8.4 Quantum Well Tunneling Devices

S. Luryi and A. Zaslavsky

A INTRODUCTION

Advances in epitaxy have permitted the design of semiconductor devices with sufficient precision to make use of quantum mechanical tunneling. The potential advantages of tunneling-based devices consist of high-speed operation and higher functionality inherent in their strongly nonlinear current-voltage characteristics, while the disadvantages include the relative weakness of quantum effects at room temperature and the difficult integration into mainstream semiconductor technology. To date, these difficulties have largely impeded tunneling devices from evolving beyond the research stage, but continuing progress in epitaxy, fabrication and novel device concepts may render these devices suitable for some applications. Section B of this Datareview will cover the extensively researched double-barrier resonant tunneling structure (RTS) as a discrete device, as well as a two-terminal nonlinear component in simple circuits, and discuss the prospects of three-terminal operation. Optical sources and detectors based on intersubband transitions in quantum wells populated or emptied by tunneling are discussed in Section C. Tunneling devices sufficiently small to employ single-carrier or Coulomb blockade effects that offer distant prospects of general-purpose single-electron logic or more immediate possibilities of high-precision current sources are discussed in Section D. Finally, expanding the definition of a device to include tunneling structures that provide experimental probes of the electronic and structural properties of quantum wells, we discuss double-barrier and double-well RTS as spectroscopic systems for subband dispersions, carrier coherence lifetimes, and strain distributions.
B RESONANT TUNNELING DEVICES

B1 Two-terminal double-barrier resonant tunneling diodes

A schematic band diagram of a double-barrier RTS in the tunneling regime, the device itself, and the corresponding nonlinear current-voltage $I(V)$ characteristic are shown in FIGURE 1. Originally proposed [1] and demonstrated [2] in the 1970’s, this general structure has attracted much scientific interest. As shown in FIGURE 1 (a), confinement of carriers in the well quantizes their electronic states into two-dimensional (2D) subbands $E_i$ and transport between the doped electrodes cladding the double-barrier regions proceeds mainly by tunneling (as long as the temperature is low or the barriers high enough to suppress thermally activated above-barrier transport). Originally tunneling was analyzed in terms of coherent quantum-mechanical transmission through the double-barrier potential [1-3], which predicts very sharp resonances when the incident energy matches $E_i$. A more realistic description in the presence of scattering is the sequential tunneling model [4]: carriers tunnel from the emitter into a 2D subband conserving energy $E$ and transverse momentum $k_{\perp}$, then lose coherence and eventually tunnel out to the collector in a separate tunneling event. The $E$ and $k_{\perp}$ conservation rules determine the number of carriers that can tunnel into the well for a given alignment, set by the applied bias $V$, of the subband $E_i$ and the emitter Fermi level $E_F$: current can flow only when $E_F > E_i$ ($E$ conservation, leading to a threshold bias $V_{th}$) and $E_i > E_c$ ($k_{\perp}$ conservation, leading to a peak bias $V_p$). The sequential tunneling model [4] describes very well the experimental low-temperature $I(V)$ of high-quality GaAs/AlGaAs RTS shown in FIGURE 1 (c) [5], except for the non-zero valley current for $V > V_p$ which requires explicit evaluation of phonon and impurity scattering-assisted tunneling that need not conserve $E$ and $k_{\perp}$.

The double-barrier RTS is a two-terminal device with a highly nonlinear $I(V)$ characteristic. The strong negative differential resistance (NDR) region for $V > V_p$, with
the peak-to-valley ratio of the current reaching more than 50 even at room temperature in appropriately designed RTS [6], is useful in the design of solid-state oscillators. The principle of operation is analogous to the classic $p$-$n$ junction-based tunnel diode [7] and the maximum oscillation frequency is limited by the $RC$ time constant. In a tunnel diode both sides of the junction are doped heavily and the tunneling barrier is set by the bandgap, while optimized RTS can have tailored barriers leading to very high peak current densities (lower $R$) without high doping on both sides of the tunnel barriers (leading to lower $C$). Microwave oscillators operating at frequencies up to 420 GHz in InGaAs/AlAs/InAs and 712 GHz in InAs/AlSb RTS diodes have been fabricated [8], as well as circuit-compatible InAs/AlSb RTS diodes with 1.7 ps switching speeds [9].

