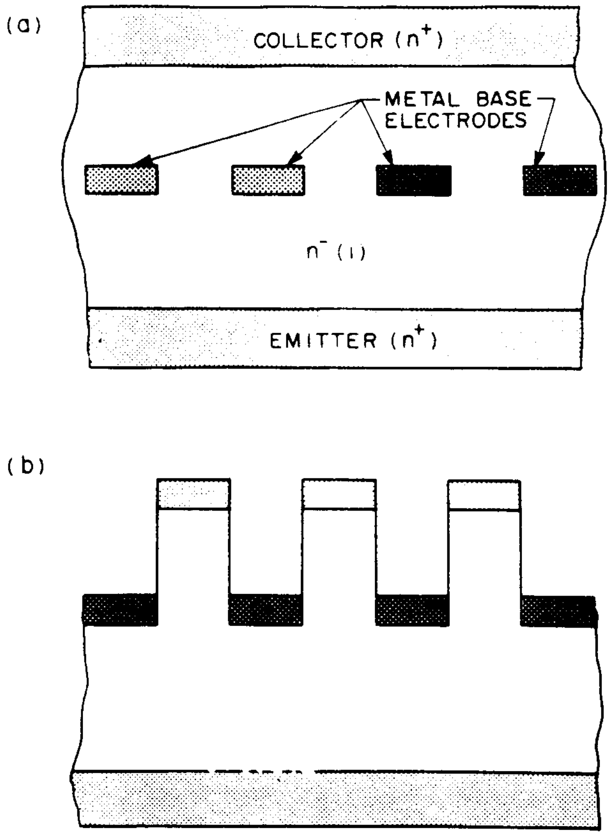

FIGURE 10. Schematic cross section of a permeable-base transistor (PBT) with (a) and without (b) overgrowth.[14,46,47]

where V_A and V_G are the potentials on the anode and gate electrodes, respectively, and α and γ are numerical coefficients which depend on the geometry of the device. The ratio $\gamma/\alpha \equiv K$ determines the voltage gain which must be greater than unity.

For any combination of (V_A, V_G) the electric field beneath the surface can be split into a uniform part which can be considered emanating from a conducting plane at a constant average potential and a nonuniform oscillating part. This procedure can be regarded as a multipole expansion of an appropriate symmetry. Close to the surface we have a "near" zone where the field is mainly multipolar and the oscillation of the potential is appreciable. Far from the surface the potential is uniform and is determined by the field of a parallel plate condenser charged to an average potential $<V>$, namely,

$$<V> = A\,V_A + G\,V_G \tag{31}$$

where $A = S_A/S$ and $G = S_G/S$ are the relative areas of the anode and the gate electrodes, respectively. The lateral inhomogeneity of the electric field averages out exponentially with distance from the surface with a characteristic length λ which is related to the period of the surface electrodes as $\lambda = d/2\pi$. An earlier version[30] of the TET contained a planar-doped charge-sheet barrier $\delta(p^+)$ which had to be built in the i layer in the process of growth by MBE. If the built-in charge sheet is located sufficiently far from the surface (far compared to λ), then the barrier height with respect to the cathode (and therefore the magnitude of thermionic current) is determined by $<V>$, and the voltage gain K reduces to the geometric

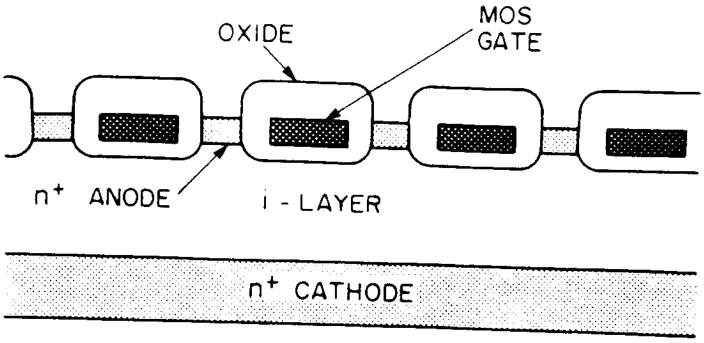

FIGURE 11. Schematic cross section of a thermionic emission transistor (TET). (From Luryi, S. and Kazarinov, R. F., *Solid State Electron.*, 25, 133, 1982. With permission.)

ratio G/A. The requirement L >> λ, where L is the total thickness of the base layer, means that the minimum vertical channel length is restricted by the lateral electrode dimensions. The situation does not appreciably change when one considers the second version of the TET which has no built-in barrier. In this case, model calculations[35] give K = γ/α = 2.5 for G/A = 3, provided one still has λ << L. As discussed in Reference 35, this small degradation of gain is not a stiff penalty to pay for the substantial simplification of the structure. It should be noted that one can obtain an additional gain by sinking the gate electrodes with respect to the anode, i.e., bringing them closer to the cathode, which would make the device similar to a variation of the PBT (this can probably be done only at the expense of substantially increasing the uncertainty in threshold voltage). Nevertheless, consideration of the voltage gain will still constrain the minimum vertical channel length by a characteristic lateral feature size. We believe this is a general limitation of any analog transistor.

Let us now discuss the speed limitations. In the low-current regime of an analog transistor, the current is due to the thermionic charge injection and the characteristics are of the form

$$I = I_0 e^{\beta(\alpha V_A + \gamma V_G)} \tag{32}$$

In this regime, the transconductance grows linearly with the current. In the high-injection regime, the current is space-charge limited and the transconductance saturates. Transition between these regimes is quite similar to that described in Section II.B.2 for PDB diodes. It occurs when the injected current density exceeds either J_{C1} or J_{C2}, given by Equations 27 and 29, respectively. One gains no further increase in speed since in this regime the output current and the entire capacitively stored charge become proportional to one another. Rigorous C/g_m analysis[30,35] leads to a characteristic intrinsic gate delay $\tau = L/v$ of a single device (here $v \sim v_S \sim v_T \sim 10^7$ cm/sec). As discussed above, in a TET one must have L > $d/2\pi$, whence

$$f_{max} = \frac{1}{2\pi\tau} < \frac{v}{d} \tag{33}$$

where, as we recall, d is the period of the interdigitated surface electrode structure. There is also the gate-anode capacitance (which is parasitic, although inherent to the device), which can be minimized to produce an extra delay of about 30%.[30] It is clear that these limitations are not fundamentally different from those of an FET designed with similar lithographic

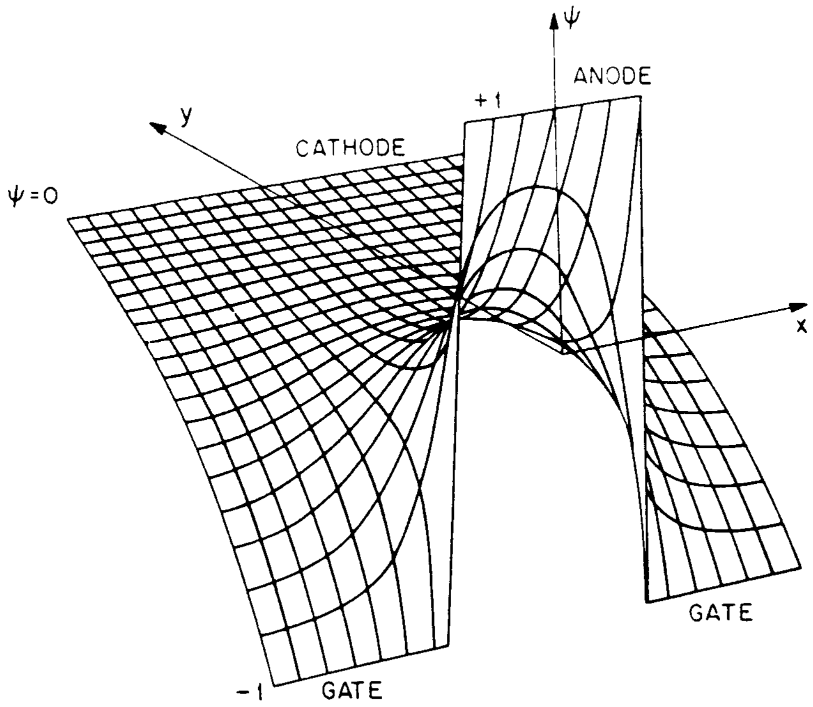

FIGURE 12. Stereometric view of the two-dimensional potential distribution $\psi(x, y)$ in a TET with $S_G/S_A = 2$, $L/d = 1$.

rules. In our view, it is unlikely that any analog transistor by Si-MBE will ultimately beat the MOSFET in the intrinsic speed of operation.

As stated above, the only possible advantage of analog transistors, in our view, is related to the possibility of reducing threshold variations. In the TET, the current transport is by charge injection into an undoped material controlled by a potential barrier removed from the surface. Because the density of injection charge in the intrinsic layer greatly exceeds the background doping, the latter should not affect the potential distribution. On the other hand, the influence of the Si-SiO_2 interface traps is also diminished because the space-charge accumulation occurs far from the surface. With the state-of-the-art MOS technology, one can expect to be able to control the threshold voltage to within a 25-mV margin. This estimate does not include the effect of possible variations in the geometry of the electrodes which may be the only serious source of uncertainty.

