the first group which includes the bipolar transistor, the so-called analog transistors, such as the static induction transistor, (SIT)[33] the permeable base transistor, (PBT)[34] and the thermionic emission transistor, (TET)[35] and a number of ballistic-injection hot-electron transistors,[36] the transistor action results from modulating the height of a potential barrier by an input electrode. The second group, containing a great variety of FETs, employs the modulation of charge in a resistance channel due to screening by the latter of an input field. This classification, of course, is quite relative since both the field-screening and the potential-modulation effects are at work in every transistor. For example, MOSFET in its subthreshold regime should be classified as a potential-effect device.

The concept of analog transistor, as discussed by Shockley,[37] is similar to a vacuum-tube triode in that the current is space-charge limited and is controlled by input electrodes, which create a potential barrier between the cathode and the anode. The similarity ends when one considers the saturation regime of the vacuum device which is due to a fixed-cathode work function. In this sense we can say that the source is a cathode with vanishing work function. Because of this, the drain current characteristics of analog transistors exhibit no saturation. It is important to distinguish the analog transistor from a conceptually different device also proposed by Shockley in 1952, the junction field effect transistor (JFET),[38] in which the current control is effected through the motion of a depletion boundary. This distinction is not always made carefully in the assessment of analog transistors, because in some of the proposed versions these devices can be, in fact, JFETs. For example, when the doping in the transit region is large compared to the injected charge, both the SIT and the PBT become equivalent to a vertical JFET or a vertical MESFET, respectively. The conceptual difference between the two types of transistor is in the physical mechanism of current transport. In the analog transistor, this is a combination of thermionic and space-charge limited current mechanisms, whereas the JFET is, basically, a voltage controlled resistor.

In our view, neither the FET nor the analog transistor has an advantage over one another in the intrinsic speed of operation, when one considers ideal structures of similar feature size. In both cases, the intrinsic delay is limited by the time of flight of carriers from the source to the drain. Strictly speaking, the effective channel length, which determines the delay, is the length of the space-charge accumulation region which coincides with the source-to-drain distance (L) only for long-channel devices. Although in analog transistors the channel is often perpendicular to the semiconductor surface and thus can be made extremely short by MBE, nevertheless the requirement of having a voltage gain usually limits the minimum effective channel length by a characteristic lateral dimension of the controlling electrodes (in the case of the PBT, for example, this dimension is the period of the base electrodes). Another important characteristic (especially for VLSI applications) is the required voltage level which crucially determines the power-delay product and thus limits the possible integration density and the speed of operation. According to this figure of merit, the analog transistors are clearly unfavored if one considers long-channel devices ($L >> 1\ \mu m$). Indeed, the potential due to space-charge accumulation at a given current density scales as L^2 (assuming saturated carrier velocity) and so does, therefore, the required drain voltage. For submicron dimensions, this is no longer a limitation, and in this range the analog transistors can operate at the supply level of 0.5 V and lower. In principle, field effect devices, notable the MOSFET, can also operate at the same low-voltage level. It should be emphasized that for submicron devices one gains no advantage in speed of operation by going beyond threshold because both the output current and the entire capacitively induced charge become proportional to each other. The real limitation is, therefore, due to insufficient accuracy of the threshold voltage control. It is in this respect that analog transistors can have an important

practical advantage. In our view, the only possible advantage of the analog transistor concept is that it can be realized as a bulk (rather than surface) device thus minimizing the threshold variation.

Ballistic hot-electron transistors represent another type of a unipolar potential effect device. The original proposals[39,40] considered a semiconductor-metal-semiconductor (SMS) structure. The problem which has plagued the SMS (and other metal-base) transistors is their poor transfer ratio α (the common-base current gain). Even assuming an ideal epitaxial SMS structure, and extrapolating the base thickness to zero, the typical calculated values of α are unacceptably low — mainly due to the quantum-mechanical (QM) reflections at the base collector interface.[41] The origin of the QM reflection problem can be traced to the large Fermi energy of electrons in a metal base[42] (Section II.C.3). This problem is largely avoided in the monolithic all-semiconductor ballistic hot-electron transistors. A number of such devices have been manufactured recently in III-V compound semiconductors, using MBE-grown planar-doped barriers,[43] heterostructure barriers,[44] or ion-implanted "camel" barriers.[45] It is not clear at this time if these devices offer any potential advantage over the most advanced heterojunction bipolar transistors. An HBT with a graded-gap base[17] can always be expected to have a better current gain, while speed limitations of both types of devices appear to be similar. On the other hand, hot-electron transistors by MBE implemented in silicon may indeed offer a superior frequency performance compared to any other Si device fabricated with a lithographic resolution of 0.5 to 1.0 μm. Hot-electron transistors and their speed limitations will be discussed in Section II.C.2.

1. Analog Transistors

The best-known analog transistors are the static-induction transistor (SIT)[33] and the permeable-base transistor (PBT).[34] In the PBT the controlling electrodes represent a metallic grid embedded in the body of the semiconductor in the transit region between the cathode and the anode (Figure 10a). The PBT represents an improvement over the SIT in which the controlling electrodes are formed by p^+ gate regions (or, generally, by heavily doped regions complementary to the source). The metallic base eliminates parasitic bipolar effects (such as minority carrier storage) and reduces the base resistance. Theoretically, the PBT with undoped transit region offers the highest possible performance of all analog transistors. However, its practical implementation faces considerable difficulties in growing high-quality material over the metallic grid. This problem is particularly severe in the case of Si where epitaxial growth is carried out at high temperatures. Recently reported Si-PBT was therefore fabricated without overgrowth[46] (Figure 10b). Instead, reactive ion etch was used to make grooves in the epitaxial n layer with tungsten deposited on the top of the ridges and the bottom of the grooves forming the collector and the base metallizations, respectively. Frequency performance of the PBT is already comparable to that obtained in the best Si bipolar transistors.[47] The use of Si-MBE can be expected to lead to further improvements.

Another example of the analog device is the thermionic emission transistor, or TET[35] (Figure 11). It differs from the PBT in that the controlling electrodes are located on the semiconductor surface. We shall describe the TET operation in a somewhat greater detail in order to bring out the characteristic limitations of all analog transistors. The device contains an intrinsic base layer grown epitaxially on an n^+ layer which serves as a cathode. An interdigitated grid of anode and gate electrodes on the surface produces a two-dimensional potential distribution in the base layer, as shown in Figure 12. The current flows near the saddle points under the anode electrodes due to thermionic emission over a potential barrier whose height is controlled by an effective potential

$$\overline{V} = \alpha V_A + \gamma V_G \tag{30}$$