**B2 Three-terminal resonant tunneling structures**

As a discrete two-terminal device the double-barrier RTS has limited utility beyond solid-state oscillators. There has been considerable research into the fabrication of three-terminal RTS devices, either by separately contacting the well of a double-barrier RTS or by adding a gate electrode. The former approach is analogous to a heterojunction bipolar transistor: the schematic band diagram of such a device is shown in FIGURE 2, which combines the bipolar [10] and unipolar [11] versions. The bipolar version permits the separate contacting of the narrow well by doping it $p$-type. Negative transconductance is observed in the collector current $I_c$ characteristic whenever the hole current in the quantum well biases the RTS into the NDR region [10]. However, the narrow quantum well necessary for large 2D subband separation leading to strong NDR also results in high base resistance unless the well is very heavily doped, but then scattering tends to wash out the tunneling characteristics. This difficulty might be overcome in polytype GaSb/AlSb/InAs RTS where bandgap blocking of the tunneling current gives excellent peak-to-valley ratios even for wide quantum wells [12]. Structures with wider $n$-type quantum wells have also been fabricated into
unipolar three-terminal tunneling hot-electron transistors, where the injected electrons traverse the well ballistically and contribute to the collector current [11]. The current gain in these devices depends on the fraction of the ballistic electrons reaching the collector rather than scattering into the quantum well; it is typically smaller than in bipolar transistors. An interesting variation on this scheme is the lateral tunneling transistor [13], where the two barriers are produced by narrow electrostatic gates deposited on top of a modulation-doped heterostructure with a 2D electron gas [14]. The high mobility of the 2D electron gas greatly enhances the fraction of ballistic electrons reaching the collector and reduces the base resistance, leading to current gains of over 100, but only at cryogenic temperatures [13].

An alternative route to a three-terminal RTS device is the fabrication of a gated structure. FIGURE 3 illustrates the cross-section and schematic band diagram of a 2D RTS, where a gate controls the density of the 2D electrons that tunnel through 1D subbands in the quantum well [15]. Negative transconductance was predicted for this device since the fringing electric fields generated by gate bias lower the subband energies $E_i$ with respect to the emitter $E_F$, shifting the current peaks towards lower drain bias values. Large-scale fabrication of such devices requires the regrowth of modulation-doping heterostructures on etched surfaces, but the negative transconductance characteristic has been confirmed [16] in a similar structure produced by cleaved-edge overgrowth [17]. The same functionality can in principle be obtained from a lateral tunneling transistor [13] if the electrostatically confined quantum well can be made sufficiently narrow for well-resolved subband quantization. Finally, narrow vertical double-barrier RTS can be controlled laterally by a Schottky gate [18], although this approach involves great fabrication complexity. It should be stressed that while it has been suggested that three-terminal RTS devices with negative transconductance can in principle perform complementary functions for logic applications [19], no RTS circuit analogous to CMOS inverter has ever been demonstrated. The difficulty lies in the fact
that complementary CMOS transistors have more than just a negative transconductance, their current flow is effected by carriers of opposite polarity, which makes it possible to connect the drains of two transistors in series, rather than source of one to drain of the other.

**B3 Integration of resonant tunneling diodes in devices and circuits**

The highly nonlinear $I(V)$ of RTS makes it a potentially useful circuit element when inserted in another device or integrated with other devices. Much research has been done on resonant hot-electron transistors (RHET's) where a double-barrier RTS is inserted into the emitter of a hot-electron transistor as shown in FIGURE 4 [20]. The RHET exhibits negative transconductance together with current gain, as the $I(V)$ characteristic of the emitter RTS is reproduced in the collector current. By adding two logic inputs at the base the RHET provides exclusive NOR functionality in a single device [20]. Various logic gates and circuits have been fabricated by combining RHET's and resistors with reduced transistor counts compared to standard bipolar or CMOS circuits. However, the reduced transistor count in such circuits as a full adder [21] typically comes at the expense of fabricating a large number of resistors.

