17. Hot Electrons in Semiconductor Devices

Serge Luryi, State University of New York, Stony Brook, NY, USA

17.1 What is a hot electron device ?

Hot-electron phenomena have become important for the understanding of all modern semiconductor devices. In many cases these phenomena are a nuisance that must be somehow cauterized, like for example is the hot carrier injection into the gate dielectric in silicon field-effect transistors. This unwelcome phenomenon gives rise to a degradation of transistor characteristics and may lead to circuit failure. Another well-known hot-electron effect in devices (with more benign consequences) is the saturation of drift velocity in high electric fields. Velocity saturation is ubiquitous in modern semiconductor devices and is sometimes of central importance for their operation. Thus, it is essential for the operation of all transit-time diode oscillators (Sze, 1990).

Commercial utilization of hot-electron phenomena began with the Gunn effect (Gunn, 1963; Kroemer, 1964), based on the intervalley transfer mechanism for a negative differential resistance, proposed by Ridley and Watkins (1961) and Hilsum (1962). The Gunn diode is undoubtedly the best-known hot-electron device, for which a mature technology has developed, see, e.g., Bosch and Engelmann (1975) and Shur (1987, 1990). Another successful application of a hot-carrier effect has been made in nonvolatile memory devices. The floating-gate avalanche injection memory device FAMOS invented by Frohman-Bentchkowsky (1971) bears conceptual similarity to some of the real-space-transfer devices discussed below. The FAMOS represents a p-channel MOSFET structure with a floating gate electrode. In the process of ‘‘writing’’ the memory, carriers, heated by the drain field, avalanche near the drain junction with hot electrons from the avalanche plasma injected into the floating gate. As the gate is charged, its potential is lowered and the p-channel conductance increases. Pioneering papers in the development of FAMOS technology have been collected by Hu (1991).

This chapter is concerned with a special narrow class of semiconductor devices whose very principle is based on hot-electron effects. We shall not attempt to review the important mature hot-electron technologies mentioned above. Instead, we shall discuss rather exotic devices none of which have been used in practical electronic applications. Most of the recent research in this area has concentrated on demonstrating the existence of an effect in question, proposals of new structures and effects, and studies of their potential physical limitations. It should be remembered, of course, that the main purpose of this type of work is to come up with a device, which will find a practical application. Our review will comment on the viability of these ideas, their promise and shortcomings, as well as attempt to identify the key hurdles to their practical implementation.

17.2 Nonequilibrium electron ensembles

The term ''hot electrons'' purports a non-equilibrium ensemble of high-energy carriers. It is often possible to pump external energy (e.g., by shining light or applying an electric field) directly into the system of carriers. If the power input into the electronic system exceeds the rate of energy loss by that system to the lattice, then the carriers ''heat up'' and their velocity distribution $f(\mathbf{v})$ deviates significantly from the equilibrium Maxwellian form.[1] In general, the time-dependent distribution function $f(t, \mathbf{r}, \mathbf{v})$ can be determined by solving the Boltzmann transport equation,

$$\frac{\partial f}{\partial t} + \mathbf{v}\cdot\frac{\partial f}{\partial \mathbf{r}} + \mathbf{a}\cdot\frac{\partial f}{\partial \mathbf{v}} = \left[\frac{\partial f}{\partial t}\right]_{\text{coll}}, \tag{1}$$

where $\mathbf{a}$ is the acceleration and the collision integral in the right-hand side is a linear functional on f. In a steady state, $f(\mathbf{r}, \mathbf{v})$ does not explicitly depend on time but it may still be a function of the spatial position.

1. Note that one can also *cool* the carrier system by making it do work against an external field at a fast rate compared to power replenished by the lattice.

Solution of Eq. (1) is a complicated task even for the simplest scattering models involving scattering of electrons among themselves as well as with phonons and impurities. It is expected from a reasonable model of interaction between the electronic system and the thermal bath that collisions will *restore* the equilibrium distribution function $f_{\rm eq}(\mathbf{r},\mathbf{v})$ from any initial distribution − although this property is often difficult to prove mathematically. Since we know that thermodynamic equilibrium cannot be destroyed by scattering of electrons by a thermal bath, the collision integral must vanish if $f_{\rm eq}(\mathbf{r},\mathbf{v})$ is substituted for f. These properties are manifestly expressed in the model which approximates the collision integral by

$$\left[\frac{\partial f}{\partial t}\right]_{\rm coll} = \frac{f(t,\mathbf{v}) - f_{\rm eq}(\mathbf{v})}{\tau(\mathbf{v})} \quad , \tag{2}$$

where for brevity we have omitted the possible position dependence. The characteristic time constant $\tau(\mathbf{v})$ is called the relaxation time and the whole model (2) is called the relaxation time approximation. In this approximation the perturbed distribution will exponentially relax to equilibrium when the perturbing influence is removed.

The relaxation time approximation (2) is too crude in practice, because different characteristics of the distribution relax with different rates. Thus when collisions are predominantly elastic, it is natural that the first moment $\langle\mathbf{v}\rangle$ of $f(\mathbf{v})$ relaxes rapidly while the second moment $\langle\mathbf{v}^2\rangle$ takes a long time to relax. The first moment characterizes the electron drift velocity (or average crystal momentum) and the corresponding time constant, $\tau_{\rm m}$, is called the momentum relaxation time, while the longer time τ_ε is called the energy relaxation time. It is often a good approximation to characterize the electron ensemble by a model distribution function, which embodies the relevant relaxation kinetics. The choice of an appropriate distribution function may also depend on time t if the electric fields rapidly vary (on the scale of τ_ε or $\tau_{\rm m}$). Similarly, the choice of a model $f(\mathbf{v})$ may depend on the position $\mathbf{r}$ in the device. Consider hot-electron models, commonly occurring in devices.

17.2.1 Quasi-thermal. An effective temperature T_e is always established in an electronic ensemble of sufficiently high concentration, when the electron-electron (ee) interaction dominates over both phonon and impurity scattering. Starting from any initial distribution, ee collisions lead to an equilibrium within the electron gas. Since the center-of-mass velocity of colliding electrons does not change, the drift velocity $\langle \mathbf{v} \rangle$ of the electron gas remains a constant of the motion.[2] The distribution function is then of the form

$$f(\mathbf{v}) = \left\{ 1 + \exp[(m\mathbf{v}^2/2 - m\langle \mathbf{v} \rangle \cdot \mathbf{v} - E_F)/kT_e] \right\}^{-1} , \tag{3}$$

called the displaced Fermi distribution. It is characterized by the effective temperature T_e, the drift velocity $\langle \mathbf{v} \rangle$, and the Fermi level E_F, determined respectively by the conservation of energy, momentum, and the number of particles. In the reference frame that travels with the velocity $\langle \mathbf{v} \rangle$, the distribution (3) looks like an ordinary Fermi function. For a non-degenerate gas, Eq. (3) reduces to the form of a displaced Maxwellian distribution,

$$f(\mathbf{v}) = e^{-\frac{m\mathbf{v}^2/2 - m\langle \mathbf{v} \rangle \cdot \mathbf{v}}{kT_e}} . \tag{3'}$$

In the nondegenerate limit $T_e \gg E_F$, the effective temperature T_e is related to the average electron energy $\langle E \rangle$ by the well-known formula,

$$\langle E \rangle = \frac{3}{2} kT_e + \frac{1}{2} m \langle \mathbf{v} \rangle^2 . \tag{4}$$

In the opposite limit, $E_F \gg kT_e$, the average energy does not depend on T_e and is determined only by the carrier density n:

$$\langle E \rangle = \frac{3}{5} E_F + \frac{1}{2} m \langle \mathbf{v} \rangle^2 , \quad \text{where } E_F = (3\pi^2)^{2/3} \frac{\hbar^2}{2m} n^{2/3} . \tag{5}$$

In the intermediate range, the expression of $\langle E \rangle$ in terms of parameters T_e and E_F can be written in the form of a quadrature (see, e.g., Landau and Lifshitz, 1980).

2. This is true, so long as Umklapp processes are insignificant. When Umklapp is included, momentum relaxation occurs even under the pure ee interaction.

Any function of the form (3) will make vanish the collision integral that contains only ee interaction. If interaction with impurities is included, the collision integral will vanish only provided $\langle \mathbf{v} \rangle = 0$. Including interaction with phonons, the collision integral will vanish only if T_e coincides with the lattice temperature T.

In the situation when energy is continuously pumped into the electron gas by an electric field or electromagnetic radiation, one cannot ignore other forms of scattering, however slow they may be compared to ee interaction. Indeed, otherwise there would be nothing to check a runaway of the average energy and/or drift velocity. The correct approach is to use the so-called *adiabatic* approximation in which the distribution function is assumed in the form (3) with the effective parameters assumed to follow the variations in the external input, as governed by the balance equations for energy and momentum,

$$\frac{d\langle E \rangle}{dt} = \langle \mathcal{P} \rangle - \frac{\langle E \rangle}{\tau_\varepsilon} \quad , \tag{6a}$$

$$\frac{d\langle \mathbf{v} \rangle}{dt} = \langle \mathbf{a} \rangle - \frac{\langle \mathbf{v} \rangle}{\tau_m} \quad , \tag{6b}$$

where $\langle \mathcal{P} \rangle$ and $\langle \mathbf{a} \rangle$ are, respectively, the average power input and acceleration. For the electronic ensemble (3) moving under a force $\mathbf{F}$, one has approximately $\langle \mathbf{a} \rangle = \mathbf{F}/m$ and $\langle \mathcal{P} \rangle = \mathbf{F} \cdot \langle \mathbf{v} \rangle$. Parameters τ_ε and τ_m in these equations are determined self-consistently from the Boltzmann equation − by substituting into the collision integral for slow processes the distribution function in the form (3) − which is assumed maintained by the rapid ee interaction.

Thus, in the adiabatic approximation parameters τ_ε and τ_m of the balance equations (6) themselves become functions of the effective temperature T_e. The momentum relaxation time is usually much shorter than the energy relaxation time, $\tau_m \ll \tau_\varepsilon$. Indeed, elastic collisions are dominant for momentum relaxation, whereas an effective energy relaxation requires several inelastic interactions with phonons. Therefore, on the time scale of the momentum relaxation, the average electron energy can be considered quasi-static and while $\tau_m = \tau_m(T_e)$ in the balance equation (6b), the effective temperature T_e itself is a function of time, ‘‘slowly’’ rising with a characteristic time τ_ε. Thus, at short times ($t < \tau_E$) after the imposition of a strong electric

field, the carrier drift occurs with a low-field mobility and the velocity can substantially overshoot its steady-state value (Ruch, 1972). The overshoot phenomenon has become quite important in determining the speed of modern transistors with ultra-short gate lengths (Sai-Halasz et al. 1988). For a 0.25 μm n-channel Si MOSFET this hot-electron effect contributes a 20% enhancement in the transistor speed (Pinto, 1991).