Potentially, the TET can be operated at very low-voltage swings, with the total supply voltage less than half the energy gap of Si. This feature is very attractive for the implementation of complementary logic circuits similar to the celebrated silicon CMOS. The low supply voltage should eliminate all parasitic bipolar (latch-up) effects. As an illustration, we shall consider below an inverter circuit[35] based on two TET devices with complementary types of conductivity. Figure 13 shows schematically the layout of one inverter gate and its circuit diagram. Let us describe the structure in more detail. One starts from a *p*-silicon substrate which serves as the source for the *p*-type transistor and is at the supply voltage level (VDD = 0.5 V). The source (cathode) of the *n*-transistor is provided by a deep diffusion of donors in the substrate and is at the ground (GND) voltage level. An epitaxial intrinsic layer of thickness L serves as a base for both transistors of the inverter pair. The gate and the anode terminals are arranged as an interdigitated pattern of electrodes on top

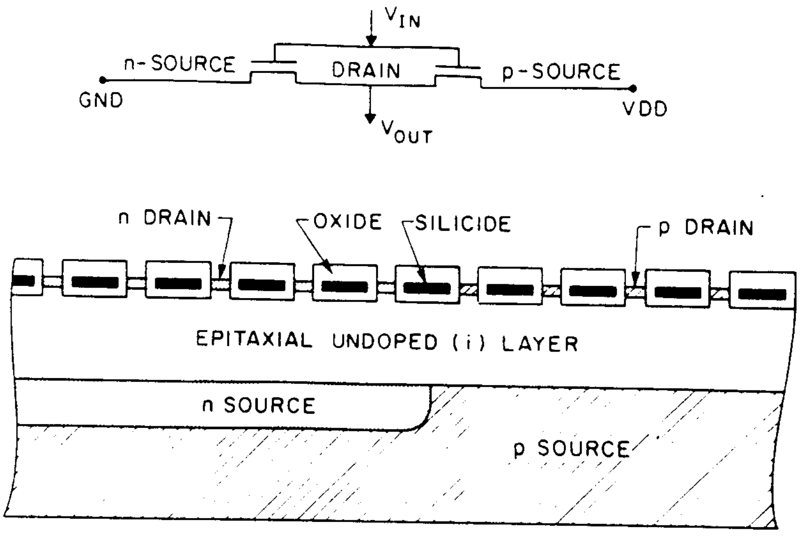

FIGURE 13. Schematic layout and equivalent circuit of a complementary TET inverter logic element.

of the intrinsic layer. The gate electrodes represent an MOS structure (silicide-gate oxide-silicon) common for both transistors. The anode electrodes are in contact with the top of the intrinsic layer. The contact is made ohmic for electrons in the n-type transistor and holes in the p-type device (e.g., by n^{+} and p^{+} polysilicon drain structures). The anodes are connected and their common potential is the output voltage of the inverter.

The power supply lines to the circuit run entirely underneath the base intrinsic layer with the voltage applied between the n-diffusion region and the p-substrate. The sources of the two transistors are biased by VDD with respect to one another and form a *forward*-biased pn^{+} junction. At VDD = 0.5 V the power dissipation associated with the forward current is negligible even at room temperature. For example, for a diode with $N_D = 3 \times 10^{20}$ cm^{-3} and $N_A \sim 5 \times 10^{16}$ cm^{-3} the experimental value of the forward current density at 0.5 V is about 10 mA/cm^2. The only negative consequence of this current is that it draws on the supply battery.

Electrically the circuit represents two variable resistances in series which divide the VDD to GND voltage depending on the potential on the gate. If an appropriate silicide (e.g., $TaSi_2$) is used for the gate metal, then the two devices may be regarded as symmetrical.* Both transistors are "normally off", i.e., either of them is in the "off" state when the gate is at zero voltage with respect to its source. Thus, the n-transistor is "off" and the p-transistor is "on" when V_G = GND. In this state V_{OUT} = VDD. Conversely, when V_G = VDD, then the p-transistor is "off" and V_{OUT} = GND.

The transfer characteristic (V_{OUT} vs. V_{IN}) of a TET inverter can be obtained by a graphical construct (Figure 14). As seen from the anode voltage (source to drain) for the n-transistor equals the output voltage, $V_A^{(n)} = V_{OUT}$, whereas for the p-transistor $V_A^{(p)}$ = VDD − V_{OUT}. Since the currents in both transistors are equal, we can superimpose the plots $I_n(V_{OUT})$ and $I_p(V_{OUT})$ as shown in Figure 14a and read off the transfer characteristic from the intersections of curves corresponding to same gate voltage. Figure 14b shows the charac-

* Asymmetry is introduced to a small degree by differences in v, and A* for electrons and holes and to a greater degree by different built-in barriers, V_{bi}. The latter can be further adjusted by a judicious choice of the doping levels N_D and N_A or by ion implantation at interface. To this end it may be advantageous to use n-substrate and p^{+} diffusion rather than p-substrate and n^{+} diffusion.

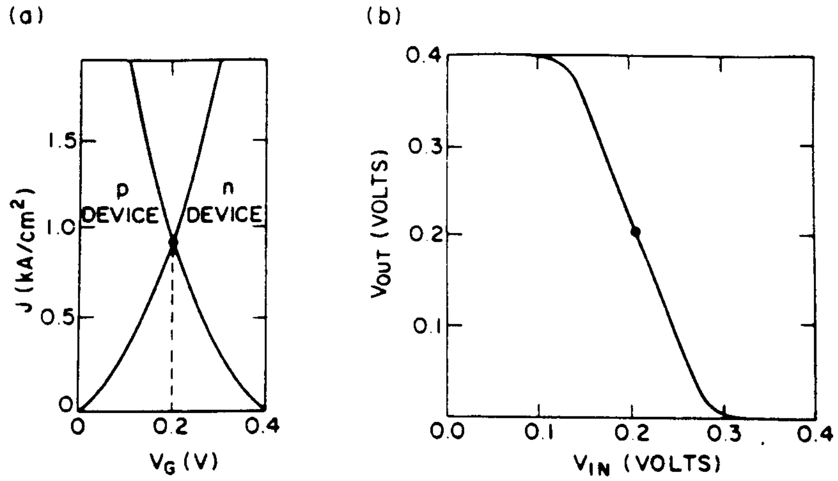

FIGURE 14. Calculated room-temperature characteristics of an inverter composed of two symmetric complementary silicon TET devices. Parameters assumed are $L = d = 0.6$ μm, $S_G/S_A = 3$, and VDD = 0.4 V. (a) Graphical superposition of I-V characteristics; (b) inverter transfer characteristic. (From Luryi, S. and Kazarinov, R. F., *Solid State Electron.*, 25, 133, 1982. With permission.)

teristic obtained in this way for an inverter composed of a complementary symmetric pair of TET devices with an exemplary set of parameters.

It may be instructive to give an approximate analytic derivation of the transfer characteristic. Note that all points of the curve in Figure 14a correspond to the low-current regime, which physically means that the inverter in its steady state draws only a small current. In this regime the current-voltage characteristics are nearly exponential and are well-described by dependences of the form in Equation 32 with $I_0 \equiv I_n$. For the p-type transistor, the dependence analogous to Equation 32 is obtained by shifting the source voltage (compare the equivalent circuit in Figure 13).

$$I = I_p\, e^{\beta[\alpha(VDD - V_A) + \gamma(VDD - V_G)]} \tag{34}$$

For a symmetric pair of devices, one has $I_n = I_p$. Equating the currents, we find in this case

$$\alpha\, V_A + \gamma\, V_G = (\alpha + \gamma)\, VDD/2 \tag{35}$$

Equation 35 correctly describes the central part of the transfer characteristic including its slope γ/α which, thus, equals the voltage gain K of a single device at low currents. It does not describe the flat portion of the transfer curve. Indeed, for $V_A \lesssim kT/q$ Equation 32 is invalid because it neglects the reverse diode current. Similarly, Equation 34 is invalid when $VDD - V_A \lesssim kT/q$. For a nonsymmetric case, $I_p \neq I_n$, the transfer curve will be shifted by the amount

$$\frac{kT}{q} \ln(I_p/I_n) = V_{bi}^{(n)} - V_{bi}^{(p)} \tag{36}$$

where $V_{bi}^{(n)}$ and $V_{bi}^{(p)}$ are the built-in voltages for the n- and the p-transistors, respectively. The inverter delay time, $\tau_{inv} = C/g_m$, where C is the total capacitance of one inverter stage and g_m is the transconductance of the driving device in its "on" state, was estimated[35] for an inverter with the period of surface electrodes $d \sim 0.6$ μm and other parameters as in Figure 14. The calculated delay is of order $\tau_{inv} = 5$ psec.

It is important to realize that in a real integrated circuit the speed of operation of a TET-based inverter can be expected to approach by an order of magnitude the above "intrinsic" gate delay. Indeed, the current flowing in this inverter during switching is of the order of that in a CMOS FET inverter while the switching voltage is an order of magnitude lower. Accordingly reduced will be the charge associated with all parasitic capacitances such as wiring, interconnect, etc. and the corresponding parasitic delay times. In other words, this means that for reasonable dimensions of the TET, say $S = 10 \times 10$ μm², the inverter capacitance C will be of the same order as the total parasitic capacitance. This situation is common to many other potential-effect transistors, whose transconductance scales with the device area and thus affords higher current-drive capabilities.