The epitaxy of several decoupled double-barrier RTS separated by wide doped regions can produce a two-terminal device with a multipeak $I(V)$ characteristic as each of the constituent RTS is biased through the resonance similar to FIGURE 1 (c). When such a structure is inserted in the emitter of a bipolar transistor, the multipeak collector current provides frequency multiplication with current gain [22]. In an optimized structure of decoupled RTS in series, a regular sawtooth $I(V)$ characteristic with a large number of resonant peaks can be produced, as shown in FIGURE 5. When biased with a constant current from a field effect transistor, such a device has a number of stable voltage operating points that can be used for multistate memory [23]: after an input voltage value is applied, the circuit moves to the nearest stable point and maintains that
value indefinitely.

A variant of static random access memory can be constructed by connecting two double-barrier RTS in series with an additional connection to the middle node, either by fabricating a double-emitter RHET with a common floating base [24] or by fabricating two RTS with a common collector on top of a single-barrier tunnel diode [25]. When two devices with NDR $I(V)$ characteristics are connected in series and the total applied bias exceeds twice the peak voltage $V_p$, the middle node becomes bistable. Most of the total applied bias drops over one of the devices, while the other takes less than $V_p$ and current continuity is preserved. The potential of the middle node (for example, the floating common base of a double-emitter RHET device) can be flipped between the two stable points by the biasing of a third terminal (the collector). As a result, very compact static memory cells consisting of two vertical RTS devices and associated contact lines have been fabricated [24, 25]. The difficulty with these RTS-based memories is their relatively large standby power consumption. The bistable operating points of two RTS in series biased just beyond $2V_p$ pass a current that is at least as large as the valley current value in the $I(V)$ characteristic of a single RTS. Yet the valley current cannot be reduced indefinitely by reducing RTS device size because in large-scale memory circuits sizable peak resonant currents are required to charge up the interconnect capacitances. Consequently, significant improvements in the peak-to-valley ratios of the RTS elements will be required before these tunneling-based static memories become competitive with low-power CMOS.

C O P T O E L E C T R O N I C T U N N E L I N G D E V I C E S

If many identical double-barrier potentials are repeated epitaxially, the result is a superlattice (SL) in which the quantized subbands broaden into minibands of width $\Delta E_i$ separated by minigaps. Consider transport in the presence of an electric field $F$ applied along the SL. If $F$ is weak, such that the potential drop $eFd \ll \Delta E$, where $d$ is the SL
period, current will flow by miniband conduction. Given sufficiently long scattering times $\tau$, the electric field would accelerate carriers into the region of negative curvature in the miniband dispersion, giving rise to NDR in the $I(V)$ characteristic [26]. For even greater $\tau$, the carriers would reach the minizone boundary and experience Bragg reflection, giving rise to Bloch oscillations [26, 27]. These effects have proven very difficult to observe in transport because of scattering, Zener tunneling between different minibands, and especially the charge-driven break-up of the superlattice into high and low-field domains [28]. For this reason, despite much research into the physics of low-field superlattice transport [29], SL $I(V)$ nonlinearities have not been used in devices to date.

In the strong $F$ limit, where $eF d > \Delta E$, the miniband breaks up into 2D subbands localized within one well and a resonant tunneling current flows when the ground subband in one well is exactly aligned with the first (or higher) excited subband in the adjacent well: $eF d = (E_i - E_0)$, $i = 1, 2 ...$ [30, 31]. As in ordinary double-barrier resonant tunneling, this energy conservation rule can be altered by an inelastic process, such as phonon or photon emission. In particular, when $eF d > (E_i - E_0)$ a carrier can tunnel into the adjacent well by emitting a photon of energy $\hbar \omega = eF d - (E_i - E_0)$, implying the exciting possibility of an infrared laser tunable by an applied electric field, an idea dating back to 1971 [30]. A conceptually similar device can also operated at resonance, $eF d = (E_2 - E_0)$ and $\hbar \omega = (E_2 - E_1)$, provided a population inversion between the subbands $E_2$ and $E_1$ is maintained by some means, such as a longer nonradiative lifetime in the $E_2$ subband compared to the $E_1$ subband. In this case, the radiation frequency cannot be tuned by the electric field but is set by the intersubband energy separation determined by the quantum well parameters.