The qualitative explanation given above is quite adequate for describing the overshoot in Si and Ge, where there is a strong dependence of the phonon scattering rates on the carrier energy. In GaAs and some other III-V compounds, the dominant scattering process in the Γ valley is due to polar optical phonons, and the rate of these processes at sufficiently high electron energies becomes nearly independent of energy. In such materials the mobility degradation at high energies is associated with the transfer of electrons to the lower-mobility upper valleys (Ridley and Watkins, 1961). Consequently the overshoot is not seen below the threshold field for the negative differential mobility effect ($F \gtrsim 3.2$ kV/cm in GaAs and $\gtrsim 11$ kV/cm in InP).

17.2.2 Ballistic. If the force acting on electrons suddenly changes, their subsequent motion for a short time may be considered without taking collisions into account.[3] The time interval $t \lesssim \tau_\beta$ when this is possible, is determined not only by the momentum relaxation time τ_m but also by the characteristic time τ_{ee} of interelectron scattering. By definition, the ee scattering rate is included neither in the rate $1/\tau_m$ which enters Eq. (6b) nor in the expression $\mu = (e/m)\tau_m$ for steady-state mobility, because ee collisions have no direct effect on $\langle \mathbf{v} \rangle$. Of course, they can have a very strong effect indirectly, by influencing other collision processes through the shape of the distribution function $f(\mathbf{v})$. At carrier concentrations $n \gtrsim 10^{17}\ \mathrm{cm}^{-3}$ the characteristic time τ_β for ballistic transport,

3. The common term ‘‘ballistic’’ applied to this time of motion, conjures up the image of a projectile moving in airless space. This image is not very apt, since interesting properties of the electronic motion in the ballistic regime often depend crucially on the band structure, which defies cannonball analogy.

$$\frac{1}{\tau_\beta} = \frac{1}{\tau_m} + \frac{1}{\tau_{ee}} \ , \tag{7}$$

may be considerably shorter than τ_m.

The concept of ballistic motion also applies to the *steady-state* transport. In this case one considers regions a short distance $d \lesssim \lambda_\beta$ away from an abrupt potential variation, Fig. 1. The characteristic length λ_β is related to τ_β by $\lambda_\beta = \langle v \rangle \tau_\beta$ and both quantities depend on the shape of the distribution function f (**v**). Instead of parameters pertaining to an electron ensemble, one often defines similar parameters for a given state of electron motion, viz. its lifetime $\tau(\mathbf{v})$ and the mean free path $\lambda(\mathbf{v})$, as limited by collisions with the lattice and other electrons. These definitions practically coincide for narrow distributions.

Steady-state ballistic transport was first demonstrated experimentally in unipolar heterostructures, illustrated in Fig. 2. The idea of ballistic-electron spectroscopy (Hesto et al., 1982) consists in the following: one measures the dependence of the collector current I_C on the collector-base bias V_{CB} at a fixed emitter-base bias V_{BE} and plots $(\partial I_C/\partial V_{CB})_{V_{BE}}$ versus V_{CB}. If certain important conditions are met, the resultant curve is proportional to the number of carriers arriving at the collector barrier with a normal component of the kinetic energy (i.e., portion of the energy corresponding to the motion normal to the barrier) equal to the barrier height Φ_C. For this to be true, one must ensure that the collector barrier height linearly depends on the bias, $\delta\Phi_C \propto \delta V_{CB}$, and that the collector bias does not affect the hot-electron energy distribution in the base, the emitter injection efficiency, and the above-barrier QM reflection. Studying the $I_C(V_{CB})$ dependence of a GaAs/AlGaAs heterostructure similar to that illustrated in Fig. 2, Heiblum et al (1985) found a sharp peak in the derivative characteristic. According to authors interpretation, the majority of electrons contributing to the peak arrive at the analyzer without a single scattering event − otherwise, the peak would be displaced by at least 36 mV (corresponding to the optical-phonon energy $\hbar\omega_{op}$) to lower voltages, which could be accounted for only by postulating an unrealistically low barrier height.

In an external field ballistic electrons can be accelerated to velocities much higher than the steady-state saturated velocity v_{sat} [up to the maximum band velocity $dE/d(\hbar\mathbf{k})$; for electrons in the conduction band of GaAs accelerated in a <100> direction this limit is $\simeq 10^8$ cm/sec] and such an enhancement is important and beneficial for the performance of semiconductor devices. It appears attractive (Shur, 1987) to realize ballistic transport in short-channel field effect transistors, so that the carrier velocity at any point in the channel would be determined by the conservation of energy. Such a situation is realized in vacuum diodes, where the current is space charge limited and is described by the Child-Langmuir law. Despite considerable efforts, no semiconductor device structure has been demonstrated to-date in which the current-voltage characteristics would convincingly conform to a similar law. In practice, conditions for collisionless transport in a 2D channel (where sheet carrier concentrations typically much exceed 10^{11} cm^{-2}) are very difficult to realize because of ee collisions (which typically result in $\lambda_\beta < 1000$ Å at these concentrations).

When the rate of ee collisions is very high, as is usually the case in the channel of a field-effect transistor in its *on* state, the electron ensemble behaves rather like a fluid than a gas. In the absence of *other* collisions (with phonons and impurities), the electronic fluid in the FET channel is described by hydrodynamic equations similar to those for shallow water (Dyakonov and Shur, 1993). Based on this analogy, Dyakonov and Shur (1995a,b) discussed several new effects related to plasma oscillations in the 2D electron fluid. In particular, a short-channel high-moblity transistor has a resonant response to an electromagnetic radiation at the plasma wave frequency of the 2D electrons (Dyakonov and Shur, 1996). This effect can be used to implement detectors, mixers and multipliers at terahertz frequencies. As pointed out by the authors, these devices should operate at much higher frequencies than conventional, transit-time limited devices, since the plasma waves propagate much faster than electrons. Moreover, their responsivities and conversion efficiencies can be expected to greatly exceed those of Schottky diodes currently used as detectors, mixers and multipliers in the terahertz range.

17.2.3 Mesoscopic. When the carrier concentration is sufficiently low and ee collisions are rare, the shape of their distribution function under external perturbation may depart considerably from the quasi-equilibrium form (3). The Maxwellian shape (Gaussian in velocities) can be viewed − in the spirit of the well-known central limit theorem of statistics − as resulting from large number of independent scattering events, each contributing or withdrawing a random amount of energy.[4] Besides ee collisions, the electron distribution function can be maxwellized by other interactions as well, provided independent scattering events exchange random energies with electrons. Scattering by acoustic phonons has this property, while optical phonon scattering does not. If the latter were the only inelastic interaction, the equilibrium shape of the electron distribution function would be rather strange and thermodynamic properties of the electron ensemble rather different. The peculiarity of interaction with optic phonons stems from their largely monochromatic nature, which quantizes the energy exchange in units of $\hbar\omega_{op}$.

At sufficiently high temperatures, the electron energy relaxation rate due to optical phonons ($1/\tau^{(op)}$) is higher than that due to acoustic phonons ($1/\tau^{(ac)}$) by several orders of magnitude. Typically, in semiconductors $\tau^{(op)} \lesssim 10^{-12}$ s and $\tau^{(ac)} \gtrsim 10^{-9}$ s (Conwell, 1967). This disparity of the inelastic relaxation times can lead to the formation of an electronic ensemble that is in equilibrium with the optical-phonon system but has not yet appreciably interacted with acoustic phonons. Manifestation of these properties in electronic transport can be conveniently referred to as the ‘‘classical mesoscopic effects’’ (drawing a parallel to and a distinction from the quantum mesoscopic effects that occur when the coherence length or time of an electronic wave function exceeds characteristic system dimensions. Grinberg and Luryi (1990) considered the kinetics of an electron ensemble initially characterized by a Maxwellian distribution with $T_e = T_i$, subject to interaction with the lattice at equilibrium temperature T. Since the initial

4. The form of distribution (3) does not contradict the central limit theorem. In the presense of the quantum correlation between electrons, maintained by the Pauli exclusion principle, multiple collisions lead to a Fermi rather than Maxwell distribution.

distribution is not the equilibrium Boltzmann function, it evolves in time because of the electron interactions with optical phonons, acoustic phonons, and due to the e-e scattering. Under the assumption that $\tau^{(op)}/\tau^{(ac)} \ll 1$ and $\tau^{(op)}/\tau_{ee} \ll 1$, electrons rapidly establish a quasi equilibrium with the optical phonon field (the mesoscopic state) and then − on a longer scale − the true equilibrium is established by other inelastic scattering processes. Even though the e-e scattering does not change the average electron energy, it counts as an inelastic interaction, because it changes the shape of the non-stationary distribution.

The distribution function of the mesoscopic state can be determined from the statistical consideration alone, without actually solving the kinetic equation. The shape of this function is illustrated in Fig. 3 both for the narrow ($T_i < T$) and broad ($T_i > T$) initial distributions. Even though the electron system is in perfect equilibrium with optical phonons at temperature T, thermodynamic properties of electrons in the mesoscopic state are very different from those in true equilibrium, e.g., the average energy $<E> \neq (3/2)kT$ and the specific heat deviates from the classical value $3k/2$ (Grinberg and Luryi, 1990). Also the electron mobility in the mesoscopic state shows a strong overshoot. The time scale of this effect is very different from the conventional velocity overshoot; the mesoscopic state is established in less than 1 ps and persists for a long time (up to nanoseconds!) − controlled by acoustic phonon and interelectron scattering. Systematic numerical study of the classical mesoscopic kinetics has been carried out by Grinberg et al (1991).

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

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

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:

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

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),

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

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

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

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

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

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

$$J(T_e) = e\,n\,\mu(T_e)F(T_e) \equiv en\,\bar{v}(T_e) . \tag{13}$$

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 $T_{\rm e}$ and hence the injection current $I_{\rm C}$ by the input voltage $V_{\rm D}$.

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 $I_{\rm D}$ 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 ($\Delta E_C \approx 0.25\,\text{eV}$) 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 $\langle \mathbf{v} \rangle$ 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

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

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).