2. *Hot-Electron Transistors*

As the dimensions of semiconductor devices shrink and the internal fields rise, a large fraction of carriers in the active regions of the device during its operation are in states of high kinetic energy. At a given point in space and time, the velocity distribution of carriers may be narrowly peaked, in which case one speaks about "ballistic" electron packets. At other times and locations, the nonequilibrium electron ensemble can have a broad velocity distribution — ususally taken to be Maxwellian and parameterized by an effective electron temperature $T_e > T$, where T is the lattice temperature. Hot-electron phenomena have become important for the understanding of all modern semiconductor devices. Moreover, a number of devices have been proposed whose very principle is based on such effects. This group of devices will be reviewed in the present section.

We shall be concerned only with the hot-electron injection devices, i.e., such devices in which hot carriers are physically transferred between adjacent semiconductor layers. Two distinct classes of such devices can be identified — depending on which of the two hot-electron regimes is essentially employed (the ballistic or the T_e regime).

In the electron-temperature devices,[48] the heating electric field is applied parallel to the semiconductor layers, with hot electrons then spilling over to the adjacent layers over an energy barrier. This process is quite similar to the usual thermionic emission — but at an elevated effective temperature T_e — and the carrier flux over a barrier of height Φ can be assumed proportional to exp $(-\phi/kT_e)$. Even though a small fraction of electrons — those in the high-energy tail of the hot-carrier distribution function — can participate in this flux, their number is replenished at a fast rate determined by the energy relaxation time, so that the injection can be very efficient.

In the ballistic devices,[36] electrons are injected into a narrow base layer at a high initial energy in the direction normal to the plane of the layer. The typical ballistic semiconductor transistors are illustrated in Figure 15. Their performance is limited by various energy-loss mechanisms in the base and by the finite probability of a reflection at the base-collector barrier. Most of these structures employ heterostructure band discontinuities available in GaAs/AlGaAs systems, but the PDB transistor is quite within the immediate reach of Si-MBE.

One should understand the main trade-off involved in the design of all ballistic transistors with a doped semiconductor base: cooling of hot electrons by phonon emission and other inelastic processes (minimized by thin base layers) against the increasing base resistance for thinner layers. It is easy to estimate the RC delay associated with charging the working base-emitter capacitance and the parasitic base-collector capacitance through the lateral base resistance:

$$RC = \tau_b = \frac{\kappa L^2}{\ell \mu \sigma} \tag{37}$$

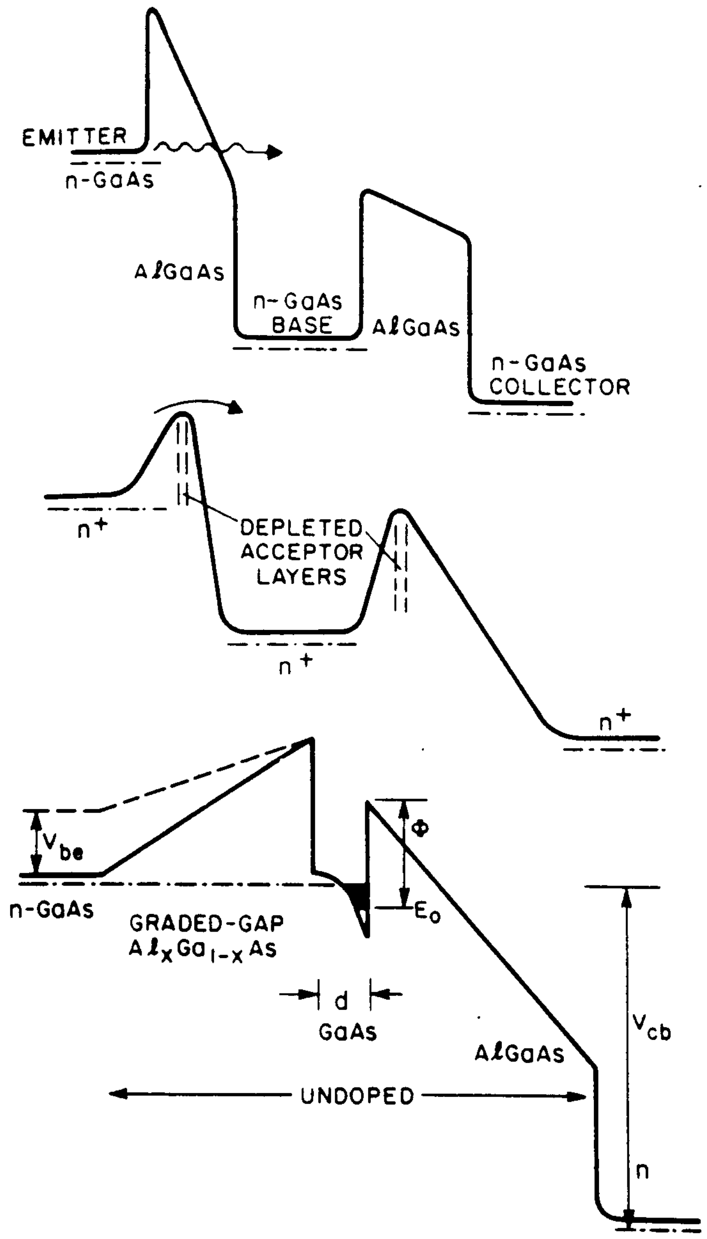

FIGURE 15. Ballistic hot-electron transistors with a monolithic semiconductor structure. (a) Tunnel-emitter transistor (THETA);[36,44] (b) planar-doped-barrier (PDB) transistor;[43,45] (c) induced-base transistor (IBT).[42]

where ℓ is the thickness of the emitter or the collector barriers, $\ell \sim 10^{-5}$ cm, L is the characteristic lateral base dimension (shortest distance to the base contact from the geometric center of the base, L $\sim 10^{-4}$ cm), μ is the mobility in the base, σ is the mobile charge density per unit base area, and κ is the dielectric permittivity. For a hot-electron transistor to be competitive, one must have $\tau_b \approx 1$ psec, which means that th : sheet resistance in the base must be $(\mu\sigma)^{-1} \lesssim 1$ $k\Omega/\square$. The base thickness cannot be made larger than the hot-electron mean free path (séveral hundred angstrom) in silicon, otherwise a strong degradation in the transfer ratio α (the common-base current gain) will occur due to various energy-loss mechanisms. This places stringent requirements on the MBE grower to provide a heavy doping in the base without too much degradation of the mobility. The limitation expressed by Equation 37 is rather severe. The minimum value of L is governed by the lithographic resolution. One cannot really make the barrier thicknesses ℓ much larger than 1000 Å, since this would introduce the emitter and the collector delays of more than 1 psec.

Before leaving the subject of ballistic transistors, let us briefly discuss their potential frequency performance. It is sometimes stated that hot-electron transistors are capable of subpicosecond operation because such is the time of flight of ballistic electrons across the base. That is a much too often repeated fallacy: the time of flight through the base has nothing to do with the intrinsic device speed. Like the bipolar, the FET, and the analog transistor, hot-electron devices have a regime in which their output current I rises exponentially with the input (base-emitter) voltage. In this regime, the maximum speed of operation is proportional to I. However, like every exponent in nature, this dependence eventually saturates and goes over into a linear law. One gains no further advantage in speed by increasing I since the charge stored in all input capacitances will rise proportionally. Ultimately, the speed of a transistor is determined by the current level at which one has a crossover between the exponential and the linear regimes. In transistors with a thermionic emitter, this crossover occurs because of the accumulation of the mobile charge diffusing up the emitter barrier and drifting down the collector barrier. This always leads to the characteristic delays $\tau_e = \ell_e/ v_T$ and $\tau_c = \ell_c/v_S$, where ℓ_e and ℓ_c are the thicknesses of the emitter and the collector barriers, respectively, v_T is the thermal velocity of carriers, and v_S is their saturated drift velocity. Of course, neither of the ℓ's can be shrunk below, say, 1000 Å — because of the complementary limitation expressed by Equation 37. We conclude that an ideally optimized ballistic transistor will be a roughly 3-psec device.

3. Metal-Base Transistors

A variety of MBTs have been proposed differing by the materials employed and by the physical mechanism of hot-electron injection into the base (Figure 16). The original MOMOM proposal by Mead[49] (Figure 16a) was based on electron tunneling from a metal emitter through a thin oxide barrier into a high-energy state in a metal base. Another insulating barrier separated the base from a metal collector electrode. Later versions of this device[50] had the second MOM replaced by a metal-semiconductor junction, resulting in a transistor structure called the MOMS (Figure 16b). Tunnel-emitter MBT concepts have not gained much development in recent years.

Metal-based transistors (MBT), which employ thermionic rather than tunneling injection of hot carriers into the base, were first proposed by Atalla and Kahng[39] and Geppert[40] in the form of a metal-semiconductor (SMS) structure. The basic SMS transistor is illustrated in Figure 16c. Experimental studies of the SMS device are actively pursued to this oay: recent advances[51,52] have been associated with the development of epitaxial techinques for the growth of monolithic single-crystal silicon-metal silicide-silicon structures. This continued interest is explained not only by the scientific usefulness of the SMS structure (it is an excellent tool for studying fundamental properties of hot-electron transport through thin films), but also by lingering hopes to produce a transistor which is faster than the bipolar or FET devices.

