17.3 Hot electron diodes and transistors

Device applications of hot electrons can be classified according to the type of hot electron ensemble involved. Most of the practical applications, beginning with the Gunn effect, are based on *quasi-thermal* rather than *ballistic* ensembles, for the simple reason that the former arise naturally − and can be easily maintained − in many situations when power input in the electronic system temporarily exceeds the energy relaxation rate. In contrast, ballistic ensembles are ‘‘capricious’’ and can be maintained only under special conditions.

Much of the pioneering research in hot-electron devices has been driven by the hope to develop new devices with high-frequency applications. Increasing carrier temperature often drastically changes the nature of carrier transport in an electronic device, and thus may lead to a negative differential resistance (NDR) and an unstable current-voltage characteristic. Both current driven (S-shape NDR) and voltage driven (N-shape NDR) instabilities are possible with quasi-thermal electron ensembles. As first shown by Ridley (1963) the former type of instability leads to current filamentation and the latter to formation of electric field domains (cf. e.g. Sze, 1981, Chap. 11). The simplest heterostructure device exhibiting an S-shape instability is the hot-electron diode proposed by Hess et al. (1986), illustrated in Fig. 4. Direction of carrier transport in this diode is ‘‘vertical’’ , i.e. normal to the plane of heterostructure layers. At low currents electrons are cold and the current across the barrier is by tunneling. At high currents the conduction mechanism is thermionic emission of hot electrons over the barrier. The high-current regime has a lower resistance. The Hess diode was studied experimentally by Emanuel et al. (1988) who demonstrated an S-shape NDR connecting two stable regimes. Transition between these two regimes is accompanied by an instability which can be used for the generation of high frequency oscillations. Spatio-temporal dynamics associated with this novel instability has been studied by Wacker and Schöll (1994).

Another type of hot-electron instability arises in heterostructures with parallel transport

along the layers. Here the hot-electron regime has the higher resistance and the instability is voltage-driven. At high voltages carriers become hot and redistribute themselves differently among parallel layers, leading to N-shape NDR. This effect, known under the name of real-space transfer (Hess et al. 1979), will be discussed in the next section.

17.3.1 Real-space transfer diodes. The idea of real-space transfer (RST) as a mechanism of NDR in layered heterostructures was first proposed by Gribnikov (1972). Subsequently, Hess et al. (1979) proposed a practical way for the implementation of RST diodes by using modulation doped multilayers. Shortly after this proposal, the RST effect was discovered experimentally by Keever et al. (1981). Microwave generation in the RST diode was first demonstrated by Coleman et al. (1982) in the same modulation-doped structure. For the history of the RST diode development the reader is referred to an excellent recent review by Gribnikov et al. (1995). The idea of RST bears a strong similarity to the Ridley-Watkins mechanism of the Gunn effect, as illustrated by the following simple model.

Consider a periodic multilayer heterostructure with narrow-gap layers of thickness d_1 and wide-gap layers of thickness d_2. Assume a field-independent mobility μ_i in each layer ($\mu_1 > \mu_2$) and effective electron masses m_1 and m_2, respectively. The total density of electrons per unit area is fixed by the overall neutrality: $n_1 + n_2 = n = \text{const}$. Assume further that layers are so thin that T_e is not a local temperature but pertains to the whole electron gas and hence when layers 1 and 2 exchange electrons, the mean energy is not being transferred from one layer to another.

In the steady state, characterized by T_e, the electron imref E_F must be constant in the direction perpendicular to the heterostructure layers; otherwise there would be a current flow and a redistribution of charge between the layers. Using non-degenerate bulk statistics, the sheet carrier density in each layer is given by

$$n_i/d_i = 2(2\pi m_i kT_e/\hbar^2)^{3/2} e^{-(E_{Ci}-E_{Fi})/kT_e} . \quad (8)$$

Taking $E_{F1} = E_{F2} = E_F$, and denoting $\Delta E \equiv E_{C2} - E_{C1}$, we find the ratio n_1/n_2 in the form:

$$z \equiv \frac{n_1}{n_2} = \left[\frac{m_1}{m_2}\right]^{3/2} \frac{d_1}{d_2} e^{\Delta E_C/kT_e} = z(T_e) . \quad (9)$$