The fabrication of such a device faces two fundamental obstacles. First, efficient amplification of a given frequency requires $F$ to be uniform over many SL periods, whereas the current-carrying charge tends to break up the SL into high and low-field
domains [28]. Second, the higher-lying states of a quantum well are confined by lower tunneling barriers, which works against population inversion. Recently, intersubband laser action was demonstrated in a quantum cascade laser (QCL) [32], illustrated in FIGURE 6. The QCL is a periodic structure alternating between a short-period SL and a double-well active region in which population inversion is established. As shown in FIGURE 6, under operating bias carriers flow through a superlattice miniband and tunnel into highest $E_2$ subband of the double-well active region. Tunneling out of the $E_2$ subband is impeded by the minigap of the downstream SL, so the carriers relax by radiative and nonradiative processes down to the $E_1$ and $E_0$ subbands, which can tunnel out into the downstream SL. Active region parameters are chosen to fix the $(E_1 - E_0)$ energy separation close to the optical phonon energy, leading to a much shorter lifetime of the $E_1$ subband and establishing population inversion and laser action at $\hbar\omega = (E_2 - E_1)$. Finally, the QCL retains overall charge neutrality under bias by proper doping of the SL regions, hence avoiding the domain formation problem. As a result, infrared lasers at $\lambda \approx 4.5$ and 8.4 µm have been demonstrated at operating temperatures above 100 K [32], leading to prospects of devices competitive with other sources in that frequency range.

Another application of tunneling-based optoelectronic devices comes in the area of quantum well intersubband IR photodetectors. Although in most of these devices IR radiation excites a transition from the lowest quantum well subband $E_0$ to the above-barrier continuum final states and current flows by hot-electron transport rather than tunneling [33], there have been versions where the final state is a higher-lying subband that is emptied by tunneling into an adjacent SL miniband [34]. In this case the photodetector structure is essential the inverse of the QCL in FIGURE 6, except that the active region consists of a single quantum well and the SL blocks tunneling out of the ground-state subband $E_0$. 
D TUNNELING纳米STRUCTURES AND COULOMB BLOCKADE

If a double-barrier RTS diode of FIGURE 1 (b) is made sufficiently narrow in one or both lateral directions, the electronic states in the quantum well will be further quantized, leading to 1D quantum wire subbands or fully quantized, atomic-like states. The \( I(V) \) characteristics of such RTS nanostructures are still governed by the same physics of tunneling into reduced dimensionality states [4], but with another effect coming into play: the tunneling of a single electron can tangibly alter the electrostatic field distribution over the entire structure. Modeling the double-barrier RTS as a small parallel-plate capacitor, the addition of a single electron to the quantum well population requires the energy of \( \frac{e^2}{2C} \), where \( C \) is the total capacitance between the quantum well and the rest of the structure. This discrete energy barrier arises from charge quantization and must be overcome for a carrier to tunnel into the well — a phenomenon known as the Coulomb blockade [35]. At low temperatures and in sufficiently small RTS, \( \frac{e^2}{2C} \) can be the largest parameter in the system and, instead of a smooth rise in the \( I(V) \) characteristic above \( V_{th} \), the current increases in discrete steps corresponding to the opening of additional single-electron tunneling channels [36]. These \( I(V) \) measurements, as well as low-frequency capacitance measurements on double-barrier RTS with impenetrable collector barriers [37] have been employed to probe the energy spectrum of artificial few-electron atoms in and out of magnetic fields.

Given an additional gate electrode to control the potential of the quantum well, the tunneling of single electrons into the well can be controlled in a transistor-like manner. Effective gate control is easier in the planar geometry, where electrostatically biased gates on top of a modulation-doped heterostructure [14] deplete the 2D electron gas and create a small 2D island isolated from the rest of the electron gas by potential barriers. An additional gate electrode can alter the effective size and capacitance of the island. As additional single-electron tunneling channels are opened by changing the gate voltage, very regular conductance peaks as a function of \( V_g \) have been observed [38, 39].

at low temperatures. There have been numerous proposals of single-electron transistors (SET) and other devices [40, 41] as the ultimate limit of miniaturization-driven semiconductor technology. It should be emphasized, however, that in addition to the extremely stringent fabrication requirements faced by large-scale SET circuitry at non-cryogenic temperatures, it is not clear that semiconductor SET realizations have any advantage over metal tunnel junctions, where the Coulomb-blockade phenomena were originally observed [42]. Thus, the first SET with voltage gain was realized in small Al tunnel junction capacitors [43].