The potential merits of the SMS transistor had been appraised long ago by Sze and Gummel.[41] They predicted that despite its possibly superior frequency performance, this device would hardly ever replace the bipolar junction transistor. The problem which has plagued the SMS (and all other metal-base) transistors is their poor transfer ratio α (the common-base current gain). Even assuming an ideal monocrystalline 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) reflection of electrons at the base-

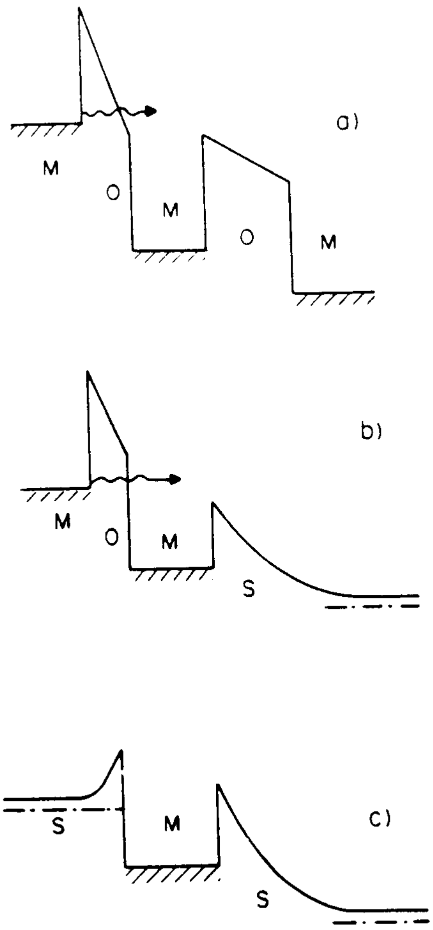

FIGURE 16. Metal-base transistors. (a) MOMOM; (b) MOMS; (c) SMS.

collector interface. In our view, these conclusions of the 1966 paper[41] remain valid today. The origin of the QM reflection problem can be traced to the large Fermi energy of electrons in a metal base.[42] Indeed, consider an (over-) simplified model of a metal-semiconductor barrier, Figure 17, and assume parabolic energy-momentum relationships in both materials. The well-known solution of this QM problem gives for the above-barrier reflection coefficient R the following expression:

$$R = \left(\frac{1 - \zeta}{1 + \zeta}\right)^2, \quad \text{where} \quad \zeta \equiv \sqrt{1 - \Phi/E} \tag{38}$$

E is the hot-electron energy in the base, and Φ is the barrier height. Note that it is not the clearance E − Φ, but the ratio E/Φ which enters the expression for R, and hence one must correctly choose the zero energy level — including a large Fermi energy, E_F. Typically, Φ/E is close to unity and the reflection is large. For a ballistic electron in Al incident on the interface with GaAs at 0.4 eV above the Schottky barrier (Φ ≈ 12 eV), the probability of reflection predicted by Equation 38 is about 50%. Of course, estimates based on the simplest free-electron model of reflection do not hold even approximately for metals with a complicated band structure and indirect-gap semiconductors, such as silicon. One should expect, however, that the band-structure effects would only further inhibit the QM transmission.[53]

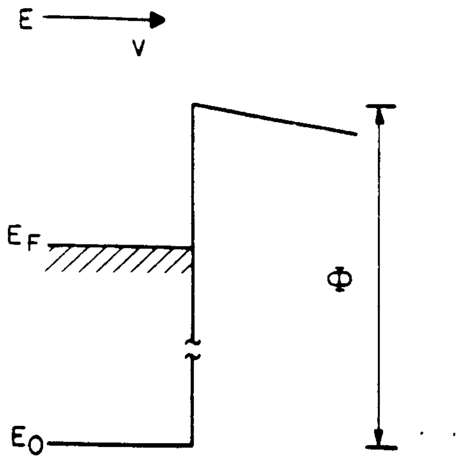

FIGURE 17. Illustrative model for the above-barrier reflection of ballistic electrons.

We are not aware of any metal-semiconductor pair where the exact solution would predict a lower reflection of hot electrons than that given by Equation 38.

Nevertheless, there have been recent reports[51,52] of a transistor action in monocrystalline Si/CoSi$_2$/Si structures with α as high as 0.6. One cannot, of course, rule out some "accidental" band-structure resonance in such systems which could aid the QM transmission of hot electrons. Such an interpretation, however, appears to us unlikely. A more probable explanation is related to the existence of pinholes in the base metal film, i.e., continuous silicon "pipes" between the emitter and the collector. A careful analysis[54] of the correlation between the pinhole sizes and the device characteristics found no evidence for any hot-electron component of the current through the base. On the other hand, the pinhole conduction in some cases gives α as high as 0.95. Such a device then becomes a "natural" version of the permeable-base transistor. This version had, in fact, been fabricated by Lindmayer[55] prior to PBT. In our view, the thermionic-emission mechanism of the current through a permeable base has a greater device potential than the hot-electron transport through a metal base. Thin silicide films may offer an attractive way of fabricating the PBT — if one learns how to control the statistics of pinhole sizes, making it sharply peaked at a desired area scale.

III. HETEROEPITAXIAL DEVICES

Heteroepitaxial growth studies by MBE have included various combinations of silicon with insulators, metals, and other semiconductors. The ultimate thrust of such studies is to create a new range of possibilities for VLSI and optoelectronics. It is impossible to overestimate the importance of this research. One group of the proposed applications assumes an active role of Si layers in the compound structures and utilizes the special properties of epitaxial junctions and superlattices. The other treats the silicon wafer as a carrier vehicle for growing self-contained device structures based on a foreign heteroepitaxial material. In the present section we shall be concerned mainly with the former category of heteroepitaxial device structures (which may be termed "commensurate"), while the latter category will be discussed in Section IV. We begin with semiconductor heterojunctions; devices based on the growth of epitaxial insulators and metals will be discussed in Section III.B.

A. Semiconductor Heterostructures

As mentioned in Section I, the only lattice-matched semiconductors (GaP and AlP) are chemically incompatible with Si and relatively little has been reported on the device use of

GaP/Si combinations. An interesting attempt to circumvent some of the difficulties associated with the growth of GaP on Si was made by Wright et al.[56] who obtained a higher-quality GaP/Si interface, by using <211> silicon substrates. According to their theoretical arguments, this unconventional orientation may suppress the formation of antiphase domains and reduce the interface charge density. Wright et al.[56] fabricated GaP/Si heterojunction bipolar transistors with a common-base current gain, $\alpha \approx 90\%$. Of course, in this approach one has to abandon the hope for integrating the heterojunction devices with VLSI circuits on the same wafer, since present-day Si technology uses almost entirely wafers of <100> orientation. On the other hand, a single-domain GaP layer may prove to be useful as an intermediate buffer for a subsequent growth of device quality III-V compound semiconductor layers. This would be an example of the incommensurate-epitaxy program — in which Si is used merely as a low-cost substrate for the growth of compound semiconductors. Still, we feel there is more to be gained by using <100> substrates even in such applications, since that would open a way for the most attractive VLSI chip architecture in which III-V compound semiconductor layers will form special-purpose islands — enhancing rather than replacing Si circuits (Section IV).

The nearest thing to bandgap engineering in silicon has been associated with Ge/Si systems. Attempts have been made to reproduce some of the successes of the GaAs/AlGaAs heterostructure technology, such as the modulation doped transistors and heterojunction bipolar transistors. This work (discussed in Section III.A.3) is yet at an early stage and it is not clear if it will ultimately lead to the development of a useful device. Another important application of Ge/Si MBE technology has been associated with attempts to fabricate an integrated receiver system for fiber-optic communications, discussed next.

1. Ge IR Photodetectors on a Si Chip

As is well known, the celebrated silicon technology has not been able to produce an on-chip IR photodetector for long-wavelength fiber-optic communications. The obvious difficulty lies in the fact that silicon bandgap E_G is wider than the photon energy in the range of silica-fiber transparency ($\lambda = 1.3 - 1.55$ µm). Attempts have been made to overcome this difficulty by using MBE-grown Schottky-barrier structures with photoexcitation of carriers from metal (or silicide) into silicon.[57] The threshold for such a photoeffect is determined by the Schottky-barrier height and can easily match the required IR range; however, the quantum efficiency of absorption in such structures is usually low. So far, the only practical way of employing silicon technology for fiber-optic communications has been to combine silicon integrated circuits with germanium or InGaAsP detectors on a separate chip.