Eq. (9) is analogous to the ratio of valley populations in the Ridley-Watkins-Hilsum model of the Gunn effect. From the energy balance equation (6a),

$$\frac{k(T_e - T)n}{\tau_\varepsilon} = eF^2(n_1\mu_1 + n_2\mu_2) , \quad (10)$$

we can express the electric field as a function of T_e:

$$F(T_e) = \left[\frac{k(T_e - T)}{e\tau_\varepsilon \mu(T_e)}\right]^{1/2} \quad (11)$$

where $\mu(T_e)$ is the average mobility of electrons in layers 1 and 2,

$$\mu(T_e) \equiv \frac{n_1\mu_1 + n_2\mu_2}{n_1 + n_2} = \frac{z\mu_1 + \mu_2}{z + 1} . \quad (12)$$

The current density per unit width of the diode is also a function of T_e:

$$J(T_e) = enu(T_e)F(T_e) \equiv en\bar{v}(T_e) . \quad (13)$$

Eliminating T_e from Eqs. (11) and (13) we find the velocity-field dependence $\bar{v}(F)$, Fig. 5. For deep NDR the factor $(m_1/m_2)^{3/2}(d_1/d_2)$ must be as small as possible (the density of states in the low-mobility layers must be high). Only then the increasing T_e would lead to a strong decrease in $\mu(T_e)$ and $\bar{v}$.

The above simple model does not take into account the electric polarization of the structure arising from electron redistribution between the layers; this polarization is of particular importance in modulation-doped structures, Fig. 6. If the device is used as an oscillator, electrons must cycle back and forth between the high and low mobility layers. The maximum oscillation frequency is limited by the delay due to ‘‘cold’’ electrons returning from the potential ‘‘pockets’’ in layers 2. This process occurs mainly by thermionic emission over the

potential barrier created by the space-charge of ionized donors. For a modulation-doped AlGaAs/GaAs heterostructure at room temperature the return time can be estimated to be at least 10^{-11} sec and may be substantially longer at lower temperatures. This suggests that some of the observations of multi-GHz microwave activity in modulation-doped structures at 77 K may in fact be due to Gunn oscillations, rather than RST.

Schöll and Aoki (1991) considered the nonlinear nature of RST dynamically coupled to space-charge polarization effects in modulation-doped heterostructures and found a surprisingly rich nonlinear dynamics. As the carrier density in the widegap layers increases due to RST, the net positive space charge in these layers decreases and the band bending in potential pockets flattens out. This leads to an increased backward thermionic emission and depletion of the potential pockets. The system may thus cycle back and forth with some delay due to the finite rate of dielectric relaxation of the electric field and of the space charge in the widegap layers. Schöll and Aoki (1991) predicted the existence of both periodic and chaotic self-oscillations at frequencies in the range of 20 to 100 GHz. Existence of the distinct time-scale separation between the fast RST and slow dielectric relaxation gives rise to rather complicated spatio-temporal evolution scenarios in this system, e.g., the formation of traveling field domains (Döttling and Schöll, 1994).

17.3.2 Multi-terminal real-space transfer devices. Transistor applications of the RST began with the proposal by Kastalsky and Luryi (1983) of a three-terminal structure where hot-electron injection occurs between *separately contacted* conducting layers. This structure, called (interchangeably) the charge injection transistor (CHINT) or negative resistance FET (NERFET), is illustrated in Fig. 7. One of the two layers (‘‘emitter’’) has source and drain contacts and plays the role of a hot-electron cathode. The other layer (‘‘collector’’) is separated by a potential barrier. When the emitter electrons are heated by the source-drain field most of them do not reach the drain but are injected over the barrier into the collector layer; a

strong NDR develops in the drain circuit. The transistor action results from an efficient control of the electron temperature T_e and hence the injection current I_C by the input voltage V_D.

Since the first demonstrations of three-terminal RST devices in GaAs/AlGaAs heterosystem (Kastalsky et al. 1984; Luryi et al. 1984a,b) considerable progress was made in the technology of device fabrication. Implementation of InGaAs/InAlAs heterostructures with epitaxial contacts enabled observation of deep NDR at room temperature (Mensz et al. 1990a,b). These devices also showed high frequency performance with the current and the power gain cut-offs of about 40 GHz (Mensz et al. 1990c). Static characteristics of InGaAs/InAlAs RST transistors are shown in Fig. 8. The quest for highest peak to valley ratios of the NDR in the drain circuit turned out rather unexpectedly, when an infinite ratio and then even *negative* valley values of I_D were observed. Explanation of this strange behavior was given by Luryi and Pinto (1991a) in terms of the formation of a hot-electron domain in the channel with its electrostatic potential exceeding that of the drain. Simulation by Pinto and Luryi (1991) of RST transistors with the help of continuation techniques revealed the existence of multiply-connected characteristics with many loops and folds − which qualitatively account for experimentally observed nonlinearities. These features arise due to an interplay between the fast RST process and relatively slower dielectric relaxation in the channel; one can draw a conceptual similarity to nonlinear instabilities in modulation doped RST diodes described by Schöll and Aoki (1991).