One advantage of semiconductor-based Coulomb-blockade RTS with barriers created by electrostatic gating of 2D electron gas [39] is the separate tunability of barriers. Given small emitter-collector biasing in the Coulomb-blockade regime, only one electron at a time can pass through the device. If the emitter and collector barriers are lowered and then raised sequentially, a single electron will tunnel onto the island when the emitter collector is lowered and tunnel out when the collector barrier is lowered, resulting in the transport of a single electron through the device with every barrier biasing cycle. A very accurate current source that supplies a current $I = ef$, where $f$ is the barrier biasing cycle frequency, has been constructed [44, 45]. Such devices may find metrological applications as current standards.

E TUNNELING SPECTROSCOPY OF QUANTUM WELL PARAMETERS

Insofar as an experimental test structure for semiconductor characterization may be termed a device, tunneling structures have found a number of applications in the study of quantum well parameters. Here we briefly describe two quantities of interest accessible via tunneling measurements — quantum well coherent lifetimes and in-plane dispersions.

An optically pumped double-well system for measuring coherent lifetimes is illustrated in FIGURE 7 [46]. By adjusting the quantum well parameters and applying
an external electric field it is possible to bring the electron subbands into resonance (with the electron eigenstates of the double-well system becoming symmetric and antisymmetric combinations of the single-well states) without aligning the hole subbands. In this case, radiative recombination of electrons and holes occurring in different wells will produce luminescence at different frequencies. If one of the wells is selectively populated with a short optical pulse of the appropriate wavelength, the photoexcited electrons will execute Rabi oscillations between the quantum wells with a frequency \( \omega \) corresponding to the symmetric-antisymmetric energy splitting. The time-resolved photoluminescence from the double-well system then contains two oscillating signal frequencies that are out-of-phase and decay on a time scale of the coherent lifetime in the quantum well [46] — an effect that was observed in high-quality GaAs/AlGaAs quantum wells by femtosecond spectroscopy [47]. Similar information on the coherent lifetimes in a quantum well can be extracted from a modulation-doped double-well system with separate contacts to the two quantum wells and a gate \( V_g \) to control subband alignment [48]. At low temperatures, tunneling transport between the two quantum wells separated by a thick tunneling barrier is governed by the same \( E \) and \( k_{\perp} \) conservation laws as in standard RTS, with the difference that tunneling occurs between fully quantized 2D subbands. Hence tunneling is possible only when the subbands are perfectly aligned and in an ideal system the tunneling \( I(V_g) \) peak would be a delta function. The measured width of the tunneling peak, combined with independently measured in-plane 2D electron gas mobility, provides an experimental measure of the coherent lifetimes in the 2D subband.

Magnetotunneling measurements in \( p \)-type double-barrier RTS have been used to extract the complex, anisotropic valence subband dispersions in quantum wells. If the tunneling \( I(V) \) characteristics are measured in a transverse magnetic field \( B_{\perp} \), the \( k_{\perp} \) conservation rule for tunneling into the well is modified by \( \Delta k_{\perp} = eB_{\perp}<z> / \hbar \), where \(<z>\) is the distance traversed in the direction of tunneling [49]. As a result, the in-plane
dispersion of the 2D subbands $E_i(k_{\perp})$ can be calculated from the measured shifts in the magnetotunneling $I(V,B_{\perp})$ peak positions [50]. Furthermore, since the $B_{\perp}$-induced $\Delta k_{\perp}$ shift affects the $k_{\perp}$ component perpendicular to $B_{\perp}$, magnetotunneling measurements as a function of $B_{\perp}$ orientation should reveal the dispersion anisotropy between different in-plane crystallographic directions [51]. While measurements on GaAs/AlGaAs RTS did not show a significant anisotropy at attainable $B_{\perp}$ values, measurements on strained InGaAs/InAlAs $p$-type RTS [52] have been successful in extracting the in-plane band structure anisotropy (and stronger effects have been observed by this technique in the highly anisotropic Si/SiGe quantum wells [53]).

As a final note, another application of tunneling spectroscopy is the determination of strain redistribution in small strained heterostructures. As semiconductor devices push deep into the submicron regime, their size becomes comparable to strain relaxation length scales. Since the energy separation between the 2D subbands in RTS is partially determined by strain, $I(V)$ peak positions and lineshapes of small RTS devices should reveal the strain distributions in the quantum well. In this way, measurements of submicron $p$-Si/SiGe RTS have been employed to extract the strain relaxation in quantum wells of small devices [54].