A different approach to this problem is based on growing single-crystal germanium *pin* junction on a silicon substrate.[58,59] The MBE-grown diodes reported in Reference 58 had a quantum efficiency $\eta \approx 40\%$ at $\lambda = 1.3$ µm. Figure 18 shows the measured photoresponse spectra at 300 and 77 K. The photocurrent threshold, as well as its shift with the temperature, agree with the well-known absorption data for bulk Ge. One can see the transition from indirect to direct absorption which occurs with increasing photon energy at $\lambda = 1.45$ µm (77 K) and $\lambda = 1.59$ µm (300 K). However, the devices suffered from a relatively high reverse-bias parasitic leakage at room temperature. This leakage resulted from threading dislocations originating at the Ge/Si interface due to a large lattice mismatch and propagating through the germanium *pin* junction. From transmission electron micrograph (TEM) of the structure, the density of threading dislocations in the working region of the diodes was estimated to be as high as 10^9 to 10^{10} cm.$^{-2}$

In the subsequent work,[59] the dislocation density was reduced by a novel trick called "glitch grading". Both the original and the improved structures are schematically shown in Figure 19. For the intended application, one was concerned only with the possibility of growing high-quality Ge layers — with whatever intermediate layers were necessary to

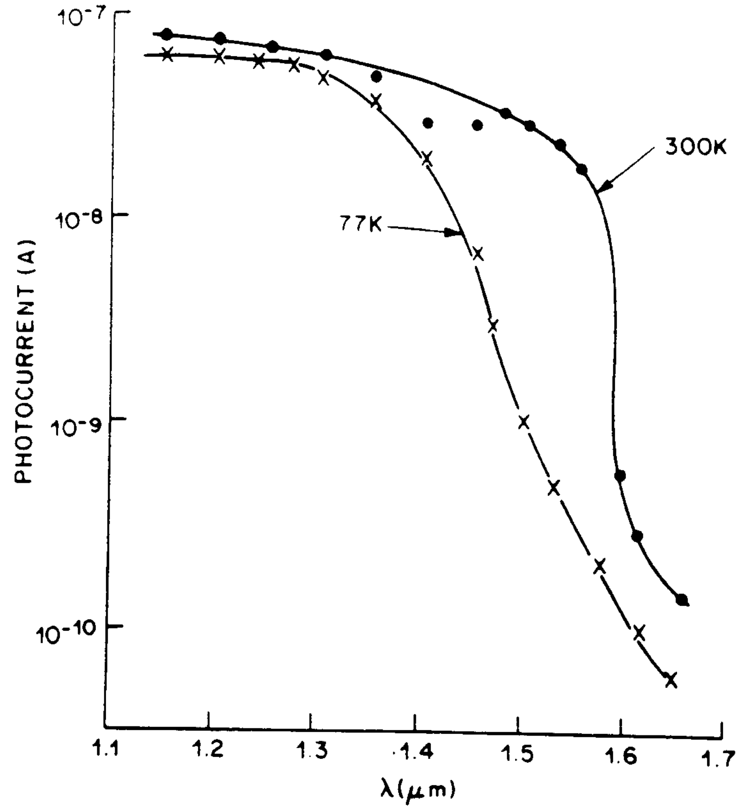

FIGURE 18. Photoresponse spectra of a Ge *pin* diode epitaxially grown on a silicon substrate. (From Luryi, S.., Kastalsky, A., and Bean, J. C., *IEEE Trans. Electron. Devices*, 31, 1135, 1984. With permission.)

achieve the material quality in the optically active working layers. In the original structure (Figure 19a), this was attempted to achieve by inserting a thick (1.5-μm) homogeneous Ge buffer layer between the active germanium layer (region where photogenerated carriers are separated by the electric field) and the Ge/Si transition region. A novel feature of the improved diode is the addition of a $Ge_{0.70}Si_{0.30}$/Ge superlattice (referred to as "glitches") within the buffer layer (Figure 19b). It had been shown previously[12] that such superlattices could be grown, despite lattice mismatch, without nucleating dislocations. Mismatch is instead accommodated by a distortion of the alloy layers — such that the alloy and Ge lattice parameters match in the plane of growth. It was hoped that the strain associated with these distortions could then be used to trap dislocations propagating up from the Si-Ge/Si interface. Different periodicities and Ge_xSi_{1-x}/Ge superlattice compositions were grown and studied by TEM. It turned out that ten periods of 100 Å $Ge_{0.70}Si_{0.30}$/500 Å Ge were indeed effective at drastically reducing dislocation propagation up into the active region of the detector: the dislocation density above the glitch region was reduced by two orders of magnitude. Similar improvement was seen[58] in the room-temperature current-voltage characteristics. At comparable reverse-bias voltages, the leakage current dropped by a factor of more than 100 and became within an order of magnitude from the theoretical diffusion-limited saturation current (approximately 4×10^{-4} A/cm^2) in an ideal Ge *pn* junction.

These results had demonstrated the power of the glitch-grading method to improve the quality of heteroepitaxial germanium. Ultimately, the MBE should be able to produce on a

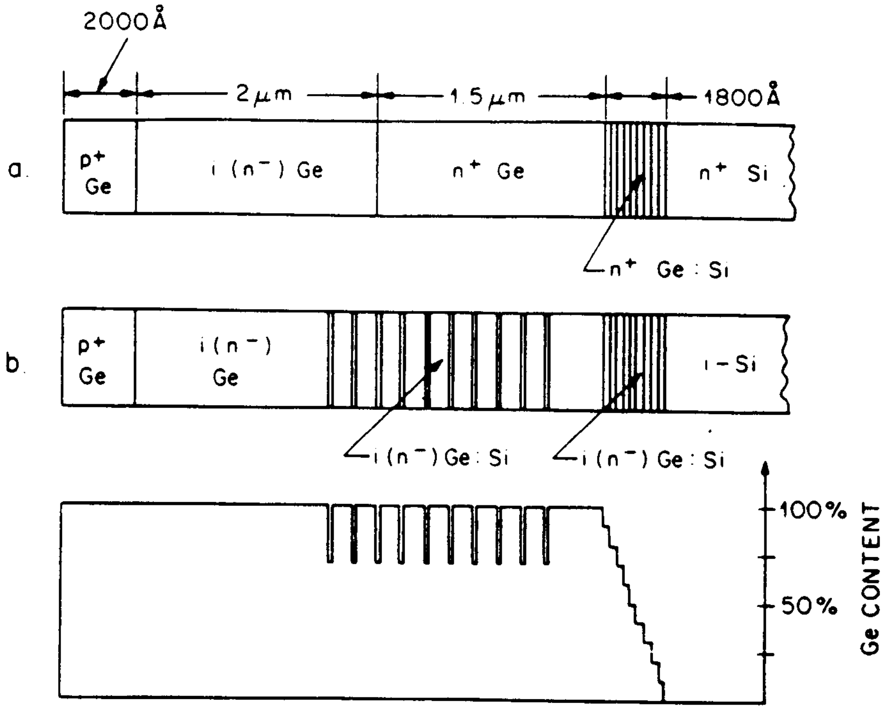

FIGURE 19. Schematic illustration of the composition of epitaxial layers in Ge/Si photodetectors. (a) The original structure[58]; (b) glitch-graded structure.[59]

silicon substrate germanium layers comparable in quality to bulk Ge samples. It is essential that the fabrication sequence is quite compatible with Si-VLSI technology.

The described Ge *pin* detectors should, of course, be classified as an example of the incommensurate epitaxy, since no use has been made of the properties of a Ge/Si interface. This section has been placed here (rather than in Section IV where it belongs) in order to open a discussion of the general concept of epitaxial detectors on silicon. It is clear that germanium is not an ideal high-speed photodetector. Indeed, in order to satisfy the requirements of both speed and efficiency, one must use direct optical transitions in germanium ($h\nu \lesssim 0.8$ eV). For these transitions, the absorption coefficient is of the order of 10^4 cm^{-1}, which corresponds to an effective absorption length ≈ 1 μm and an intrinsic delay of about 10 psec. However, the intrinsic carrier concentration in Ge and therefore the ideal reverse-bias saturation current are associated with the indirect bandgap of 0.66 eV. The fundamentally larger dark current of any germanium detector (compared to direct-gap InGaAs detectors which absorb light and leak dark current through the same bandgap) results in an inferior noise performance. There is no fundamental reason why InGaAs detectors could not be grown by MBE on Si substrate and integrated on-chip with silicon amplifying circuits. Again, this sort of development would be within the realm of incommensurate epitaxy.

However, there is one property of silicon, which is very attractive for use in fiber-optic communications and whose utilization requires commensurate epitaxy. Silicon is an ideal material for avalanche multiplication of photogenerated signals. Neither Ge nor InGaAs are ideal avalanche photodetector (APD) materials from the point of view of the so-called excess noise factor F, which describes the stochastic nature of avalanche multiplication.[60,61] The F factor generally depends on the avalanche gain M and the ratio of the impact ionization coefficients K = α_n/α_p for electron and holes. If K ≈ 1, then F ≈ M and the total noise power scales ∝ M^3. Such is the situation for Ge with $\alpha_p/\alpha_n \lesssim 2$ and InGaAs, where α_n/α_p

≤ 2. On the other hand, if $K >> 1$ or $K << 1$, then $F \approx 2$ even for $M >> 1$, provided avalanche is initiated by the type of carrier with higher α. It is well established [61] that in Si at not too high electric fields ($\leq 3 \times 10^5$ V/cm) the electron ionization coefficient is substantially greater than the hole ionization coefficient. Thus, properly designed Si APDs can have the noise performance near the theoretical minimum. At present, there are commercially available silicon devices with $K \approx 20$ to 100 (of course, these APDs do not operate in the range of interest for fiber-optical communications). It would be very attractive to implement a heterostructure device with separate absorption and multiplication regions (SAM APD),* in which electrons photogenerated in a Ge or InGaAs layer would subsequently avalanche in Si. An example of such a structure[64] is shown in Figure 20. It combines Si multiplication with Ge absorption layers. It also contains a depleted layer of acceptors (charge sheet) built in silicon in the vicinity of the Ge interface, whose purpose is to separate the low-field region in the optically active Ge layer from a high-field Si layer where avalanche multiplication occurs. Similar SAM APD structures with high-low electric field profiles have been successfully fabricated in III-V compound semiconductors[65] however, the implementation with Ge/Si heterojunctions requires far better material quality in the interfacial layers than that presently available with any crystal growth technique. We shall return to the discussion of possible Ge/Si SAM APD structures in Section III.A.3.