The use of emitter channels made of InGaAs where the satellite valley separation is high is one way to disentangle the RST from intervalley transfer effects that were shown by Kizillyalli and Hess (1989) to be very important in GaAs/AlGaAs structures. To achieve ultimate speed of RST transistors it is important to get rid of the parasitic momentum space transfer which slows down the device operation. The record speed in RST transistors (extrapolated current-gain cutoff of 115 GHz) was demonstrated by Belenky et al. (1994a) in InGaAs/InP heterostructures where the barrier height for RST ($\Delta E_C \approx 0.25$ eV) is nearly twice lower than intervalley

separation in the channel ($E_{L\Gamma} = 0.55\,\text{eV}$). These devices had a top-collector configuration, advantageous from the point of view of parasitic capacitances. Similarly processed top-collector InGaAs/InAlAs devices (where $\Delta E_C \approx 0.5\,\text{eV}$ is nearly as high as $E_{L\Gamma} = 0.55\,\text{eV}$) had substantially lower cutoff frequencies. Belenky et al. (1994a) attributed the difference to a parasitic effect of momentum space transfer slowing down the RST process.

Unwelcome momentum space transfer effects can also be eliminated by using p-type heterostructures (Favaro et al. 1990; Mensz et al. 1990d). However, heating of heavy holes is not as efficient as heating of Γ-valley electrons. In a given field **F** the steady-state effective carrier temperature T_e is proportional to carrier drift velocity <**v**> along the field (cf. Eq. 6a). Since the latter is higher for lower mass carriers in any given field,[5] the RST of holes requires higher heating fields. This penalty may not be excessive in silicon-based heterostructures, where one may contemplate advantageous integration of RST transistors with traditional VLSI circuits. Thus, Mastrapasqua et al. (1994, 1996) reported the implementation of powerful logic elements using RST of hot holes in strained-layer Si/GeSi heterostructures.

The lattice matched InGaAs/InP system is nearly ideal for the implementation of RST transistors. For many contemplated applications, however, it is attractive to stay within technologies based on GaAs substrates. To this end, several groups (Favaro et al. 1989; Hueschen et al. 1990; Maezawa and Mizutani, 1991) have reported RST transistors with the emitter channel implemented in strained InGaAs layers, see also references cited in the reviews by Luryi (1990a) and Gribnikov et al. (1995). Progress in the design of strained-layer InGaAs/GaAs RST transistors continues to this day (Lai and Lee, 1994, 1995; Wu et al. 1995a,b).

---

5. For lower fields, one is in the mobility regime $\mu = e\,\tau_m/m$; in the saturated velocity regime, one has $<v> \propto (\hbar\omega_{op}/m)^{1/2}$.

Maezawa and Mizutani (1991) studied the microwave performance of a 1μm-channel AlGaAs/InGaAs/GaAs heterostructure device − comparing its operation as an insulated-gate FET and as a RST transistor (in the latter mode, the gate plays the role of a collector). They demonstrated experimentally that the RST mode is faster (by nearly a factor of three) than the FET mode. This conclusion was supported by subsequent Monte Carlo simulations by Akeyoshi et al. (1992). This advantage results from the fact that the ultimate speed of a RST transistor is limited by the time of flight of hot electrons over distances of order the barrier thickness rather than the source-to-drain time of flight, characteristic of a FET. Nevertheless, at this time I do not see bright prospects for the use of RST transistors as high-speed amplifiers, since the time-of-flight advantage proves illusory when the device is scaled into deep submicron regime. Another reason for pessimism is an extremely sharp dependence of the small-signal parameters, such as the short-circuit current gain h_{21} and the unilateral gain U, on the electron heating bias V_D (Mensz et al. 1990c; Belenky et al. 1994a). Even though the extreme sensitivity of the gain to quasi-static DC bias may itself find useful applications, circuits taking advantage of this unique property are yet to be demonstrated.

A fundamental property of RST transistors is the *symmetry equivalence* (Luryi and Pinto, 1992) between the internal states $S[V_D, V_C]$ of the device at different external bias configurations:

$$S[V_D, V_C] \rightleftarrows S[-V_D, (V_C - V_D)]\,. \tag{14}$$

This correspondence follows from the reflection symmetry in the plane normal to the source-drain direction which cuts the channel in the middle, cf. Fig. 7. Thus the output current is invariant under an interchange of the input voltages VS and V_D and the device exhibits an exclusive-OR (**xor**) dependence of the collector current on the input voltages, regarded as binary logic signals.