Maezawa and Mizutani (1991) studied the microwave performance of a $1\,\mu$m-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 $V_{\rm D}$ (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

$$\begin{aligned} \mathbf{ornand}\,(\{\mathbf{V_j}\}) &\equiv (\mathbf{V}_1 \bigcup \mathbf{V}_2 \bigcup \mathbf{V}_3) \bigcap (\bar{\mathbf{V}}_1 \bigcup \bar{\mathbf{V}}_2 \bigcup \bar{\mathbf{V}}_3) && (15)\\ &= \mathbf{V}_1 \bigcup \mathbf{V}_2 && \text{for } \mathbf{V}_3 = 0 \quad (\mathbf{or})\\ &= \bar{\mathbf{V}}_1 \bigcup \bar{\mathbf{V}}_2 = \overline{\mathbf{V}_1 \bigcap \mathbf{V}_2} && \text{for } \mathbf{V}_3 = 1 \quad (\mathbf{nand}) \end{aligned}$$

where the symbols $\bigcap$, $\bigcup$, and $\bar{\mathbf{A}}$ stand for logic functions **and**, **or**, and **not A**, respectively. When $\mathbf{V}_3$ is low, $\mathbf{ornand}\,(\{\mathbf{V_j}\}) = \mathbf{or}\,(\mathbf{V}_1, \mathbf{V}_2)$, and when $\mathbf{V}_3$ is high, $\mathbf{ornand}\,(\{\mathbf{V_j}\}) = \mathbf{nand}\,(\mathbf{V}_1, \mathbf{V}_2)$. 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.

---

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.

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 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

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 $\mathbf{k}_{\|} \equiv (k_x, k_y)$ 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).

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 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

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.

perpendicular to the base layer, is beneficial for downscaling the device area (Jalali et al. 1989). 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.

17.4 Hot electrons in lasers

Carrier heating in double heterostructure lasers arises mainly due to injection of electrons and holes into the narrow-gap active region "over the cliff" from the wide-gap cladding layers,

cf. Fig. 16. Another contribution to heating, which becomes appreciable far above the threshold, is free carrier absorption of the coherent radiation in the laser cavity. The energy balance in the active region is described by Eq. (6a) with the average power input per carrier approximately given by

$$\mathcal{P} = \frac{\alpha S\, \hbar\Omega}{2} + \frac{J\,\Delta}{2n} \quad , \tag{21}$$

where $J \equiv I/eVA$ is the electron flux per unit volume VA of the active layer, I the pumping current, S the photon density in the active layer, Ω the optical frequency, Δ the kinetic energy per carrier injected into the active region, and $\alpha > 0$ the free carrier absorption rate. It is assumed in Eq. (21) that effectively only one type of carriers absorbs radiation and that the densities of electrons and holes are the same in the active layer, $n = p$, hence the factor of 2.

For the free-carrier absorption coefficient Henry et al. (1983) cite $\alpha_0 = 25\,\mathrm{cm}^{-1}$ at $p = 10^{18}\,\mathrm{cm}^{-3}$ and $\lambda = 1.6\,\mu\mathrm{m}$. Our α above is related to α_0 by $\alpha_0 = \alpha p\, \bar{c}$, where $\bar{c} \equiv c/\kappa_g$ is the group velocity of light. The energy Δ depends on the laser structure, e.g., the fraction of carriers that get into the active region by tunneling; in general we can say that $\Delta \lesssim \Delta E_G$, where ΔE_G is the bandgap difference between the cladding and the active layers. For InP cladding ($E_G = 1.35\,\mathrm{eV}$) and $\lambda = 1.55\,\mu\mathrm{m}$ active layer, $\Delta E_G = 0.55\,\mathrm{eV}$.

Taking the average electron energy in the approximate nondegenerate form (4) with $\langle \mathbf{v} \rangle = 0$, we can estimate the carrier overheating near the threshold ($S \approx 0$, $J \approx n/\tau_{sp}$):

$$k\,(T_e - T) = \frac{\tau_\varepsilon}{\tau_{sp}}\,\frac{\Delta}{3} \quad , \tag{22}$$

where τ_{sp} is the lifetime of carriers (including nonradiative processes). With $\tau_{sp} \approx 0.5\,\mathrm{ns}$ and $\tau_\varepsilon \approx 1\,\mathrm{ps}$ (Hall et al. 1992) we find from Eq. (22) an overheating by $(T_e - T) \approx 4\,\mathrm{K}$.

It should be noted that experimental estimates for τ_ε vary widely, because in a high-density bipolar plasma one can expect a strong dependence of τ_ε on the carrier concentration and temperature. Indeed, the dominant energy-loss mechanism in long-wavelength semiconductors is due to the emission of polar optical phonons and typically leads to $\tau_\varepsilon < 1\,\mathrm{ps}$. However, at

high carrier concentrations the hot-carrier cooling rates may be much slower, especially in quantum wells, as has been found by a number of workers who studied luminescence of photo-excited carriers on a picosecond scale (Westland et al. 1988; Lobetanzer et al. 1989). Similar results were found by Belenky et al. (1994b) by measuring thermionic emission of heated carriers from a quantum well. The anomalously long relaxation times τ_ε, substantially exceeding the characteristic time of optical-phonon emission, can be explained by hot phonon effects (Kumekov and Perel, 1988). Because of the small group velocity, optical phonons do not move appreciably during their lifetime and are likely to be re-absorbed by the free carriers, provided the carrier plasma has a high enough concentration (typically, higher than $10^{18}\,\mathrm{cm}^{-3}$). In this situation, termed the *optical phonon bottleneck*, the optical phonon gas and the carriers are characterized by the same effective temperature T_e and the energy relaxation of the coupled carrier-phonon system is determined by the process of polar optical phonon decay into acoustic phonons. The above estimate of $4\,\mathrm{K}$ overheating at room temperature near the laser threshold is rather modest compared to calculations that include the bottleneck effect (see, e.g., Tsai et al. 1993). Nevertheless, even this small amount of carrier overheating may have important consequences for applications.

A primary example of such applications is the subcarrier-multiplexed optical transmission system used for cable TV distribution. These systems require source lasers with an extremely *linear* functional relationship between the input current $I(t)$ and output light intensity $L(t)$. As functions of time, these quantities encode all the information carried by multichannel cable TV networks. Even a small amount of nonlinearity brings about some intermodulation distortion (e.g., TV channels operating at 180 and 240 MHz produce unwelcome parasitic signals around 60 and 420 MHz). Channel capacity in current systems is about 40 channels, but future cable networks are expected to carry much larger number of channels. Linearity specifications for an 80-100 channel system become extremely stringent.

The unwelcome consequences of hot-carrier effects arise from the dependence of the optical gain g on the carrier temperature. This dependence is typically of the form

$$g\,(T_e, n, \Omega) = g_0\,(f_e + f_h - 1)\ , \tag{23}$$

where f_e and f_h are the Fermi functions of electrons and holes, respectively, at energies selected by the incident photons $\hbar\Omega$ and g_0 is a constant. Dynamics of the laser can be described by the following system of rate equations (Gorfinkel and Luryi, 1995):

$$\frac{dn}{dt} = J - g\,S - \frac{n}{\tau_{sp}}\ ; \tag{24a}$$

$$\frac{dS}{dt} = \Gamma S\,(g - \alpha n) - \frac{S}{\tau_{ph}}\ , \tag{24b}$$

$$\frac{dT_e}{dt} = \frac{2}{3}\mathcal{P} - \frac{T_e - T}{\tau_\varepsilon}\ , \tag{25}$$

where τ_{ph} is the lifetime of photons in the laser cavity and $\mathcal{P}$ is the power input given by Eq. (21). Nonlinear effects originate from the terms $g \cdot S$ and $S \cdot n$ in Eqs. (24).

If the carrier temperature were constant, the gain in a given optical mode of the laser would depend only on the concentration n. Since above the threshold the value of n is "pinned" in a static regime by the condition of stable generation, g = loss, the dependence of the output light power ($\propto S$) on the pumping current J is linear to a good approximation. Small nonlinear effects are possible at high modulation frequencies, when the carrier concentration deviates from the pinned value. These effects contribute the so-called "intrinsic" distortion (Darcie et al. 1985). The latter goes as $(f/f_r)^2$, where $(2\pi f_r)^2 = S_0\, g_n'/\tau_{ph}$ is the resonant frequency of the laser operated around a steady-state output level S_0 (cf. Eq. 32 below). The intrinsic distortion decreases with S_0 and becomes small at sufficiently high bias currents. However, increasing the steady-state bias introduces additional nonlinearities, associated with carrier heating.

Because of the T_e variation above threshold, carrier concentration is no longer "pinned" even in the static regime. Since normally $g_T' < 0$ and $g_n' > 0$, the increasing T_e with increased pumping current is accompanied by an increase of carrier concentration n in the active region. An important adverse consequence of the T_e variation is a current-dependent carrier leakage out of the active region in laser heterostructure. This leakage contributes an unwelcome intermodulation distortion. It can be minimized in properly designed double-heterostructure lasers (Belenky et al. 1995).

Recently, Gorfinkel and Luryi (1995) considered nonlinear distortions in systems arising from carrier heating effects which cannot be avoided in any laser structure, since they are brought about by the very existence of the modulated optical and electrical signals. These effects are described by Eq. (21) − namely, the free-carrier absorption of the coherent radiation in the cavity and the power flux into the active layer, associated with the input current. It turns out that the resultant intermodulation distortion approaches the tolerance limit for an 80-channel cable TV system − even if all other "parasitic" nonlinearities are eliminated, including the above mentioned carrier leakage. Thus, hot carrier effects present a fundamental limitation for the number of channels available on cable television.

Another fundamental nonlinearity in lasers is associated with carrier heating by the power released in Auger recombination processes. In these processes the potential energy of an electron-hole pair is transferred to free carriers. Gorfinkel and Luryi (1995) showed that even though this power is typically comparable to that due to free carrier absorption, its contribution to the intermodulation distortion is negligible, being of higher order in the optical modulation depth. The rate of Auger recombination goes as $C_A n^3$, see e.g., Agrawal and Dutta (1993), and hence above threshold it is independent of the signal level.