E CONCLUSION

Intensive research over the past two decades has examined a number of interesting quantum-well structures in which carrier tunneling phenomena are harnessed to perform potentially useful electronic or optoelectronic functions. While to date none of these structures has progressed to real technological implementation because of fabrication and high-temperature operation problems, as well as competition from ever-improving standard transistor-based technology, further advances in epitaxial and lithographic techniques should enable tunneling-based devices to occupy niche applications (such as metrological current standards or infrared lasers). The possibility of tunneling-based memories or logic circuitry finding widespread technological
application appears less likely, unless major breakthroughs in materials and device performance are accomplished.

REFERENCES


[14] [Data review in this book: 7.2 Application of GaAs-based modulation doped heterostructures]


[33] [Datareview in this book: 8.3 Quantum well and intersubband photodetectors]


    K.K. Likharev, T. Claeson [Scientific Amer. (USA) v.266 (1992) p.80-85]


FIGURE CAPTIONS

FIGURE 1. (a) Schematic band diagram of an \( n \)-type double-barrier resonant tunneling structure under bias. (b) Cross-sectional diagram of a two-terminal RTS device pillar, doped regions shaded. (c) Low-temperature \( i(V) \) characteristic of a GaAs/AlGaAs double-barrier RTS. The valley current for \( V > V_p \) is due to scattering or phonon-assisted tunneling.

FIGURE 2. (a) Bipolar three-terminal RTS (after [10]), \( p \)-type doping permits a separate contact to the narrow quantum well. (c) Unipolar \( n \)-type three-terminal RTS, with a separate contact to a wider quantum well (analogous to the base in a heterojunction bipolar transistor, after [11]).

FIGURE 3. Gated two-dimensional RTS structure, after [15]. The electrons tunnel from the 2D electron gas (\( E_0 \) labels the bottom of the 2D subband) at the modulation-doped interface through the 1D subbands of the quantum wire (only the lowest 1D subband \( E_{0}' \) is shown). The gate potential controls both 2D gas density and \( E_{0}' \) [15], for experimental realization see [16].
FIGURE 4. Schematic resonant hot-electron transistor (RHET) band diagram under bias. The resonant peak current of the RTS emitter structure is reproduced in the collector current (after [20]).

FIGURE 5. Multipeak $I(V)$ characteristic of a cascaded RTS structure with a field-effect transistor load employed for a nine-state memory at room temperature [23]. Inset shows the schematic memory circuit: the field-effect transistor load provides a constant current to set the various stable points.

FIGURE 6. Schematic diagram of one period in the quantum cascade laser (QCL). The radiative transition frequency is indicated, it can be varied by adjusting the double quantum well parameters [32].

FIGURE 7. Schematic diagram of the double-well tunneling structure for measuring coherent lifetimes in a quantum well implemented in GaAs/AlGaAs. Electrons oscillate between the quantum wells giving rise to out-of-phase luminescence signals at frequencies $v_1$ and $v_2$. The decay of the oscillating luminescence gives the coherent lifetime [46, 47].
Superlattices and Microstructures is a journal disseminating the science and technology of synthetic heterostructures, including individual and collective use of semiconductors, metals and insulators for the exploitation of their unique properties. The journal hosts papers dealing with fundamental and applied experimental research as well as theoretical studies. Fields of interest, including emerging ones, cover: Heterostructures, nanowires, quantum dots. Thin films, surfaces and interfaces. Photonics. Quantum electronics, optoelectronics. Plasmonics. Spintronics. Wide bandgap semiconductors, @inproceedings{Bhattacharya2014PropertiesOI, title={Properties of III-V Superlattices and Quantum Wells}, author={Pratip K. Bhattacharya and Serge Luryi and Alexander Zaslavsky}, year={2014} }.

A INTRODUCTION Advances in epitaxy have permitted the design of semiconductor devices with sufficient precision to make use of quantum mechanical tunneling. The potential advantages of tunneling-based devices consist of high-speed operation and higher functionality inherent in their strongly nonlinear current-voltage characteristics, while the disadvantages include the relative weakness of quantum effects at room temperature and the difficult integration into mainstream semiconductor technology.