Even more powerful logic functionality is obtained in a RST device with three input

terminals (Luryi and Pinto, 1991b). This device, which we shall refer to as the ORNAND

$$\textbf{ornand}(\{\mathbf{V_j}\}) \equiv (\mathbf{V}_1 \cup \mathbf{V}_2 \cup \mathbf{V}_3) \cap (\bar{\mathbf{V}}_1 \cup \bar{\mathbf{V}}_2 \cup \bar{\mathbf{V}}_3) \tag{15}$$

$$= \mathbf{V}_1 \cup \mathbf{V}_2 \qquad \text{for } \mathbf{V}_3 = 0 \qquad (\textbf{or})$$

$$= \bar{\mathbf{V}}_1 \cup \bar{\mathbf{V}}_2 = \overline{\mathbf{V}_1 \cap \mathbf{V}_2} \qquad \text{for } \mathbf{V}_3 = 1 \qquad (\textbf{nand})$$

where the symbols $\cap$, $\cup$, and $\bar{\mathbf{A}}$ stand for logic functions **and**, **or**, and **not A**, respectively. When $\mathbf{V}_3$ is low, $\textbf{ornand}(\{\mathbf{V_j}\}) = \textbf{or}(\mathbf{V}_1, \mathbf{V}_2)$, and when $\mathbf{V}_3$ is high, $\textbf{ornand}(\{\mathbf{V_j}\}) = \textbf{nand}(\mathbf{V}_1, \mathbf{V}_2)$. Operation of an ORNAND gate was first demonstrated by Luryi et al. (1990) using an assembly of three discrete RST transistors. Recently, Mastrapasqua et al. (1994, 1996) reported a monolithic ORNAND gate implemented in Si/SiGe heterostructure and operating at room temperature.

Figure 10 shows ORNAND gate implemented in a InGaAs/InAlAs heterostructure lattice matched to InP substrate (Mastrapasqua et al. 1993). Besides the logic function (15) in the collector current $I_C(\{\mathbf{V_j}\})$, Fig. 10a shows a similar function in the output light. The idea of using RST for the implementation of light emitting functional devices (light emitting ‘‘triodes’’ and multiterminal lasers) was first proposed by Luryi (1991) and demonstrated by Mastrapasqua et al. (1992). Operation of such devices is based on RST of *minority carriers* into a collector layer of complementary conductivity type, cf. the structure cross-section in Fig. 10b. This structure has been designed to be an efficient light emitter. Optical logic operation in InGaAs/InAlAs heterostructures at room temperatures proved to be even better than electrical operation (in terms of the on/off ratio of output in different logic states) because light output is generated only by the RST flux of hot electrons and is not sensitive to parasitic leakage of majority holes. Operation of light emitting RST logic devices of the kind illustrated in Fig. 10 has been reviewed by Luryi and Mastrapasqua (1993). Potentially more promising (but also more challenging technologically) are light-emitting RST transistors in the top-collector configuration. Parallel investigation of such structures with InAlAs and InP barriers

(with both emitter channel and collector implemented in InGaAs) enabled Belenky et al (1993) to clarify the physics issues associated with the radiative efficiency of RST and hot-carrier effects within the collector layer. Recently, Lai et al. (1996) reported light emitting RST devices implemented on GaAs substrates and utilizing strained-layer $In_{0.2}Ga_{0.8}As$ emitter channels. I believe the most attractive potential applications of RST injection devices will be found in the area of optoelectronics.

General purpose logic operation based on the ORNAND gate is not likely in view of its relatively large power dissipation (even in the *off* state there is a substantial current − flowing to the drain). In this context, an interesting innovation was recently introduced by Koscica and Zhao (1995) who integrated an RST transistor in series with an FET. In principle, their ''FERST'' structure allows to eliminate power dissipation in the *off* state. Tian et al. (1992, 1993) proposed another class of logic RST devices and studied their operation by Monte Carlo simulations. The key novel idea is to employ *two* output electrodes C_1 and C_2 (which is quite possible in the top-collector configuration). Depending on the heating field, the RST will flow predominantly either in C_1 or in C_2. The concept od a fast routing function based on RST is very appealing. In a sense, the original CHINT device is doing just that (see Fig. 8) but its routing function can be hardly used, since the output current is switched between an input and output electrodes rather than between two output electrodes. On the other hand, the FERST device of Koscica and Zhao (1995) is free of this problem, as it performs the routing function controlled by an insulated gate.