Carrier heating by Auger recombination processes may play a very important role in long wavelength semiconductors, especially in single QW lasers where, due to a small value of the optical confinement factor Γ the carrier concentration at the lasing threshold reaches rather high values, typically much exceeding $10^{12}\,\mathrm{cm}^{-2}$. At high injection current I, this may lead to a substantial carrier heating. Inclusion of this effect in the energy balance equation is accomplished by adding a term $\sim C_A n^2 E_G$ into Eq. (21). The increasing carrier temperature $T_e(I)$ suppresses the optical gain g and may lead to the appearance of a maximum g_{max} in the dependence $g(I)$ of gain on the pumping current.[9] If the total losses in the laser cavity exceed

9. This discussion follows an unpublished work by V. B. Gorfinkel and S. Luryi (1992).

g_{max} then the structure will not lase at any pumping current. Since the Auger coefficient C_A itself increases with the temperature, one can expect a sharp disappearance of lasing at some critical temperature. Similar nonmonotonic $g(I)$ dependence with a maximum, which results from electron heating and is accompanied by a sharp disappearance of lasing with increasing temperature, is important in the operation of the new *unipolar* laser and will be discussed in the next section. Note that for a constant T_e the dependence $g(I)$ is always monotonic. If the losses do not exceed g_{max}, then the laser generation regime can be reached, but the negative slope of $g(I)$ characteristic results in peculiar instabilities for currents exceeding $I_{cr} \equiv I(g_{max})$. For the same value of g one can have two regimes that differ in n and T_e and, most importantly, in the output radiation power. The high-T_e regime corresponds to higher n and lower $\mathcal{P}$. This regime is metastable. Illumination of the laser by a sufficiently powerful pulse of external light temporarily suppresses the Auger recombination and switches the laser into the stable regime with a high $\mathcal{P}$.

17.4.1 Quantum cascade lasers. The QCL is a new mid-infrared laser, based on unipolar transitions of electrons between energy levels created by quantum confinement. Since the first demonstration of QCL (Faist et al. 1994), its design has continuously improved through a series of elegant innovations by the Bell Labs group (Faist et al. 1995a,b) culminating in their recent report of a high power room-temperature operation (Faist et al. 1996).

Kinetics of the carrier transport in the QCL and the device structure are schematically illustrated in Fig. 17. The device has multiple periods (25 in reported structures) and each electron performs a cascade of transitions down the periodic ladder. Under the operating conditions, each period must have net zero charge: by Gauss' law this is necessary in order that subsequent periods could replicate each others electrostatic state. To avoid the space charge accumulation, associated with current flow, one therefore needs a reservoir of positive fixed charge, compensating the negative mobile charge in each period. Introduction of such a reservoir, implemented as a doped superlattice region is the key design innovation that led to

the successful implementation of a unipolar laser. Subsequent QCL designs entrusted the reservoir with an additional mission to suppress the unwelcome tunneling from the upper level into the continuum. For this purpose, the superlattice is implemented as an electronic Bragg filter with a stop band in the range of energies of upper level states (Faist et al. 1995). Thus, most carriers enter the QW only via the upper and leave only via the lower subband, while leakage from the upper subband directly into the reservoir is negligible.

Theoretical analysis of the QCL operation (Gorfinkel et al. 1996) suggests that it is dominated hot-electron effects. Non-equilibrium electron distributions arise from the power $\mathcal{P} \approx J \cdot \hbar\Omega$ per unit area, dissipated in each cascade period. The energy stored in the *transverse degrees of freedom*, corresponding to in-plane motion of carriers, fundamentally changes the lineshape of intersubband resonance and the spectral characteristics of gain. Depending on the laser design, there may be different scenarios of how this energy is distributed among various electrons and dissipated into the lattice. The simplest (but not the most advantageous) scenario arises at high carrier concentrations, when the electron-electron interaction is sufficiently fast to equilibrate the input power among all free carriers in a QCL period. In this case, the kinetic energy of two-dimensional electrons can be characterized by an effective temperature T_e, which is moreover the same in both subbands. This temperature can be found from the balance equation,

$$\frac{k\,\mathrm{d}T_e}{\mathrm{d}t} = \mathcal{P} - \frac{n_D k\,(T_e - T)}{\tau_\varepsilon} \quad , \tag{26}$$

where $n_D \sim 10^{11}\ \mathrm{cm}^{-2}$ is the doping concentration per period. In the steady state the overheating is directly proportional to the current: $k\,(T_e - T) = J\,\tau_\varepsilon \hbar\Omega / n_D$.

The subband concentrations in the QCL are found from the following rate equations:

$$\frac{\partial n_2}{\partial t} = J - \frac{n_2}{\tau_{21}} - g\bar{S} \quad , \tag{27a}$$

$$\frac{\partial n_1}{\partial t} = \frac{n_2}{\tau_{21}} + g\bar{S} - \frac{n_1}{\tau_{1\,\mathrm{out}}} \quad , \tag{27b}$$

where J is the current flux and S is the photon density per unit area in the lasing mode

(frequency Ω_L). The gain $g = c\bar{g}\,(\Omega_L)$ is the intersubband gain (in units of $\sec^{-1}$) at the lasing frequency, where $c = c/\sqrt{\kappa_\infty}$ is the speed of light in the lasing mode. The time constants τ_{21} and $\tau_{1\,\mathrm{out}}$ describe, respectively, the rate of nonradiative intersubband transitions and carrier removal from the bottom subband. The ratio of these constants, $\xi_0 \equiv \tau_{1\,\mathrm{out}}/\tau_{21}$, is an important design parameter of an intersubband laser, as it determines the carrier density ratio in a steady state below threshold, $n_1 = \xi_0 n_2$.

To illustrate the importance of hot-electron effects, consider an oversimplified model of the QCL which neglects the *nonparabolicity* in the conduction band. In this case one can use the well-known expression for the gain of a two-level system (Yariv, 1991, Chap. 5)

$$g \equiv \frac{4\pi e^2 |z_{12}|^2 \Omega}{\hbar c \sqrt{\kappa_\infty}\,\gamma_0} \frac{n_2 - n_1}{a} \quad , \tag{28}$$

where z_{12} is the transition matrix element, $\sqrt{\kappa_\infty}$ the refractive index, a the quantum well width and $\hbar\gamma_0$ is the spontaneous linewidth. For high T_e the dominant contribution to the linewidth results from the emission of optic phonons by electrons making transitions within the same subband. One can therefore assume that γ_0 depends on the effective temperature of the electron ensemble.

From the generation condition $g = \alpha_{\mathrm{Loss}}$ we find that the threshold is described by the following equation for T_e:

$$\frac{\hbar\gamma_0(T_e)}{k\,(T - T_e)} = \mathcal{R} \quad , \tag{29}$$

where $\mathcal{R}$ is a dimensionless constant,

$$\mathcal{R} = \frac{4\pi e^2 |z_{12}|^2 \Gamma n_D}{\hbar c \sqrt{\kappa_\infty}\, a\, \alpha_{\mathrm{Loss}}} \frac{\tau_{21} - \tau_{1\,\mathrm{out}}}{\tau_\varepsilon} \quad . \tag{30}$$

Consider a graphical solution to Eq. (29), Fig. 18, assuming that the intersubband linewidth γ_0 is an increasing and *concave* function of the effective carrier temperature.[10] The threshold

10. Intersubband photoluminescence experiments indicate that $\gamma_0(T)$ is an increasing concave function of the *lattice* temperature (J. Faist, private communication).

carrier temperature is determined by the first intersection of a line $(T_e - T)\cdot\mathcal{R}$ with the concave upward curve $\hbar\gamma_0(T_e)$. Depending on the value of $\mathcal{R}$ and the ambient temperature, equation (29) may have no solutions at all or − if the value of $\mathcal{R}$ is sufficiently large − *two* solutions. For a given $\mathcal{R}$ the highest lattice temperature $T = T_{max}$ at which threshold is possible, is determined by the condition that the line is *tangent* to the curve. Finally, there is a minimum value of $\mathcal{R}$ such that $T_{max} > 0$. If $\mathcal{R}$ is less than this minimum value, then there can be no lasing at any temperature.

From this qualitative example we conclude that both the transverse relaxation of the intersubband resonance and the hot-carrier effects are indispensable for the correct physical understanding of the temperature behavior of quantum cascade lasers. Quantitatively, the two-level model (28) does not adequately describe the intersubband gain function in experimentally realized (Faist et al. 1994-1996) QCL heterosystems. The reason is two-fold. Firstly, one cannot neglect the nonparabolicity: its inclusion suppresses the peak gain value by more than an order of magnitude. Secondly, the relaxation rate of intersubband resonance, which is mainly due to transverse intrasubband scattering processes, is strongly dependent on electron kinetic energy, $\gamma = \gamma(E)$. Indeed, the rate of optical phonon scattering has a threshold nature: when $E > \hbar\omega_{op}$ it increases by nearly an order of magnitude due to the onset of intrasubband emission. The generalized expression for the intersubband gain at an optical frequency Ω is of the form (Gelmont et al. 1996):

$$g(\Omega) = \frac{4e^2 |z_{12}|^2 m_2 \Omega}{\hbar^3 a\, c\sqrt{\kappa_\infty}} \int_0^\infty \frac{d\varepsilon\, \gamma(\varepsilon)\,[f_2(\varepsilon) - f_1(\varepsilon_1)]}{[\Omega - \Omega_\varepsilon]^2 + [\gamma(\varepsilon)]^2} \quad , \tag{31}$$

where Ω_ε is the optical transition frequency for the in-plane electron momentum $\hbar k = \sqrt{2m_2\varepsilon}$, viz. $\hbar\Omega_\varepsilon \equiv \hbar\Omega_0 + \varepsilon_2 - \varepsilon_1$, where $\varepsilon_2 \equiv \varepsilon$ and $\varepsilon_1 = \hbar^2k^2/2m_1$ are kinetic energies in the upper and lower subbands, respectively, characterized by the effective masses m_2 and m_1 and the distribution functions f_2 and f_1. The function $\gamma(\varepsilon)$ describes the transverse phase relaxation rate due to intrasubband scattering, both elastic and inelastic.

It is worthwhile to stress that Eq. (31) does not rely on the electron temperature approximation and remains valid for general electron energy distributions within each subband. The most attractive situation in fact arises at low carrier concentrations ($n_D \ll 10^{11}$ cm^{-2}) when electron-electron collisions are not fast enough to establish a common T_e between the two subbands. It turns out (Gorfinkel et al. 1996) that in this regime the lower-subband electron distribution can be approximately characterized by a "negative" temperature, so that states near the subband bottom are mostly unoccupied.[11] In this case it is possible to achieve positive values of the gain at room temperature even in the absence of an overall population inversion between the two subbands.