2. *Strained-Layer Ge_xSi_{1-x}/Si Systems and Modulation-Doped Transistors*

As discussed in the preceding section, germanium layers grown on a silicon substrate are replete with dislocations originating from defects of the lattice-mismatched interface between Ge and Si. The situation is substantially similar when one attempts to grow thick Ge/Si alloy layers on Si. On the other hand, it had been established long ago[66] that under ultrahigh vacuum conditions thin Ge_xSi_{1-x} alloy layers can be grown on Si almost pseudomorphically (i.e., with few dislocations) and then capped by another Si layer. These results, however, were restricted only to layers with low germanium contents ($x \leq 0.2$), and thicknesses in the range 10 to 100 Å, which correlated with the existing theoretical calculations.[67] based on thermodynamic stability considerations. Recent advances[12,68] in low-temperature MBE growth techniques have radically altered this situation. It has been found that strained-layer Ge_xSi_{1-x}/Si heterostructures and superlattices can be grown virtually free of dislocations with various compositions and thicknesses up to 1 µm. At the same time, new and unexpected electronic and optical properties arising from strain were discovered in these material systems, thus substantially enhancing the potential versatility of Si-based technologies. The most salient features of the new findings are summarized below; for a complete review the reader is referred to the chapter by Bean in this book.

Let us first discuss the questions of stability. The maximum thichness h_c of a single strained Ge_xSi_{1-x} alloy layer which can be grown pseudomorphically on Si depends on the germanium content, decreasing with x. People and Bean[69] have calculated $h_c(x)$ on the assumption that the film grows initially without dislocations, which are then generated at the interface, as the strain energy density per unit area of the film exceeds the areal energy density associated with an isolated dislocation. **Their result, which implicitly gives $h_c(x)$ in angstroms by the equation:

* Considerable research has been devoted to the use of III-IV compound-semiconductor SAM APDs for fiber-optic communications (Reference 62 and references therein). Excellent performance has been demonstrated by $InP/Ga_{0.47}In_{0.53}As$ APDs of this type.[63]

** Recently, Luryi and Suhir[72] considered theoretically the growth of lattice-mismatched materials on *patterned* silicon surfaces. They showed that in such structures the strained region of the heteroepitaxial film can be confined to a narrow zone near the interface, so that the total strain-energy density per unit area of the film remains finite as $h \rightarrow \infty$. This opens the possibility of dislocation-free growth of lattice-mismatched films of *arbitrary* thickness.

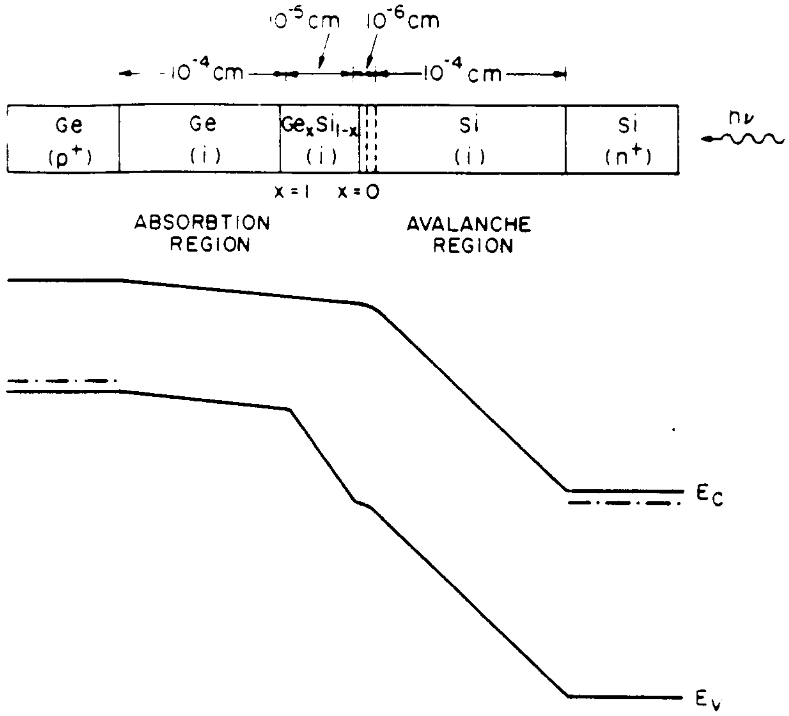

FIGURE 20. A possible Ge/Si hi-lo SAM APD structure[64] with separate absorption and multiplication regions and high-low electric field profile. Its implementation requires further improvement in the quality of the interfacial germanium layers. Large number of misfit defects, resulting in a high parasitic dark current, may be very difficult to avoid.

$$x^2 h_c = 10.8 \ln (h_c/4) \tag{39}$$

is in an excellent agreement with the empirical data. Raman scattering studies[70] have shown that most of the strain in such structures resides in the alloy layer, with Si cladding layers being nearly unstrained. A second Ge_xSi_{1-x} layer can then be grown on the Si cap layer (provided the latter is two to three times thicker than the alloy layer), and the sequence can be repeated many times without a noticeable incommensurate growth (as many as 100 periods have been reported). The maximum total thickness of such strained-layer-superlattices (SLS) can be estimated from the semiempirical rule[71] that the stability of the SLS against the formation of dislocations is equivalent to that of a single alloy layer of same thickness, but average Ge content. This rule can be represented by the following expression:

$$h_{SLS}^{MAX}(x, r, T) \approx h_c(xr) \tag{40}$$

where $r \equiv h/T$ is the ratio of the thickness of the alloy layer to the superlattice period (i.e., the "duty cycle" of the superlattice), and h_{SLS} is the total superlattice thickness. Note that $h_{SLS}^{MAX} \neq f\,(T)$, which means that a coarse superlattice with few periods will be as stable as a fine superlattice with many periods, provided they have the same total thickness and the same values of x and r.

Next we discuss the effects of strain on the band structure of an alloy layer. An important finding in this regard is the theoretical calculation of People,[73] who considered the bandgap narrowing in strained Ge_xSi_{1-x} alloys grown on Si (100) substrates and found that the gap is substantially reduced in comparison with the unstrained alloy. Lang et al.[74] have measured

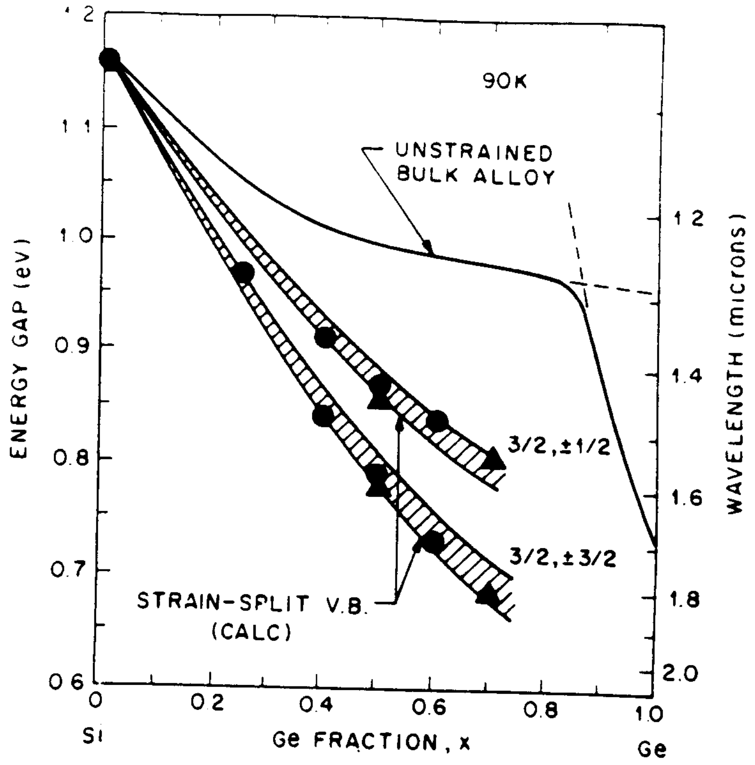

FIGURE 21. The energy-gap values in strained Ge_xSi_{1-x} alloy layers, including quantum-well corrections (circles correspond to 75-Å wells; triangles correspond to 33 Å wells) at 90 K. (Courtesy of R. People.) The data points were obtained from optical absorption experiments.[74] The double points at the same values of x correspond to a splitting of the valence band by strain. Theoretical calculations[73] are indicated by the cross-hatched bands. Also included in the figure are the data calculated for unstrained bulk alloys.

the fundamental absorption threshold in the Ge_xSi_{1-x}/Si SLS as a function of the Ge content in the alloy layers and found a good agreement with the theoretical predictions.[73] Summary of the energy-gap values for strained alloys is presented in Figure 21. At x = 0.6, the bandgap E_G is narrower than that of pure unstrained Ge, and for $x \gtrsim 0.5$, one has $E_G \leq$ 0.8 eV. The absorption edge is thus brought down by the strain to below the photon energy at wavelengths of silica-fiber transparency (λ = 1.3 − 1.5 μm). These findings are of great importance for optical applications of the Ge/Si SLS discussed in the next section.