17.3.3 Ballistic transistors. Ballistic-injection hot-electron transistors discussed in this section, belong to the group of potential-effect transistors (PET) which also includes all bipolar and analog transistors. In these devices transistor action results from *modulating the height of a potential barrier for carrier injection.* The barrier is controlled electrostatically by the charge supplied to the base by an input electrode. Different types of potential-effect transistors

employ different ways of supplying charge to the base.

The most important property of the base is that it must be ‘‘transparent’’ for carriers injected from the emitter. It is a notable feature of all PETs that *carriers participating in the output current are distinct from those that control the barrier height*. Any mixing between these groups of carriers leads to a degradation of the transistor performance. The classic example of a PET is the bipolar transistor, where the identity of the two groups is determined by the band in which the carriers move. Transparency of the base for the minority carriers relies on the fact that their lifetime is substantially longer than the time they need to travel across the base and form the output current. In a vacuum triode, which is the prototype of all PET’s, the two groups of carriers move in different regions of the ‘‘real space’’; similar situation occurs in most analog transistors.

In ballistic hot-electron transistors, each group of carriers is maintained in a distinct narrow momentum range within the same band. One group of carriers is injected from the emitter at high energy and travels across the base ballistically; the other group of carriers are ‘‘native’’ to the base and are distributed there in an equilibrium fashion. As in a bipolar transistor, one can say that the controlling and the injected carriers move in different portions of the momentum space. Base transparency for ballistic carriers relies on the short time required to cross the narrow base compared to the momentum relaxation time τ_m, which is the time required for the two groups of carriers in the base to lose their identity.[6]

Ballistic-injection transistors differ by the materials employed and by the physical mechanism of hot-electron injection into the base. The first proposal of a hot-electron injection device by Mead (1960) was based on electron tunneling from a metal emitter through a thin oxide barrier into a high energy state in a metal base, Fig. 11a. Another insulating barrier separated the base from a metal collector electrode and the whole structure was called the MOMOM (metal-oxide-metal-oxide-metal) transistor. Subsequent versions of this device had

the second MOM replaced by a metal-semiconductor junction, resulting in a transistor structure called the MOMS (Fig. 11b). Attempts have also been made to employ a vacuum collector barrier (MOMVM). Metal-base transistors, which employ thermionic rather then tunneling injection of hot carriers into the base, were first proposed in the form of a semiconductor-metal-semiconductor (SMS) structure. Schematic band diagram of an SMS transistor is illustrated in Fig. 11c.

The basic principles of all ballistic hot-electron transistors can be discussed using the SMS as an example. The SMS is a direct unipolar analog of the bipolar junction transistor. The emitter-base and the base-collector junctions are Schottky diodes biased, respectively, in the forward and the reverse direction. As $V_{\rm BE}$ increases, the emitter current rises exponentially. The hope is that electrons injected into the metal base traverse it without losing too much energy, so that at the plane of the base-collector junction the injected electron energy is sufficiently high to clear the metal-semiconductor barrier. Following the bipolar terminology, the fraction of injected electrons that make it to the collector (the transfer ratio) is referred to as the common-base current gain α. If α is close to unity, then the transistor power gain is approximately equal to the ratio of the large output resistance to the small input resistance, which becomes progressively smaller as the emitter-base Schottky diode is forward biased.

The tunnel-emitter (MOMOM, MOMS) structures operate somewhat differently. The emitter current is not and does not have to be a strong function of the input bias. What is being controlled is the electron injection energy relative to the top of the collector barrier. Depending on that energy, the injected electrons end up mostly in the base or in the collector. In principle, these transistors are devices with a controlled transfer ratio $\alpha -$ switching a relatively

6. Strictly speaking, one can argue that distinction between the two groups persists for a longer − energy relaxation − time τ_ε. However, I am not aware of any unipolar ballistic transistor design that would take advantage of this opportunity. A potential effect transistor with ‘‘robust’’ transport properties would be very attractive but such a device could hardly employ a ballistic electron ensemble.

constant emitter current between the base and the collector circuits.

The problem that has plagued the SMS (and all other metal-base) transistors is their poor transfer ratio α. 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-collector interface. In most cases, transmission between electronic states in materials with different band structure is suppressed kinematically.[7] Exact calculations by Stiles and Hamann (1989) for the case of electron transmission through $NiSi_2$-Si interfaces indicate that over 50% of electrons are reflected even in the most ideal situation. Metal-semiconductor interfaces in practice conduct due to imperfections, finite size of polycrystalline boundaries, phonons and other effects (such as diffraction) that permit violating the conservation of the crystal momentum parallel to the interface. These mechanisms do not seriously affect the operation of Schottky diodes (except by lowering the effective value of the Richardson constant A^* to A^{**}, cf., e.g., Sze, 1981) but they are prohibitive for the operation of metal-base transistors, where α must be near unity.