For higher (but still moderate, $n_D \lesssim 10^{11}$ cm^{-2}) concentrations, the T_e approximation becomes applicable, but it is still possible (at least semi-quantitatively) to regard optical phonon scattering as the dominant phase breaking mechanism. In this range, Gorfinkel et al. (1996) predict a peculiar hot electron effect manifested in the dependence of the lasing wavelength λ on the pump current, including a regime where λ switches "digitally" between two stable values. Still higher concentrations require a special consideration, as it becomes necessary to explicitly include electron-electron scattering in the calculation of $\gamma(\varepsilon)$. Also the energy balance equation in this range should include the dependence of the electron cooling rate on the carrier density and effective temperature due to optical phonon bottleneck. Validity of Eq. (31) itself is not restricted to any particular scattering mechanism responsible for the transverse phase relaxation.

17.4.2 Hot-electron lasers. Due to the dependence $g(T_e)$, Eq. (23), variations in the carrier temperature can produce rapid changes in the optical output of a semiconductor laser (Rivlin, 1985). Possible device applications of this effect stem from the observation (Gorfinkel and

11. For high positive temperatures the electron distribution becomes uniform over the entire subband. For $T_e < 0$ the distribution is inverted, the lower energy states having smaller occupation probability; one can say that negative temperatures are higher than $T_e = \infty$,

Filatov, 1990) that controlling the laser by an independently modulated T_e offers a higher modulation bandwidth than the conventinal pump current control. In general, the modulation response of a semiconductor laser $\mathcal{R}_J(\omega) \equiv \delta S/\delta J$ (at a constant T_e) is limited in bandwidth by an intrinsic resonance in the nonlinear laser system (the electron phonon resonance). Due to this effect, the response $\mathcal{R}_J$ reaches a maximum near the resonance frequency ω_r where

$$\omega_r^2 = \frac{S_0 g_n'}{\tau_{ph}} \quad , \tag{32}$$

and then decreases by 20 dB/dec, i.e. as $1/\omega^2$. The analogous response function for T_e modulation $\mathcal{R}_T(\omega) \equiv \delta S/\delta T_e$ (at a constant J) rolls off at high frequencies only as $1/\omega$, i.e. by 10 dB/decade (Gorfinkel et al. 1991). This result can be demonstrated by a small-signal analysis of rate equation (24) taking both J and $\mathcal{P}$ as externally varied parameters.

Rapid electron heating can be accomplished in a variety of ways. The original suggestion by Gorfinkel et al. (1991) was to employ a four terminal laser structure with a carrier-heating current driven laterally through the active layer. High-frequency modulation of T_e by several degrees was demonstrated in such a structure (Bagaeva et al. 1991). Other proposals include power delivery to the electronic system by intersubband absorption in quantum wells (Noda et al. 1990; Gorfinkel and Luryi, 1992) and by a vertical hot-electron injection in a transistor-like heterostructure (Tolstikhin and Mastrapasqua, 1995).

Modulation of the laser output power via carrier heating at a constant pumping has the potential of offering higher bandwidth of modulation. Still higher potential for applications lies in the possibility of a *synchronous* control of both the T_e and the pumping current (Gorfinkel and Luryi, 1993, 1994). This technique, called the dual modulation, permits in principle a complete elimination of electron-photon oscillations and linearization of the light-current characteristics in a wide frequency band. Indeed, with an independent control of gain via T_e one can impose the condition of constant n even as both J and S vary. Setting $dn/dt = 0$ in Eqs. (24). we find

$$0 = J - J_{\rm th} - gS \quad ; \tag{33a}$$

$$\frac{{\rm d}S}{{\rm d}t} = \Gamma (J - J_{\rm th}) - \tau_{\rm ph}^{-1} S \quad , \tag{33b}$$

where $J_{\rm th} \equiv n_{\rm th}^2 / \tau_{\rm sp}$ and $n_{\rm th}$ is the pinned carrier concentration above threshold. For the sake of simplicity, we have neglected the free carrier absorption term. Equation (33b) establishes a *linear* relationship between the functions $S(t)$ and $J(t)$:

$$S = = \Gamma {\rm e}^{-t/\tau_{\rm ph}} \int_0^t [J(t') - J_{\rm th}]\, {\rm e}^{t'/\tau_{\rm ph}}\, {\rm d}t' \quad . \tag{34}$$

For this property to hold we must satisfy Eq. (33a) with the help of a simultaneous variation of $T_{\rm e}$, viz.

$$g[T_{\rm e}(t), n_{\rm th}] = \frac{J(t) - J_{\rm th}}{S(t)} \quad . \tag{35}$$

Having solved Eq. (35) for the time dependence $T_{\rm e}(t)$, we can determine the required heating power signal from Eq. (25). Linearity of the $S[J(t)]$ relationship is of great value for optical communication systems.

Small signal analysis of dual modulation can be done by linearizing Eqs. (24):

$$J(t) = J_0 + \delta J\ {\rm e}^{{\rm i}\omega t} \quad ; \tag{36a}$$

$$g(t) = g_0 + \delta g\ {\rm e}^{{\rm i}\omega t} \quad , \tag{36b}$$

where $\delta g = g_n' \, \delta n + g_T' \, \delta T_{\rm e}$. Under the condition ${\rm d}n/{\rm d}t = 0$ we find the following response function:

$$\delta S = \frac{\delta J}{g_0 (1 + {\rm i}\omega\tau_{\rm ph})} \quad , \tag{37a}$$

while the required relation between the dual inputs δg and δJ is of the form

$$\delta g \equiv g_T' \, \delta T_{\rm e} = \frac{{\rm i}\omega\tau_{\rm ph}}{1 + {\rm i}\omega\tau_{\rm ph}} \, \frac{\delta J}{S_0} \quad . \tag{37b}$$

We remark that the "target condition" (37b) is practically frequency-independent for $\omega\tau_{\rm ph} > 1$ and that the magnitude of the required dual modulation input is inversely proportional to S_0. When this relation is fulfilled, then there is no electron-photon resonance in the system and the modulation efficiency decays with frequency as $1/\omega$. Small-signal response of the laser output

power is plotted in Fig. 19 for three types of modulation: (1) purely by the pumping current, (2) purely by the electron temperature, and (3) by their coherent combination as in Eq. (37b).

Suppression of relaxation oscillations of the carrier density makes possible a high repetition rate coding of information with short pulses. Calculated laser response to a 10 Gb/s series of dual current-temperature pulses (Gorfinkel and Luryi, 1993) is practically undistorted compared to the single pulse situation. It is clear that small-signal pulses δJ of *any* shape, as well as analog signals, can be transmitted in a regime of constant n, provided the system can Fourier analyze $\delta J(t)$ and form in real time an appropriate complementary pulse $\delta T_e(t)$. The dual modulation method is not limited to the variation of T_e. Any of the parameters of the system of rate equations (24), such as the modal gain and the photon lifetime, may be varied externally.

Control of the semiconductor injection laser by modulating T_e is not the only way of utilizing hot electrons in lasers. In fact, there is a far more advanced laser application of hot electrons. *Injectionless* far infrared lasers based on *hot hole* effects in germanium have been demonstrated several years ago.[12] In these devices population, the inversion of hot holes arises due to the very different dynamical behavior of heavy and light holes in crossed electrical and magnetic fields. Discussion of these effects would be out of scope of this chapter.[13]

17.5 Conclusion

We have reviewed a number of possible device applications of hot electrons in semiconductors. The wide class of hot-electron devices were classified into two groups, the ballistic devices and the quasi-thermal (T_e) devices, depending on the type of a hot-electron ensemble essentially employed in their operation. At this time, the quasi-thermal group appears more promising for applications, if only because the corresponding nonequilibrium electronic ensemble is easier to

13. For an introductory discussion, the reader is referred to an early review by Andronov (1987) as well as to several papers in the special issue on far-infrared semiconductor lasers, *Opt. Quantum Electron.* **23**, No. 2 (1991).

maintain in a robust and reproducible fashion. Indeed, modern electronic devices operate typically with the concentration of carriers in the active region exceeding 10^{17} cm^{-3}. In this case, the electron-electron interaction is very fast, dominating most other scattering and transport properties and maintaining a well defined effective temperature T_e.

References

Agrawal, G. P. and Dutta, N. K. (1993) *Semiconductor Lasers*, 2nd edition, Van Nostrand Reinhold, New York.

Akeyoshi, T., Maezawa, K., Tomizawa, M., and Mizutani, T. (1992) "Monte Carlo study of charge injection transistors (CHINTs)", *Japanese Journal of Applied Physics* **32**, 26-30.

Andronov, A. A. (1987) "Hot electrons in semiconductors and submillimeter waves (review)", *Fizika i Tekhnika Poluprovodnikov* **21**, 1153-1187 [English translation: *Soviet Physics - Semiconductors* **21**, 701-721 (1987)].

Bagaeva, T. Yu., Filatov, I. I., Gorbovitsky, B. M., Gorfinkel, V. B., Avrutin, E., and Gurevich, S. A. (1991) "High frequency modulation of quantum well heterostructure diode lasers by carrier heating in microwave electric field", *Proceedings of the 1991 International Semiconductor Device Research Symposium*, 403-406.

Belenky, G. L., Garbinski, P. A., Luryi, S., Mastrapasqua, M., Cho, A. Y., Hamm, R. A., Hayes, T. R., Laskowski, E. J., Sivco, D. L., and Smith, P. R. (1993) "Collector-up light-emitting charge injection transistors in n-InGaAs/InAlAs/p-InGaAs and n-InGaAs/InP/p-InGaAs heterostructures", *Journal of Applied Physics* **73**, 8618-8627.

Belenky, G. L., Garbinski, P. A., Luryi, S., Smith, P. R., Cho, A. Y., Hamm, R. A., and Sivco, D. L. (1994a) "Microwave performance of top-collector charge injection transistors on InP substrates", *Semiconductor Science and Technology* **9**, 1215-1219.

Belenky, G. L., Kastalsky, A., Luryi, S., Garbinski, P. A., Cho, A. Y., and Sivco, D. L. (1994b) "Measurement of the effective temperature of majority carriers under injection of hot minority carriers in heterostructures", *Appl. Phys. Lett.* **64**, pp. 2247-2249 (1994).

Belenky, G. L., Kazarinov, R. F., Lopata, J., Luryi, S., Tanbun-Ek, T., and Garbinski, P. A.

(1995) "Direct measurement of the carrier leakage out of the active region in InGaAsP/InP laser heterostructures", *IEEE Transactions on Electron Devices* **TED-42**, 215-218.

Bosch, B. G. and Engelmann, R. W. H. (1975) *Gunn Effect Electronics*, Academic Press, New York.

Chang, C.-Y., Liu, W. C., Jame, M. S., Wang, Y. H., Luryi, S., and Sze, S. M. (1986) "Induced base transistor fabricated by molecular beam epitaxy", *IEEE Electron Device Letters* **EDL-7**, 497-499.