Another question of great importance for applications is that of heterojunction band offsets. This question has often been the subject of intense controversy,[75] both theoretical and experimental. Even in the best studied $GaAs/Al_xGa_{1-x}As$ heterojunctions, the conduction (ΔE_C) and valence (ΔE_V) band discontinuities, which for years had been believed to be related in the proportion 85:15, were recently[76] found to obey a substantially different rule,* ΔE_C to $\Delta E_V \approx$ 60:40. For Si/Ge heterojunctions, the situation is still more uncertain. A simple electron affinity rule (known to have severe limitations),[79] suggests that virtually all of the band discontinuity should fall in the valence band and that the band alignment may be slightly of Type II (staggered), with the conduction band of Ge being above that of Si by 55 meV. The first observation of two-dimensional holes in selectively doped Ge_xSi_{1-x}/

* Even though the new rules have been confirmed by a variety of techniques, e.g., Reference 77 and references therein, there have also been conflicting reports.[78]

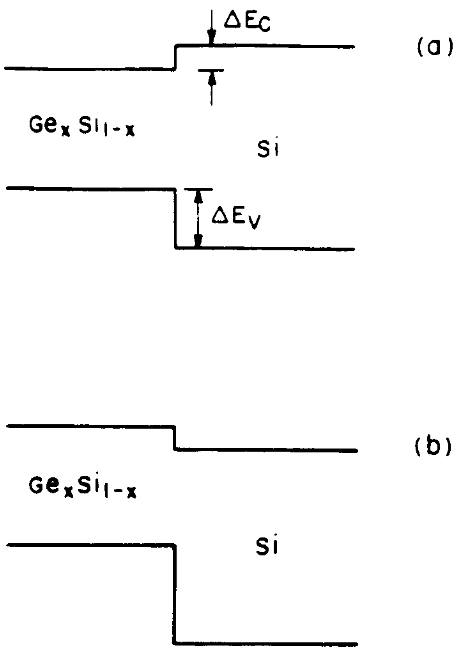

FIGURE 22. Possible heterojunction lineups in Ge/Si systems. (a) Type I and (b) Type II alignments.

Si heterostructures[80] indicated that indeed $\Delta E_V >> \Delta E_C$, but the alignment is of Type I (i.e., the narrower Ge gap within the wider Si gap). On the other hand, photoemission measurements of ΔE_V at the University of Wisconsin[81] implied that the discontinuities were split in the proportion ΔE_C to ΔE_V = 0.63:0.37, again with a Type I alignment (Figure 22a). It was a major surprise, therefore, when first observations of two-dimensional electrons[82,83] revealed a Type II alignment of bands with the Ge_xSi_{1-x} conduction band edge lying higher in energy than that in Si layers, as illustrated in Figure 22b.

This apparent paradox seems to have been resolved[83,84] by a careful consideration of strain in general $Ge_xSi_{1-x}/Ge_ySi_{1-y}$ systems ($x > y$). Depending on the composition parameters, the growth sequence, and possibly the growth conditions, the wider gap y layers may be nearly unstrained (maintaining the Si cubic symmetry), with most of the strain residing in the narrow-gap layers, or the y layers may also be strained. According to calculations of People and Bean,[84] the Type I band alignment results whenever y layers are cubic (unstrained), while in the case when both the x and y layers are strained, one may find a Type II alignment. Recently, Kasper et al.[85] introduced the concept of strain-symmetric growth in which a strained-layer superlattice is grown on a relaxed buffer layer of intermediate lattice constant (e.g., $Ge_{0.25}Si_{0.75}/Ge_{0.75}Si_{0.25}$ on a strain-relaxed $Ge_{0.5}Si_{0.5}$ layer pregrown on a silicon substrate. The electronic structure of an SLS grown in this way usually correpsonds to a Type II band alignment.

These findings are of particular importance for the implementation of enhanced-mobility devices in Si-based structures. The effect of mobility enhancement by modulation doping was first discovered by Störmer et al.[86] in AlGaAs/GaAs superlattices. In those modulation-doped structures, the low-field electron mobility parallel to the layers is greatly enhanced (especially at lower temperatures) because of the suppressed Coulomb scattering of electrons by ionized impurities — due to (1) spatial separation from the scatterers and (2) higher than thermal electron Fermi velocity in a degenerate two-dimensional electron gas, which reduces the scattering crosssection in accordance with the Rutherford formula.

One can expect similar effects in modulation-doped strained-layer Si/Ge systems. Ideally,

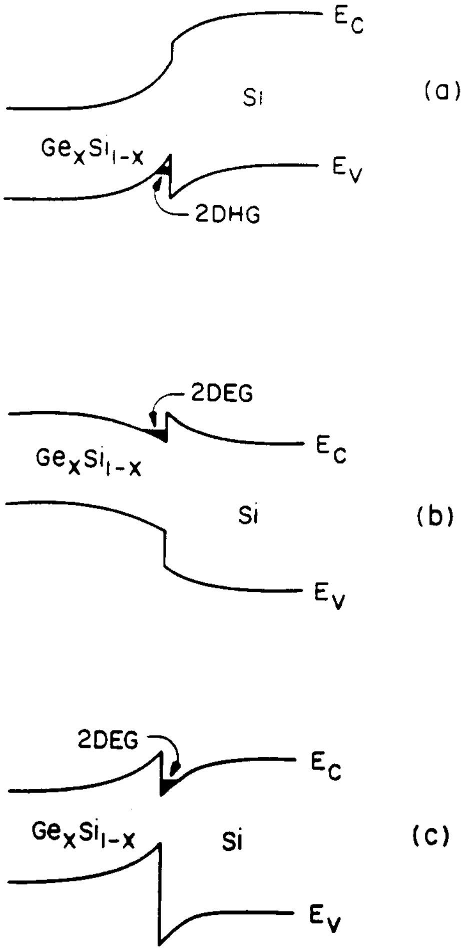

FIGURE 23. Possible enhanced-mobility two-dimensional carrier systems at heterointerfaces in strained-layer Ge_xSi_{1-x}/Si systems.

the enhanced mobility will be limited only by the phonon and the alloy scattering mechanisms. The latter mechanism, which is not important for two-dimensional electron system in GaAs, may become dominant in Ge_xSi_{1-x} alloys. The alloy-scattering-limited electron mobility in Ge_xSi_{1-x} was recently considered by Krishnamurthy et al.;[87] the effects of strain, however, were not included. According to the recent work, both experimental and theoretical,[80,82-85,88] three types of enhanced-mobility two-dimensional systems can be obtained (Figure 23):

1. Holes in the low-gap Ge_xSi_{1-x} alloy, separated from their parent acceptors in Si
2. Electrons in the low-gap alloy separated from their parent donors in Si in a Type I heterostructure
3. Electrons in silicon, separated from donors in a low-gap alloy in a staggered Type II heterostructure

There is hope, therefore, that Si-MBE may be able to duplicate some of the achievements of the MODFET technology[5,6] in III-V compound semiconductor systems. This goal is very attractive and certainly deserves extensive study. One has to understand clearly, however, that successes of the MODFET (also known as the high electron mobility transistor, or

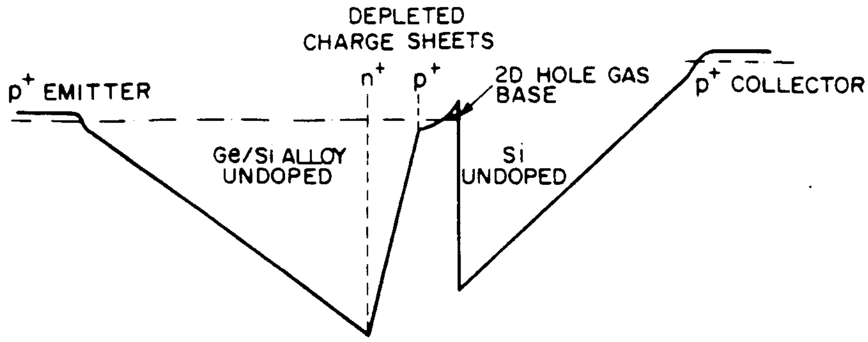

FIGURE 24. Valence-band diagram of a Ge/Si induced-base transistor. (From Luryi, S., *Physica*, 134B, 466, 1985. With permission.)

HEMT) are not really based on the enhanced mobility effect, since the latter takes place only at low electric fields (≤ 10 V/cm in AlGaAs/GaAs systems), while at fields of the order 200 V/cm, the electron drift velocity is already nearly saturated.[6] The speed advantage of HEMT results mainly from its higher saturation velocity of electrons and the lower source resistance. It remains to be investigated, both experimentally and theoretically, whether similar advantages can be obtained in silicon-based systems.

Recently, a transistor device was proposed[42] which could take a direct advantage of the high carrier mobility in a two-dimensional metal at an undoped heterostructure interface. This device, called the induced base transistor or IBT, belongs to the category of ballistic hot-electron transistors. A version of the IBT which could in principle be implemented using Si/Ge heterojunction technology[89] is illustrated in Figure 24. It contains an injecting emitter barrier of triangular shape produced by two built-in charge sheets — planar-doped donors and acceptors. The dopant concentration and the geometry must be designed so as to have both sheets depleted of mobile carriers. The two-dimensional hole gas induced at the undoped heterointerface (by the collector field and, possibly, by a selectively doped acceptor sheet within the collector barrier) should be separated from both acceptor sheets by undoped setback layers of about 50 Å — to ensure the benefit of enhanced mobility in the base. Note that the implementation of IBT with indirect-gap semiconductor heterojunctions requires the use of the ballistic injection of hot holes rather than electrons because, unlike the conduction-band minima, the valence-band maxima are located at the same $k = 0$ point in both semiconductors. The main idea of the IBT is to circumvent the basic trade-off (between the base conductivity and the transfer ratio, Section II.C.2) involved in the design of all hot-electron transistors — by using the enhanced carrier mobility at low lateral fields in the base. No experimental results on the IBT concept in silicon have been reported to date.