Nevertheless, there have been experimental reports of a transistor action in monocrystalline Si/$CoSi_2$/Si structures with α as high as 0.6. These reports caused some stir in the device community since one could not rule out some ''accidental'' resonance that aids the QM transmission of hot electrons in these devices, even though such an interpretation is not very likely. 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. Careful analysis by Tung et al. 1986) of the correlation between the pinhole sizes and the device characteristics

7. It is instructive to consider an idealized interface between Si and Ge, neglecting strain and assuming that conduction band edges coincide in both materials. In this case, conservation of the crystal momentum $\mathbf{k}_\parallel \equiv (k_x, k_y)$ parallel to the interface requires that both the initial and final states of the electronic motion must lie in the Brillouin zones of both materials on the same perpendicular to the interface plane, (k_x, k_y). If the initial electron state is near the bottom of the conduction band, it will not have an allowed counterpart in the other material since the ellipsoids of Si and Ge do not project onto one another. Consequently, the coherent transmission probability will vanish (Luryi, 1990b).

revealed no evidence for a hot-electron component of the current through the base. On the other hand, the pinhole conduction in some cases gives an α as high as 0.95. The SMS structure, therefore, works like a permeable base transistor (Lindmayer, 1964) − which is not a hot-electron device. Thermionic emission through a permeable base has, in our opinion, a greater potential for use in transistors than the hot-electron transport through a metal base.

References to the history of metal-based hot-electron transistor concepts can be found in my review (Luryi, 1990b). There has been little development of these concepts in recent years.

The interest in ballistic hot-electron transistors was revived by the tremendous progress in the epitaxial growth of semiconductor heterojunctions. An influential paper by Heiblum (1981) offered eloquent support to the concept of all-semiconductor analogs of MBT. A number of such devices have been implemented (cf. the above cited review by Luryi, 1990b). Figure 12 shows their schematic energy-band diagrams. The problem of QM reflections can be largely avoided in such structures − provided the carrier transport occurs in similar Brillouin-zone points on both sides of the interface. In the design of all hot-electron transistors with a doped base one faces a trade-off between the degradation of α due to scattering (gets worse in thicker layers) and the increasing base resistance for thinner layers. This trade-off leaves nearly no room for the implementation of a high-speed ballistic transistor.

In an attempt to circumvent this problem, I had proposed an *induced-base transistor* (Luryi, 1985). This device, illustrated in Fig. 12c, can be regarded as a metal-base transistor − with the notable difference that the base ‘‘metal’’ is two-dimensional (induced by the collector field at an undoped heterointerface). The induced-base conductivity is virtually independent of its thickness down to $d \lesssim 100\,\text{Å}$. At such short distances the loss of hot electrons due to scattering is small. Injected hot electrons, traveling across the base with a ballistic velocity of order 10^8 cm⁄sec, lose their energy mainly through the emission of polar optic phonons. For $d = 100$ Å the attendant decrease in α is estimated to be about 1%. Energy losses to the

collective and single-electron excitations of the 2D electron gas are negligible. The first experimental implementation of the induced-base transistor (Chang et al. 1986) showed a common-base current gain of $\alpha \approx 0.96$ at room temperature. This value of gain approaches the estimated limit due to both the QM reflection at GaAs/AlGaAs interface and scattering in a 100 Å-thick GaAs base.

The induced base transistor can be viewed as the "ultimate" unipolar ballistic device. Nevertheless, even leaving aside the difficulties of fabrication, its practical use does not seem very likely. As stated in my original paper (Luryi, 1985), the speed limitation of an induced base transistor are quite similar to those of heterojunction bipolar transistors (HBT) − devices that are manufactured by an incomparably more mature technology and at the same time enjoy much more robust and reproducible characteristics. The idea of a HBT with a wide-gap emitter was first proposed by Shockley (1951) and developed theoretically by a number of workers, most notably by Kroemer (1957, 1982). The advantage of the wide-gap emitter concept is that the minority-carrier injection into the emitter can be practically suppressed. This allows the use of heavily doped base layers without degrading the gain.