Chen, Y.-K., Nottenburg, R. N., Panish, M. B., Hamm, R., and Humphrey, D. A. (1989) "Subpicosecond InP/InGaAs heterostructure bipolar transistor," *IEEE Electron Device Letters* **EDL-10**, 267-269

Coleman, P. D., Freeman, J., Morkoç, H., Hess, K., Streetman, B. G., and Keever, M., (1982) "Observation of a new oscillator based on real-space transfer in heterojunctions," *Applied Physics Letters* **40**, 493-495.

Conwell, E. M. (1967) *High Field Transport in Semiconductors*, Academic Press, New York.

Darcie, T. E., Tucker, R. S., and Sullivan, G. J. (1985) "Intermodulation and harmonic distortion in InGaAsP lasers", *Electronics Letters* **21**, pp. 665-666.

Döttling, R. and Schöll, E. (1994) "Domain formation in modulation-doped GaAs/$Al_x Ga_{1-x}$As heterostructures", *Solid-State Electronics* **37**, 685-688.

Dyakonov, M. I. and Shur, M. S. (1993) "Shallow water analogy for a ballistic field effect transistor and new mechanism of plasma wave generation by DC current", *Physical Review Letters* **71**, 2465-2467.

Dyakonov, M. I. and Shur, M. S. (1995a) "Choking of electron flow - a mechanism of current saturation in field effect transistors", *Physical Review* **B 51**, 14341-14345.

Dyakonov, M. I. and Shur, M. S. (1995b) "Two dimensional electronic flute", *Applied Physics Letters* **67**, 1137-1139.

Dyakonov, M. I. and Shur, M. S. (1996) "Detection, mixing, and frequency multiplication of terahertz radiation by two-dimensional electronic fluid", *IEEE Transactions on Electron Devices* **43**, 380-387.

Emanuel, M. A., Higman, T. K., Higman, J. M., Kolodzey, J. M., Coleman, J. J., and Hess, K. (1988) "Theoretical and experimental investigations of the heterostructure hot electron diode", *Solid-State Electronics* **31**, 589-592.

Faist, J., Capasso, F., Sivco, D. L., Sirtori, C., Hutchinson, A. L., and Cho, A. Y. (1994) "Quantum Cascade Laser," *Science* **264**, 553-556.

Faist, J., Capasso, F., Sirtori, C., Sivco, D. L., Hutchinson, A. L., and Cho, A. Y. (1995a) "Vertical transition quantum cascade laser with Bragg confined excited state", *Applied Physics Letters* **66**, 538-540.

Faist, J., Capasso, F., Sirtori, C., Sivco, D. L., Hutchinson, A. L., and Cho, A. Y. (1995b) "Continuous wave operation of a vertical transition quantum cascade laser above $T = 80\,\mathrm{K}$", *Applied Physics Letters* **67**, 3057-3059

Faist, J., Capasso, F., Sirtori, C., Sivco, D. L., Baillargeon, J. N., Hutchinson, A. L., Chu, S.-N. G., and Cho, A. Y. (1996) "High power mid-infrared ($\lambda \sim 5\,\mu\mathrm{m}$) quantum cascade lasers operating above room temperature", *Applied Physics Letters* **68**, 3680-3682.

Favaro, M. E., Fernández, G. E., Higman, T. K., York, P. K., Miller, L. M., and Coleman, J. J. (1989) "Strained layer InGaAs channel negative-resistance field-effect transistor" *Journal of Applied Physics* **65**, 378-380

Favaro, M. E., Miller, L. M., Bryan, R. P., Alwan, J. J., and Coleman, J. J. (1990a) "p-channel

negative resistance field-effect transistor", *Applied Physics Letters* **56**, 1058-1060.

Frohman-Bentchkowsky, D. (1971) "Memory behavior in a floating-gate avalanche injection MOS (FAMOS) structure", *Applied Physics Letters* **18**, 332-334.

Gelmont, B., Gorfinkel, V., and Luryi, S. (1996) "Theory of the spectral lineshape and gain in quantum wells with intersubband transitions", *Applied Physics Letters***68**, 2171-2173.

Gorfinkel, V. B. and Filatov, I. I. (1990) "Heating of an electron gas by an electric field in the active region of a semiconductor laser", *Fizika i Tekhnika Poluprovodnikov* **24**, 742 [English translation: *Soviet Physics - Semiconductors* **24**, 466 (1990)].

Gorfinkel, V. B., Gorbovitsky, B. M., and Filatov, I. I. (1991) "High frequency modulation of light output power in double-heterojunction laser", *International Journal of Infrared & Millimeter Waves* **12**, 649-658.

Gorfinkel, V. B. and Luryi, S. (1992) "Rapid modulation of interband optical properties of quantum wells by intersubband absorption", *Applied Physics Letters* **60**, 3141-3143.

Gorfinkel, V. B. and Luryi, S. (1993) "High-Frequency Modulation and Suppression of Chirp in Semiconductor Lasers", *Applied Physics Letters* **62**, 2923-2925.

Gorfinkel, V. B. and Luryi, S. (1994) "Dual modulation of semiconductor lasers", in *Physics and Simulation of Optoelectronic Devices II*, ed. by M. Osinski, *Proceedings of SPIE* **2146**, 204-209; US patent **5,311,526**.

Gorfinkel, V. B. and Luryi, S. (1995) "Fundamental limits for linearity of CATV lasers", *IEEE Journal of Lightwave Technology* **13**, pp. 252-260.

Gorfinkel, V., Luryi, S., and Gelmont, B. (1996) "Theory of gain spectra for quantum cascade lasers and temperature dependence of their characteristics at low and moderate carrier concentrations", *IEEE Journal of Quantum Electronics* **32**, No. 11.

Gribnikov, Z. S., (1972) "Negative differential conductivity in a multilayer heterostructure," *Fizika i Tekhnika Poluprovodnikov* **6**, 1380-1382 [English translation: *Soviet Physics - Semiconductors* **6**, 1204-1205 (1973)].

Gribnikov, Z. S., Hess, K., and Kosinovsky, G. A., (1995) "Nonlocal and nonlinear tansport in semiconductors: real space transfer effects", *Journal of Applied Physics* **77**, 1337-1372.

Grinberg, A. A., Luryi, S., Schryer, N. L., Smith, R. K., Lee, C., Ravaioli, U., and Sangiorgi, E. (1991) "Adiabatic approach to the dynamics of non-equilibrium electron ensembles in semiconductors", *Physical Review* **B 44**, 10536-10545.

Grinberg, A. A. and Luryi, S. (1990) "Nonstationary quasiperiodic energy distribution of an electron gas upon ultrafast thermal excitation", *Physical Review Letters* **65**, 1251-1254.

Grinberg, A. A. and Luryi, S. (1992) "Ballistic versus diffusive base transport in the high-frequency characteristics of bipolar transistors", *Applied Physics Letters* **60**, 2770-2772.

Grinberg, A. A. and Luryi, S. (1993) "Coherent transistor", *IEEE Transactions on Electron Devices* **ED-40**, 1512-1522.

Gunn, J. B. (1963) "Microwave oscillations of current in III-V semiconductors", *Solid State Communications* **1**, 88-91.

Hall, K. L., Lenz, G., Ippen, E. P., Koren, U., and Raybon, G. (1992) "Carrier heating and spectral hole burning in strained-layer quantum-well laser amplifiers at 1.5μm", *Applied Physics Letters* **61**, 2512-2514.

Heiblum, M. (1981) "Tunneling hot electron transfer amplifiers (THETA): amplifiers operating up to the infrared," *Solid-State Electronics* **24**, 343-366

Heiblum, M., Nathan, M. I., Thomas, D. C., and Knoedler, C. M. (1985) "Direct Observation of Ballistic Transport in GaAs," *Physical Review Letters* **55**, 2200-2203.

Henry, C. H., Logan, R. A., Merritt, F. R., and Luongo, J. P. (1983) "The effect of intervalence band absorption on the thermal behavior of InGaAsP lasers", *IEEE Journal of Quantum Electronics* **QE-19**, 947-952.

Hess, K., Morkoç, H., Shichijo, H., and Streetman, B. G. (1979) "Negative differential resistance through real-space electron transfer," *Applied Physics Letters* **35**, 469-471.

Hess, K., Higman, T. K., Emanuel, M. A., and Coleman, J. J. (1986) "New ultrafast switching mechanism in semiconductor heterostructures", *Journal of Applied Physics* **60**, 3775-

Hesto, P., Pone, J.-F., and Castagne, R. (1982) "A proposal and numerical simulation of N^+NN^+ Schottky device for ballistic and quasiballistic electron spectroscopy", *Applied Physics Letters* **40**, 405-406.

Hilsum, C. (1962) "Transferred electron amplifiers and oscillators", *Proceedings of the IRE* **50**, 185-189.

Hu, C., Editor (1991) *Nonvolatile semiconductor memories: technologies, design, and applications*, IEEE Press, New York.

Hueschen, M. R., Moll, N., and Fischer-Colbrie, A. (1990) "Improved microwave performance in transistors based on real-space electron transfer", *Applied Physics Letters* **57**, 386-388.

Jalali, B., Nottenburg, R. N., Chen, Y.-K., Levi, A. F. J., Sivco, D. L., Cho, A. Y., and Humphrey, D. A. (1989) "Near-ideal lateral scaling in abrupt *AIA/IGA* heterostructure bipolar transistors prepared by molecular beam epitaxy," *Applied Physics Letters* **54**, 2333-2335.

Kastalsky, A. and Luryi, S. (1983) "Novel Real-Space Hot-Electron Transfer Devices", *IEEE Electron Device Letters* **EDL-4**, 334-336.

Kastalsky, A., Luryi, S., Gossard, A. C., and Hendel, R. H. (1984) "A field-effect transistor with a negative differential resistance", *IEEE Electron Device Letters* **EDL-5**, 57-60.

Keever, M., Shichijo, K., Hess, K., Banerjee, S., Witkowski, L., Morkoç, H., and Streetman, B. G. (1981) "Measurements of hot-electron conduction and real-space transfer in GaAs/AlGaAs heterojunction layers," *Applied Physics Letters* **38**, 36-38.

Kizillyalli, I. C. and Hess, K.(1989) "Physics of Real-Space Transfer Transistors", *Journal of Applied Physics* **65**, 2005-2013.

Koscica, T. E. and Zhao, J. H. (1995a) "Field effect real space transfer transistor", *IEEE Electron Device Letters* **EDL-16**, 196-198.