3. *Waveguide Detectors*

It would be very attractive to use a Ge$_x$Si$_{1-x}$/Si SLS for an IR photodetector — possibly with an avalanche multiplication in Si of the photogenerated signal. However, one has to find a way to circumvent the necessarily low absorption coefficient of a Ge$_x$Si$_{1-x}$ alloy at wavelengths of interest for fiber-optic communications. Even though, as discussed in the preceding section, the absorption edge is brought down by the strain to below the photon energy at wavelengths of silica-fiber transparency (λ = 1.3 to 1.5 μm), the Ge$_x$Si$_{1-x}$ alloy remains as indirect-gap semiconductor at all values of x. One can, therefore, expect a low absorption coefficient $\alpha \leq 10^2$ cm^{-1} even at photon energies above the fundamental threshold.

An efficient detector would, therefore, require an optical path length of order 100 μm or greater. This requirement appears to rule out the conventional detector designs in which photogenerated carriers travel along the direction of the propagation of light. Indeed, even if one made a 100-μm-thick SLS active layer, with p and n contacts at the top and the bottom, the response time of such a detector would be limited by the time of carrier drift across the SLS — which at best would be in a nanosecond range, i.e., too slow. In order to utilize the remarkable properties of a Ge_xSi_{1-x}/Si SLS, it is thus imperative to design a detector structure in which the direction of carrier propagation is normal to that of light. It is conceivable to use a thick planar SLS structure with laterally defined n and p contacts spaced 1 or several μm apart, so that carriers generated by incident radiation perpendicular to the surface will then drift laterally over a relatively short distance to the contacts. Such a design, however, appears impractical. A much better alternative is to use lateral propagation of light and vertical carrier drift. In that way, one can take advantage of the natural wave-guiding property of Si/Ge_xSi_{1-x}/Si heterostructures.

A novel IR photodetector structure of this type was recently proposed by Luryi et al.[90] It represents a waveguide in which the core is a strained-layer Ge_xSi_{1-x}/Si superlattice (SLS) sandwiched between Si layers of a lower refractive index. Absorption of IR radiation occurs in the core region due to interband electron transitions, and photogenerated carriers are collected in the Si cladding layers. The optimum SLS composition and thickness have been estimated[90] from the known material properties and waveguide theory. Experimentally, such structures were recently manufactured by MBE and tested.[91,92] The first SLS waveguide *pin* diodes[91] showed an internal quantum efficiency of 40% at $\lambda = 1.3$ μm and a frequency bandwidth of close to 1 GHz. The first APD structure[92] showed an avalanche gain as high as M = 50 and a quantum efficiency of 100% at M = 10. The waveguide-detector approach is entirely compatible with the Si integrated circuit technology and offers the possibility of fabricating a complete receiver system for long-wavelength fiber-optic communications on a silicon chip.

Below, we shall discuss, following Reference 90, the optimum composition of an SLS core, as determined by the trade-off between the confinement of radiation and the stability requirements for a Ge_xSi_{1-x}/Si SLS, as well as the design of a SAM APD waveguide structure, in which low-noise avalanche multiplication occurs in one of the Si cladding layers.

Consider first a waveguide-detector structure in which the core represents a single alloy layer, Figure 25a. We assume that in this layer $x > 0.5$ and that the absorption coefficient at wavelengths of interest is $\alpha \approx 10^2$ cm^{-1} (as indicated by the preliminary results[91] at $\lambda = 1.3$ μm, Figure 21). To be in the range of commensurate growth, the alloy thickness h must be less than the critical $h_c(0.5) = 100$ Å. To a good approximation, the fraction Γ of the integrated intensity of the light wave which falls within the absorbing core is given by[93]

$$\Gamma = 2\pi^2\left(\frac{h}{\lambda}\right)^2 (n_{\mathrm{core}}^2 - n_{\mathrm{clad}}^2) \qquad (41)$$

The refractive index of a Ge_xSi_{1-x} alloy is approximately given by a linear interpolation:

$$n(x) \approx n_{\mathrm{Si}} + x(n_{\mathrm{Ge}} - n_{\mathrm{Si}}) = 3.45 + 0.55\,x \qquad (42)$$

For $x = 0.5$, $h = 100$ Å, and $\lambda = 1.3$ μm, we thus find $\Gamma = 2.3 \times 10^{-3}$. The effective absorption coefficient of such a waveguide, $\alpha_{\mathrm{eff}} = \alpha\Gamma \sim 0.2$ cm^{-1}, is too low for a practical use (a detector would have to be several centimeters long and even the speed of light is not fast enough over such distances). The use of a superlattice is thus imperative. Consider the structure illustrated in Figure 25b. Ignoring in the first approximation the influence of strain on the dielectric constant, the refractive index of an SLS can be estimated as an average of $n^2(x)$ and n_{Si}^2 over one period:

$$n_{\mathrm{SLS}}^2 \approx \frac{h}{\mathrm{T}}\,n^2(x) + \frac{\mathrm{T} - h}{\mathrm{T}}\,n_{\mathrm{Si}}^2 = n_{\mathrm{Si}}^2 + r\,x(3.8 + 0.3x) \qquad (43)$$

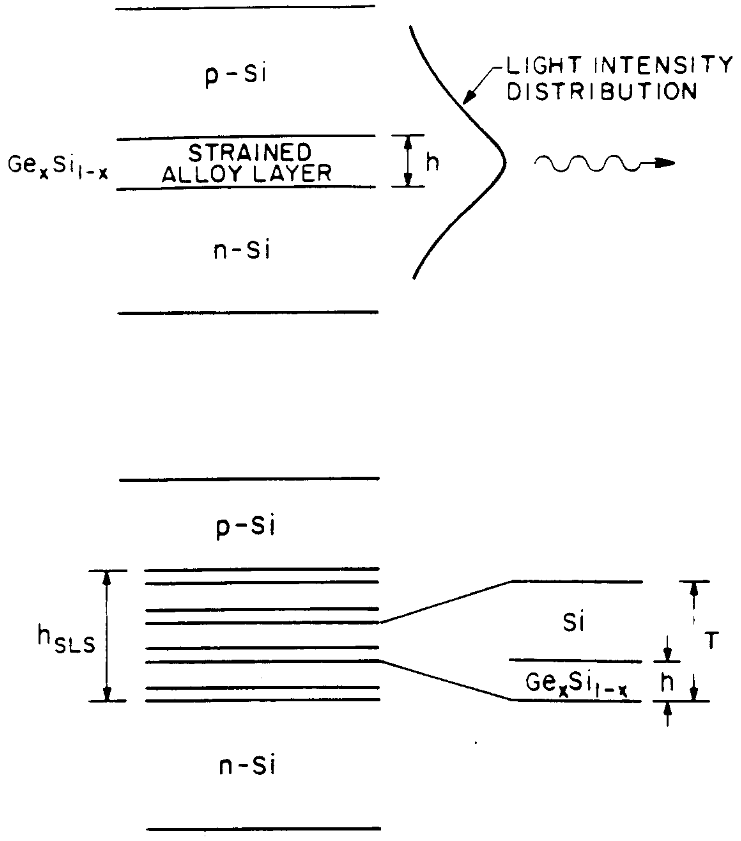

FIGURE 25. Schematic illustration of strained Ge_xSi_{1-x} layer waveguide detectors. (Top) Single-layer core; (bottom) SLS core.

The effective absorption coefficient of an SLS core is given by $\alpha_{eff} = r\Gamma\alpha$, and substituting Equation 43 into Equation 41, one has

$$\alpha_{eff} = 2\alpha(\pi/\lambda)^2 x(3.8 + 0.3x)[r\, h_{SLS}]^2 \tag{44}$$

In a practical device, the value of x will be determined by the need to absorb a specific λ, and hence for a fixed r one maximizes α_{eff} by pushing h_{SLS} to its limit given by Equation 40. As a function of r, therefore, α_{eff} is maximized together with the function $\phi \equiv zh_c(z)$, where $z = xr$. With the help of Equation 39, this function can be differentiated, yielding the result that ϕ is a monotonically decreasing function of z for all $z < 1$. One can reach the same conclusion from considering the experimental data,[12] which clearly show that $h_c(z)$ decreases faster than $1/z$. Thus, we arrive at the result that α_{eff} is maximized by smaller r, which for $h_{SLS} = h_{SLS}^{MAX}$ implies maximizing the superlattice width.

This result is obtained within the approximation Equation 41 for Γ. Of course, at very small r (and hence large h_{SLS}) the validity of Equation 41 is lost. This "narrow core" approximation is good provided $\Gamma \leq 1/2$, with the error at $\Gamma = 1/2$ being about 10%. In principle, Γ could be pushed above 1/2, which would invalidate the solution above. An inspection of the exact form of Γ for the fundamental mode in a symmetric three-layer slab dielectric waveguide[93] reveals, however, that Γ begins to saturate in this range, going over from the $\Gamma \propto h^2$ dependence at $\Gamma < 1/2$ to $\Gamma \propto h^p$ with gradually decreasing $p < 1$ for $\Gamma \gtrsim 1/2$. Therefore, α_{eff} has its optimum value for those r which correspond to $\Gamma \approx 1/2$. Physically, as the superlattice is made thicker to absorb the wings of the light intensity