Great progress has been achieved in the implementation of the hot-electron HBT − a device which employs ‘‘non-equilibrium’’ ballistic transport in the base.[8] Hot-electron HBT is the first transistor to reach subpicosecond speeds (Chen et al. 1989). In this device, illustrated in Fig. 13, carriers are launched at high energies into a small angular cone perpendicular to the base-emitter junction. The fact that the velocity distribution is sharply peaked in the direction perpendicular to the base layer, is beneficial for downscaling the device area (Jalali et al. 1989).

8. The words ‘‘non-equilibrium’’ are in quotes to stress the fact that any minority-carrier transport is non-equilibrium. Minority carriers are always ‘‘hot’’ with respect to the possibility of recombination. However, because the recombination times are typically much longer than the kinetic energy relaxation time, one can talk about a quasi-equilibrium energy distribution of minority carriers − and deviations from that equilibrium of minority-carrier packets injected at high energies and traveling ballistically across the base.

In bipolar transistors the lateral scaling is limited mostly by recombination in the extrinsic base region (both in the base bulk and on the surface adjacent to the emitter stripe). The high injection velocity in HBT − compared to the rate of lateral spread by diffusion − results in a better spatial confinement of injected carriers to the intrinsic base area.

Ballistic hot-electron ensembles are difficult to maintain because of their extreme sensitivity to scattering. At the same time their supposed speed advantage is largely illusory, because all that can be minimized in these devices is the base transit time. This time is no longer limiting the performance in modern HBT. Even with diffusive transport across a thin base ($W_B \lesssim 0.1\,\mu m$) one has typically $\tau_B \lesssim 1\,ps$. Because of this the entire concept of a unipolar ballistic transistor is no longer practical, in my view.

In a sense, the success of *bipolar* hot-electron transistors is largely due to the fact that their operation does not entirely rely on ballistic transport. Indeed, it is even hard to conclusively distinguish the diffusive and ballistic regimes in narrow-base HBT's. Thus, Ritter et al. (1991) and Levi et al. (1992) attempted to distinguish these regimes by studying the base thickness dependence of static gain β − in a series of structures with W_B varying − and came to different conclusions. In principle, it is possible to distinguish the dominant transport mechanism by studying high-frequency behavior of a single transistor (Grinberg and Luryi, 1992). The required microwave experiment is difficult and has never been performed.

17.3.4 Coherent transistor. In conventional heterostructure bipolar transistors the prevalence of ballistic transport is not crucial for the device operation. In contrast, unipolar ballistic transistors require collisionless transport essentially (as discussed above, this is their Achilles' heel). Reliance on subtle and hard-to-maintain effects may be tolerated ultimately only if accompanied by a dramatic leap in the performance. In this spirit, consider the coherent transistor, a device recently proposed by Grinberg and Luryi (1993).

In general, the base transport factor $\alpha_B(\omega) = |\alpha_B| \exp(-i\omega\tau_B)$ with increasing frequency ω describes an inbound spiral $\alpha_B \to 0$ in the complex plane. We shall apply the term coherent to any transport mechanism that provides a sufficiently slow spiraling in of the $\alpha_B(\omega)$, such that $|\alpha_B| \gtrsim 0.5$ for large values of $\omega\tau_B \gtrsim \pi$. The quantity $\phi_B \equiv \omega\tau_B$ is called the base transit angle.

In a "perfectly" coherent transistor all injected electrons travel toward the collector with the same velocity v_B. In this case, a periodic modulation of the injection at the emitter interface sets up in the base an electron density wave of the form:

$$n(z,t) = n_0 \, e^{i\omega(t - z/v_B)}, \tag{16}$$

where n_0 is the amplitude of the concentration modulation. At any point in the base, the electron current density $J(z,t)$ equals $-env_B$ and the complex transport factor α_B hence is

$$\alpha_B \equiv \frac{J(W_B, t)}{J(0, t)} = e^{-i\omega\tau_B}, \tag{17}$$

where $\tau_B \equiv W_B/v_B$. In an intrinsic transistor, the common-emitter current gain h_{21}^e is given by $\alpha_B/(1-\alpha_B)$ and equals

$$h_{21}^e = \frac{\exp(-i\omega\tau_B/2)}{2i\sin(\omega\tau_B/2)}; \tag{18}$$

$$\beta \equiv |h_{21}^e| = \frac{1}{2|\sin(\omega\tau_B/2)|} \underset{\omega\tau_B<1}{\approx} \frac{1}{\omega\tau_B}. \tag{19}$$

The usual procedure for the determination of f_T in a microwave transistor is to extrapolate to unity gain the data measured at relatively low frequencies. Since for $\omega\tau_B \lesssim 1$ the current gain rolls off very accurately as ω^{-1} (i.e., 10 dB/decade), as seen from Eq. (19), the extrapolation procedure will produce the usual value $f_T = 1/2\pi\tau_B$. However, in a coherent transistor f_T is not the fundamental current gain cutoff. Equation (19) indicates the existence of frequency windows, centered at $f_\nu = 2\pi\nu f_T$ and characterized by $\beta > 1$.