Koscica, T. E. and Zhao, J. H. (1995b) "Frequency doubling in GaAs/AlGaAs field effect transistor using real space transfer", *IEEE Electron Device Letters* **EDL-16**, 545-547.

Kroemer, H. (1957) "Theory of a wide-gap emitter for transistors," *Proceedings of the IRE* **45**, 1535-1537.

Kroemer, H. (1964) "Theory of the Gunn effect", *Proceedings of the IEEE* **52**, 1736.

Kroemer, H. (1982) "Heterostructure bipolar transistors and integrated circuits," *Proceedings of the IEEE* **70**, 13-25.

Kumekov, S. E. and Perel, V. I. (1988) "Energy relaxation of the electron-phonon system of a semiconductor under static and dynamic conditions", *Zhurnal Eksperimental′noi i Teoreticheskoi Fiziki* **94**, 346-356 [English translation: *Soviet Physics - JETP* **67**, 193-198 (1988)].

Lai, J-T. and Lee, J. Y. (1994) "Ultrahigh and controllable drain current peak-to-valley ratio in negative resistance field-effect transistors with a strained InGaAs channel", *IEEE Electron Device Letters* **15**, 333-335.

Lai, J-T. and Lee, J. Y. (1995a) "Enhanced real-space electron transfer in charge injection transistors with source-channel heterojunctions formed by graded $Al_xGa_{1-x}As$ layer and

shallow Pd/Ge ohmic contacts", *Applied Physics Letters* **66**, 1779-1781.

Lai, J-T. and Lee, J. Y. (1995b) "Enhancement of electron transfer and negative differential resistance in GaAs-based real-space transfer devices by using strained InGaAs channel layers", *Applied Physics Letters* **66**, 1965-1967.

Lai, J-T., Yeh, Y.-H., and Lee, J. Y. (1996) "Light emitting real-space transfer devices fabricated with strained GaAs/$In_{0.2}Ga_{0.8}As$/AlGaAs Heterostructures", *Electronics Letters* **32**, 1041-1042.

Landau, L. D. and Lifshitz, E. M. (1980) *Statistical Physics*, Vol. 1 3rd edn., Pergamon, Oxford.

Levi, A. F. J., Jalali, B., Nottenburg, R. N., and Cho, A. Y. (1992) "Vertical scaling in heterojunction bipolar transistors with nonequilibrium base transport", *Applied Physics Letters* **60**, 460-462.

Lindmayer, J. (1964) "The metal-gate transistor," *Proceedings of the IEEE* **52**, 1751.

Lobentanzer, H., Stolz, W., Ploog, K., Bäuerle, R. J., and Elsaesser, T. (1989) "Screening of the N=2 excitonic resonance by hot carriers in an undoped GaInAs/AlInAs multiple quantum well structure", *Solid-State Electronics* **32**, 1875-1879.

Luryi, S. (1985) "An induced base hot electron transistor", *IEEE Electron Device Letters* **EDL-6**, 178-180.

Luryi, S. (1990a) "Charge injection transistors and logic circuits", *Superlattices and Microstructures* **8**, 395-404.

Luryi, S. (1990b) "Hot-Electron Transistors", Chap. 7 in *High-speed semiconductor devices*, ed. by S. M. Sze, Wiley Interscience, pp. 399-461.

Luryi, S. (1991) "Light emitting devices based on the real-space-transfer of hot electrons", *Applied Physics Letters* **58**, 1727-1729.

Luryi, S. (1994) "How to make an ideal HBT and sell it too", *IEEE Transactions on Electron Devices* **TED-41**, 2241-2247.

Luryi, S. (1996) "Active packaging: a new fabrication principle for high performance devices and systems", in *Future of Microelectronics: Reflections on the Road to Nanotechnology*, ed. by S. Luryi, J. M. Xu, and A. Zaslavsky, NATO ASI Series **E 323**, pp. 35-43, Kluwer Academic, Dordrecht.

Luryi, S., Kastalsky, A., Gossard, A. C., and Hendel, R. H. (1984a) "Charge injection transistor based on real-space hot-electron transfer", *IEEE Transactions on Electron Devices* **ED-31**, 832-839.

Luryi, S., Kastalsky, A., Gossard, A. C., and Hendel, R. H. (1984b) "Hot electron memory effect in double-layered heterostructures", *Applied Physics Letters* **45**, 1294-1296.

Luryi, S., Mensz, P. M., Pinto, M. R., Garbinski, P. A., Cho, A. Y., Sivco, D. L. (1990) "Charge injection logic", *Applied Physics Letters* **57**, 1787-1789.

Luryi, S. and Pinto, M. R. (1991a) "Broken symmetry and the formation of hot-electron domains in real-space transfer transistors", *Physical Review Letters* **67**, 2351-2354.

Luryi, S. and Pinto, M. R. (1991b) "Logic element and article comprising the element", US Pat. **4,999,687**.

Luryi, S. and Pinto, M. R. (1992) "Symmetry of the real-space transfer and collector-controlled states in charge injection transistors", *Semiconductor Science and Technology* **7**, B520-B526.

Luryi, S. and Mastrapasqua, M. (1993) "Light-emitting logic devices based on real space transfer in complementary InGaAs/InAlAs heterostructures", in "*Negative Differential*

Resistance and Instabilities in 2D Semiconductors", ed. by N. Balkan, B. K. Ridley, and A. J. Vickers, NATO ASI Series [Physics] **B 307**, pp. 53-82, Plenum Press, New York.

Luryi, S., Grinberg, A. A., and Gorfinkel, V. B. (1993) "Heterostructure bipolar transistor with enhanced forward diffusion of minority carriers", *Applied Physics Letters* **63**, 1537-1539.

Maezawa, K. and Mizutani, T. (1991) "High-frequency characteristics of charge-injection transistor-mode operation in AlGaAs/InGaAs/GaAs metal-insulator-semiconductor field-effect transistors", *Japanese Journal of Applied Physics* **30**, 1190-1193.

Malik, R. J., Hollis, M. A., Eastman, L. F., Wood, C. E. C., Woodard, D. W., and AuCoin, T. R. (1981) "GaAs planar-doped barrier transistors grown by molecular beam epitaxy," *Proceedings of the 8th Biennial Cornell Conference on Active Microwave Semiconductor Devices and Circuits*, IEEE/Cornell.

Mastrapasqua, M., Capasso, F., Luryi, S., Hutchinson, A. L., Sivco, D. L., and Cho, A. Y. (1992) "Light emitting charge injection transistor with p-type collector", *Applied Physics Letters* **60**, 2415-2417.

Mastrapasqua, M., Luryi, S., Belenky, G. L., Garbinski, P. A., Cho, A. Y., and Sivco, D. L. (1993) "Multi-terminal light emitting logic device electrically reprogrammable between OR and NAND functions", *IEEE Transactions on Electron Devices* **ED-40**, 1371-1377.

Mastrapasqua, M., King, C. A., Smith, P. R., and Pinto, M. R. (1994) "Charge injection transistors and logic elements in $Si/Si_{1-x}Ge_x$ heterostructures", 1994-IEDM *Technical Digest*, 385-388.

Mastrapasqua, M., King, C. A., Smith, P. R., and Pinto, M. R. (1996) "Charge injection transistors and logic elements in $Si/Si_{1-x}Ge_x$ heterostructures", in *Future of Microelectronics: Reflections on the Road to Nanotechnology*, ed. by S. Luryi, J. M. Xu, and A. Zaslavsky, NATO ASI Series **E 323**, pp. 377-383, Kluwer Academic, Dordrecht.

Mead, C. A. (1960) "The tunnel-emission amplifier", *Proceedings of the IRE* **48**, 359-361.

Mensz, P. M., Luryi, S., Cho, A. Y., Sivco, D. L., and Ren, F. (1990a) "Real Space Transfer in Three-Terminal InGaAs/InAlAs Heterostructure Devices", *Applied Physics Letters* **56**, 2563-2565.

Mensz, P. M., Garbinski, P. A., Cho, A. Y., Sivco, D. L., and Luryi, S. (1990b) "High Transconductance and Large Peak-To-Valley Ratio of Negative Differential Conductance in Three-Terminal InGaAs/InAlAs Real-Space Transfer Devices", *Applied Physics Letters* **57**, 2558-2560.

Mensz, P. M., Schumacher, H., Garbinski, P. A., Cho, A. Y., Sivco, D. L., and Luryi, S. (1990c) "Microwave operation of InGaAs/InAlAs charge injection transistors", 1990-IEDM *Technical Digest*, 323-326.

Mensz, P. M., Luryi, S., Bean, J. C., and Buescher, C. J. (1990d) "Evidence for a real-space transfer of hot holes in strained GeSi/Si heterostructures", *Applied Physics Letters* **56**, 2663-2665.

Noda, S., Uemura, T., Yamashita, T., and Sasaki, A. (1990) "All-optical modulation using an n-doped quantum-well structure", *Journal of Applied Physics* **68**, 6529-6531.

Pinto, M. R. (1991) "Simulation of ULSI Device Effects", in *ULSI Science and Technology*, J. Andrews and G. K. Celler, eds., *Electrochemical Society Proc.* **91-11**

Pinto, M. R. and Luryi, S. (1991) "Simulation of multiply connected current-voltage characteristics in charge injection transistors", 1991-IEDM *Technical Digest*, 507-510.

Ridley, B. K. (1963) "Specific negative resistance in solids", *Proceedings of the Physical Society* (London) **82**, 954-966.

Ridley, B. K. and Watkins, T. B. (1961) "The possibility of negative resistance effects in

semiconductors", *Proceedings of the Physical Society* (London) **78**, 233-235.

Ritter, D., Hamm, R. A., Feygenson, A., Panish, M. B., and Chandrasekhar, S. (1991) *Applied Physics Letters* **59**, 3431-3433.

Rivlin, L. A. (1985) "‘‘Febrile’’ reaction of electrons in a semiconductor laser to an ultrashort pulse", *Kvantovaya Elektronika* **12**, 689-693 [English translation: *Soviet Journal of Quantum Electronics* **15**, 453-456 (1985)].

Ruch, J. G. (1972) "Electron dynamics in short channel filed-effect transistors", *IEEE Transactions on Electron Devices* **ED-19**, 652-654.

Sai-Halasz, G. A., Wordeman, M. R., Kern, D. P., Rishton, S., and Ganin, G. (1988) "High transconductance and velocity overshoot in NMOS devices at the 0.1 μm gate-length level", *IEEE Electron Device Letters* **EDL-9**, 464-466.

Schöll, E. and Aoki, K. (1991) "Novel mechanism of a real-space transfer oscillator", *Applied Physics Letters* **58**, 1277-1279.

Shockley, W. (1951) US Patent 2,569,347 (filed 1948).

Shockley, W. (1954) "Negative resistance arising from transit time in semiconductor diodes", *Bell System Technical Journal* **33**, 799-826.

Shur, M. S. (1987) *GaAs devices and circuits*, Plenum Publishing, New York.

Shur, M. S. (1990) *Physics of semiconductor devices*, Prentice Hall, Englewood Cliffs.

Stiles, M. D. and Hamann, D. R. (1989) "Electron transmission through $NiSi_2$-Si interfaces," *Physical Review* **B 40**, 1349-1352.

Sze, S. M. (1981) *Physics of semiconductor devices*, Wiley Interscience, New York.

Sze, S. M. (1990) "Microwave diodes", Chap. 9 in *High-Speed Semiconductor Devices*, ed. by S. M. Sze, Wiley Interscience, New York.

Tian, H., Kim, K. W., and Littlejohn, M. A. (1992) "Novel heterojunction real-space transfer logic transistor structures: a model-based investigation", *IEEE Transactions on Electron Devices* **ED-39**, 2189-2196.

Tian, H., Kim, K. W., and Littlejohn, M. A. (1993) "Novel charge injection transistors with heterojunction source (launcher) and drain (blocker) configurations", *Applied Physics Letters* **63**, 174-176.

Tung, R. T., Levi, A. F. J., and Gibson, J. M. (1986) "Control of a natural permeable base transistor," *Applied Physics Letters* **48**, 635-637.

Tolstikhin, V. I. and Mastrapasqua, M. (1995) "Three-terminal laser structure for high-speed modulation using dynamic carrier heating", *Applied Physics Letters* **67**, 3868-3870.

Tsai, C. Y., Eastman, L. F., and Lo, Y. H. (1993) "Hot carrier and hot phonon effects on high-speed quantum well lasers", *Applied Physics Letters* **63**, 3408-3410.

Wacker, A. and Schöll, E. (1994) "Spiking at vertical electrical transport in a heterostructure device", *Semiconductor Science and Technology* **9**, 592-594.

Westland, D. J., Ryan, J. F., Scott, M. D., Davies, J. I., and Riffat, J. R. (1988) "Hot carrier energy loss rates in GaIAs/InP quantum wells", *Solid-State Electronics* **31**, 431-438.

Wu, C.-L., Hsu, W.-C., Tsai, M.-S., and Shieh, H.-M. (1995a) "Very strong negative differential resistance real space transfer transistor using a multiple δ-doping GaAs/InGaAs pseudomorphic heterostructure", *Applied Physics Letters* **66**, 739-741.

Wu, C.-L., Hsu, W.-C., Shieh, H.-M., and Tsai, M.-S. (1995b) "A novel δ-doped GaAs/InGaAs real-space transfer transistor with high peak-to-valley ratio and high current driving capability",

IEEE Electron Device Letters **16**, 112-114.

Yariv, A. (1991) *Optical Electronics*, 4th Ed., Saunders College Publishing, Philadelphia.

Figure captions

Figure 1. Ballistic motion of electrons (a) accelerated by electric field; (b) thermionically injected from a wide-gap material; (c) injected by tunneling into a state of kinetic energy.

Figure 2. Conduction band profile of a heterostructure for ballistic electron spectroscopy (after Heiblum et al. 1985).

Figure 3. Electron energy distributions $(EE)^{1/2} f_{\rm m}$ (solid lines) corresponding to the mesoscopic functions $f_{\rm m}(EE)$ that evolve from the non-equilibrium Maxwellian ensembles characterized by an initial temperature $T_{\rm i}$ (after Grinberg et al. 1991). The illustrated case corresponds to the lattice temperature $T = 300\,{\rm K}$. The equilibrium distributions are indicated by stippled lines and the initial distributions by dashed lines.
(a) $T_{\rm i} = 75\,{\rm K}$; (b) $T_{\rm i} = 1000\,{\rm K}$.

Figure 4. Illustration of the Hess diode, its energy band diagram and current-voltage characteristics. Labels 1 and 2 indicate a low-current, high-resistance state and a high-current, low-resistance state, respectively.

Figure 5. Calculated current-voltage characteristics of a multilayer real-space-transfer diode. Solid line displays $v(F)$ for parameters as in the legend. The dashed line corresponds to the same set of parameters, but a different effective mass ratio: $m_1/m_2 = 0.3$

Figure 6. Band diagram of a modulation-doped real-space transfer diode (after Hess et al., 1979). At equilibrium electrons reside in narrow-gap layers 1; heated by an electric field they transfer into wide-gap layers 2 of lower mobility.

Figure 7. Schematic diagram of a charge injection transistor. Emitter electrons heated up by the field applied between electrodes S and and D undergo real-space transfer, as indicated by the arrow. By symmetry, the output current $I_{\rm C}$ at a fixed collector bias $V_{\rm C}$ is an exclusive-OR function of the input voltages: $I_{\rm C} = \mathbf{xor}(VS, V_{\rm D})$.

Figure 8. Room temperature characteristics of a CHINT/NERFET device with $W = 25\,\mu\text{m}$ and $Lch = 1\,\mu\text{m}$ (after Mensz et al. 1990b). The drain current (I_D) and collector current (I_C) are plotted versus the heating drain voltage V_D at a fixed collector bias $V_C = 3.9\,\text{V}$. Insert shows the collector-bias dependence of the peak-to-valley ratio and the leakage current, defined as the magnitude of I_D at $V_D = 0$.

Figure 9. Principle of the multi-terminal logic device ORNAND. Three input terminals arranged with a 3-fold cyclic symmetry (top figure) define three channels 1-2, 2-3, and 3-1. The RST current I_C as a function of the voltages $\mathbf{V}_1$, $\mathbf{V}_2$, and $\mathbf{V}_3$, regarded as logic signals, obeys the truth table shown below. The value of I_C is low (logic-0) in two states when $\mathbf{V}_1 = \mathbf{V}_2 = \mathbf{V}_3$ and is high (logic-1) in the other six states. By the symmetry, all the logic-1 states give the same I_C.

Figure 10. The ORNAND gate (after Mastrapasqua et al. 1993).

(a) Optical and electrical logic operation obtained in a quasi-stationary measurement at room temperature for $V_C = 2.4\,\text{V}$. Electrodes 3 and $\tilde{3}$ are tied together. The binary values "logic-0" and "logic-1" of the input signals $\mathbf{V}_1$, $\mathbf{V}_2$, and $\mathbf{V}_3$ correspond to 0 and $3\,\text{V}$, respectively. The particular grouping of the states into OR and NAND reflects the choice of $\mathbf{V}_3$ as the "control" electrode.

(b) Cross-section of the device structure. Cyclic symmetry of an ORNAND gate results from the periodic boundary condition obtained by tying together electrodes 3 and $\tilde{3}$.

Figure 11. Metal Base Transistors: a) MOMOM; b) MOMS; c) SMS. Each of the M (metal) or S (doped-semiconductor) electrodes is contacted independently and can be biased with respect to the other electrodes. The figure shows the energy-band diagrams of metal-base transistors under operating bias conditions.

Figure 12. Unipolar ballistic transistors with a monolithic all-semiconductor structure.

a) tunneling hot-electron transfer amplifier (THETA), a tunnel-emitter transistor (after

Heiblum, 1981);

b) planar-doped barrier (PDB) transistor (after Malik et al. 1981);

c) induced-base (IBT) transistor (after Luryi, 1985).

Figure 13. Schematic band diagram of an abrupt-junction heterostructure bipolar transistor (after Kroemer, 1981). Minority carriers are injected into the base "over a cliff" of energy Δ.

Figure 14. Intrinsic frequency characteristics of a coherent transistor (after Grinberg and Luryi, 1993).

(a) The common-emitter current gain. Solid line shows the calculated $\beta^2(\omega) \equiv |h_{21}^{e}|^2$, and the dashed lines indicate the gain roll-off in the conventional frequency range ($f < f_T$) and the extended range ($f > f_T$);

(b) Frequency dependences of the unilateral gain $|U|$, the output resistance $r_{22} \equiv \mathcal{R}e(z_{22}^{e})$, and the common-emitter current gain $\beta \equiv |h_{21}^{e}|$, assuming $\Delta = 10kT$, the collector transit time $\tau_C = 1$ ps, and base transit time $\tau_B = 2$ ps.

Figure 15. Common-emitter current gain $|h_{21}|$ and the unilateral power gain $|U|$ of a model coherent transistor with a special graded-gap base design, optimized for stable oscillation at 94 GHz. Base total width $W = 1\,\mu$m. Transistor is assumed loaded with the parasitics with state-of-the art equivalent circuit parameters, e.g. $C_{Cx} = 2C_C$. Conventional current-gain cutoff is $f_T \approx 32$ GHz, however the transistor also exhibits a range of current gain $|h_{21}| > 1$ at $f \approx 2\pi f_T$ (near the second peak in U). The fundamental peak in U occurs near πf_T.

Figure 16. Schematic diagram of a double-heterostructure laser. Electron injection into the narrow-gap active layer is partly by tunneling; holes which are heavier are mostly injected by thermionic emission.

Figure 17. Schematic diagram of the quantum cascade laser structure and the kinetics of carrier transport. To ensure identical conditions, each cascade period must be neutral. The

mobile charge in the undoped quantum well is compensated by the positive donor charge in the reservoir. The rates of (nonradiative) transitions between the subbands and the escape rate from the lower subband into the reservoir are characterized by time constants τ_{21} and $\tau_{1\,\mathrm{out}}$, respectively.

Figure 18. Graphical solution of the threshold equation (29) for a parabolic model of the quantum subbands. Assuming a concave upward curve $\hbar\gamma_0(T_e)$ and $T < T_{\max}$, the lasing range is defined by $T_{\mathrm{th}}^{(1)} < T_e(J) < T_{\mathrm{th}}^{(2)}$. In view of Eq. (26), the two thresholds [$T_{\mathrm{th}}^{(1)}$ and $T_{\mathrm{th}}^{(2)}$] in the effective temperature T_e determine both the lower and the upper thresholds for the pump current.

Figure 19. Frequency dependence of the optical response $\delta S(f)/\delta S(0)$ to the variation of different parameters in a stripe multiple quantum well laser (after Gorfinkel and Luryi, 1993). Curve 1: modulation by the pumping current, curve 2: modulation by the electron temperature, curve 3: dual modulation as in Eq. (37).