The physical origin of this effect is simple. Neglecting recombination in the base, the only reason that the base current is flowing in a microwave transistor is to maintain neutrality of the base layer by screening the injected charge. The resonant frequencies f_ν correspond to an

integer number ν of periods $\lambda = v_B/f$ of the density wave (16) in the base. To the extent that the wave is not decaying in amplitude in a perfect coherent transistor, the total minority-carrier charge in resonance does not fluctuate at all.

Collisionless base propagation by itself does not imply coherence. In the ballistic case the modulated signal injected at the base-emitter interface is washed out because of the thermal spread in the normal velocities of injected carriers, leading to a variance $\delta\tau$ in their base propagation time. The latter process is analogous to the Landau damping of density waves in collisionless plasmas and has an effect similar to diffusion. The coherent regime in a ballistic transistor arises when $\delta\tau \ll \tau_B$, i.e., when the injected electrons form a collimated and monoenergetic beam. A good approximation to such a beam results from the passage of electrons across an abrupt heterointerface at low temperatures. For $\delta\tau \ll \tau_B$, the minority-carrier density wave does not appreciably decay over the entire base and one finds

$$\alpha_B = e^{-(\omega\delta\tau)^2/2}\, e^{-i\omega\tau_B} . \qquad (20)$$

Figure 14a shows the calculated frequency variation of β in ballistic HBT with a finite $\delta\tau$ as it arises in a thermal ensemble. The transistor speed is limited not by the base propagation time τ_B but rather by its variance $\delta\tau$. As discussed in detail by Grinberg and Luryi (1993), a coherent transistor can have both the common-emitter current gain h_{21} and the unilateral power gain U exceeding unity at frequencies far above the usual f_T, cf. Fig. 14b. Calculations carried out for exemplary heterostructures, implemented at the state-of-the-art rules of technology, indicate that active transistor behavior can be extended to about 1 THz.

It is clear, however, that the ballistic coherent operation requires cryogenic temperatures. For a collisionless transport to hold over the entire base width, the electron kinetic energy Δ (the injection energy, corresponding to a conduction-band discontinuity in the base-emitter junction) should not exceed the optical phonon emission threshold, $\Delta \lesssim \hbar\omega_{opt}$. On the other hand, to achieve coherence, the injection energy Δ must substantially exceed the thermal

spread, whence we need $\hbar\omega_{opt} \gtrsim \Delta \gg kT$. This means that the implementation of base transport coherence in a ballistic HBT requires at least liquid N_2 temperatures. Moreover, the ballistic coherent operation is limited to ultra-high frequencies. The concept cannot even be tested at lower frequencies, because the first resonance in U appears only at $f_\pi \approx \pi f_T = 1/2\tau_B$, which must evidently be higher than the collision rate $1/\tau_m$ governing the momentum relaxation of ballistic electrons at energy Δ.

Another idea for achieving base transport coherence (which is not limited to ultra-high frequencies and also works at room temperature) can be traced back to the famous paper by Shockley (1954). In that paper he introduced the concept of transit-time diodes and suggested that the delay in minority-carrier transit across a transistor base can lead to an active device at extended frequencies. A necessary condition for this to occur is that the directed transport across the base be much faster than the diffusive transport, that tends to wash out a modulated structure of the injected distribution. Shockley (1954) suggested that this condition can be met in a minority-carrier delay diode with a variable doping in the base. However, because of a limited range of the potential variation available with an exponentially graded doping, the feasibility of this approach is marginal and it has never been realized. Recently, Shockley's argument was reconsidered (Luryi et al. 1993) in the context of HBT with a graded alloy base composition. It was shown that coherence of the base transport in such devices is feasible and may lead to useful applications. Figure 15 shows the calculated room-temperature microwave characteristics of a graded-base HBT structure designed by A. Kastalsky and myself (1995, unpublished) for the maximum power gain at 94 GHz. Reduction in the parasitic capacitance C_{Cx} to $C_{Cx} \lesssim 0.5\,C_C$ enables a still more aggressive design of the coherent transistor, optimized for stable oscillation at near $300\,\text{GHz}$ (Luryi, 1994, 1996). An attractive application for ultrafast transistor oscillators would be the implementation of submillimeter phased antenna arrays on a silicon chip.