(1)

NATO Advanced Study Institute: Granular Nanoelectronics

July 23 – August 4, 1990, Il Ciocco, Italy

Program

Poster Abstracts

List of Attendees

NATO Advanced Study Institute: Granular Nanoelectronics

July 23 – August 4, 1990, Il Ciocco, Italy

Please remember that the role of the Advanced Study Institute is one of education. The talks reflect some introductory material and make connection with the other speakers, while at the same time presenting up-to-date research. Each will try to relate in some way to the overall theme of the ASI, which is granular nanoelectronics – the future of few-electron systems for electronics applications. *Information is best transmitted by a lively question and answer period. We encourage each attendee to ask questions and make sure that the topic is made understandable to him (or her).*

Talks 1-3 are morning talks each day, while talks 4,5 are afternoon talks.

Program

Monday, 23 July

1. Opening Remarks and general welcome
2. Carlo Jacoboni, "General welcome, introduction to the school, and general aspects of quantum systems"
3. Dave Ferry, "Projections on the future of ULSI and nanolithograpy"
4. Steve Beaumont, "Nanofabrication of quantum wires and rings"
5. Karl Hess, "Path integrals and transport in mesoscopic systems"

Tuesday, 24 July

1. Markus Büttiker, "Transmission probabilities and electric resistance"
2. Alan Fowler,"Interference phenomena in devices structures"
3. John Barker, "Introduction to quantum transport in quantum waveguides"
4. Tony Leggett, "Dephasing and non-dephasing collisions in nanostructures"
5. Trevor Thornton, "Electronic fluctuations and transport in waveguides"

Wednesday, 25 July

1. Jörg Kotthaus, "Spectroscopy of electronic excitations in low dimensional systems"
2. Alan Fowler, "Waveguide devices: rings, non-local transport, etc."
3. POSTER SESSION 1
4. Mark Reed, "Tunneling investigations of low dimensional systems"
5. Erick Gornik, "Tunneling between constrained dimensionality systems"

Thursday, 26 July

1. Jörg Kotthaus, "Lateral confinement in surface superlattices"
2. Dave Ferry, "Magnetoconductance in surface superlattices"

3. Carlo Beenakker, "Ballistic and adiabatic transport I"
4. Markus Büttiker, "Quantum oscillations in loops and tunnel junctions"
5. Steve Beaumont, "Optical properties of arrays of microdevices"

Friday, 27 July

1. Carlo Jacoboni, "Numerical approaches to non-equilibrium quantum transport"
2. Mark Reed, "Non-equilibrium quantum effects – structures and devices"
3. Lino Reggiani, "Noise in small and ultra-small device structures"
4. Gerhardt Abstreiter, "Growth engineering of mesoscopic systems"
5. Antti Jauho, "Non-equilibrium Green's functions in quantum transport"

Monday, 30 July

1. Lex Akers, "Neural and constrained interconnect automata"
2. J. P. Launay, "Molecular electronics"
3. John Barker, "Granular nanoelectronic phenomena"
4. Gerhardt Abstreiter, "Quantum field effects on optical properties of superlattices"
5. Chihiro Hamaguchi, "Optical properties of short period superlattices"

Tuesday, 31 July

1. Trevor Thornton, "Resistance fluctuations and quantization in microdevices"
2. Aaron Szafer, "Fluctuations and quantum chaos in ballistic transport"
3. Carlo Beenakker, "Ballistic and adiabatic transport II"
4. Friedl Kuchar, "Microwave studies of quasi-one dimensional wires"
5. Chihiro Hamaguchi, "Non-equilibrium carrier transport in small structures"

Wednesday, 1 August

1. Aaron Szafer, "Quantum circuits and non-locality"
2. Tony Leggett, "Some considerations related to the quantization of charge in mesoscopic systems"
3. POSTER SESSION 2
4. K. K. Likharev, "Single-electronics: The correlated transfer of single electrons in ultrasmall tunnel junctions, arrays, and systems"
5. L. J. Geerligs, "Charging effects and 'turnstile' clocking of single electrons in small tunnel junctions"

Thursday, 2 August

1. Toshiaki Ikoma, "Weak localization and phase-breaking mechanisms of electron waves in quasi-one-dimensional wires"
2. Karl Hess, "Numerical simulations of electron waveguides"
3. Lex Akers, "VLSI implementations of neural networks"
4. Antti Jauho, "High field quantum transport"
5. Lino Reggiani, "Monte Carlo algorithms for non-equilibrium quantum transport"

Friday, 3 August

1. Gerry Iafrate, "Correlation and exchange in single electron events"
2. Bob O'Connell, "The few-body problem in nanoelectronics"

3. John Barker, "Interfacing molecular and biological systems with electronic mesoscopic systems"

CONFERENCE WRAPUP

Poster Manuscripts

Poster presentations are to last for 3-4 minutes, and should be viewed as an "advertisement" for the poster itself. In general, no questions will be entertained during the poster "shotgun" presentation. A decision will be made by the organizers during the ASI as to which posters should be included in the published proceedings.

Posters that are accepted for the proceedings will be allowed *4 single-spaced A4 pages* (typing space of 6.25"x10.194"), which includes all references and figures (manuscripts which exceed this limit will be arbitrarily shortened by the editors). The format can be taken from the lecture notes provided to attendees, for which the text is prepared originally as 12 point, Times-Roman. Original line drawings can be submitted with the manuscript, and will be scanned into the document itself. The manuscript is due to Prof. D. K. Ferry, Department of Electrical Engineering, Arizona State University, Tempe, AZ 85287-5706, by **September 1, 1990.** A 3.5" disk containing the manuscript as a word processing .ile should be submitted with the manuscript. Either Macintosh or MSDOS format is acceptable, although the former is preferred. Please note that the full titles and author lists are required for the cited references. Equations should be set with MacEqn (preferred to Expressionist), or written in by hand for resetting by the editors.

Abstracts

Session 1
Wednesday, July 25

1-1 25 to 85nm Gate Length GaAs MESFETs and HEMTs, J. M. Ryan, J. Han, and D. K. Ferry, *Arizona State University, U.S.A.* GaAs MESFETS and HEMTs with gate lengths from 85 to 25nm have been fabricated using electron beam lithography. Maximum transconductances of 108 mS/mm and 215 mS/mm were obtained for the MESFETs and HEMTs, respectively. These apparently low values are in the expected range for devices with gate-length-to-channel ratios much smaller than one, so that capacitance fringing is important. The transconductance demonstrates velocity overshoot is for devices with gate lengths below ~55nm. For gates shorter than ~35nm, the overshoot-related increase in transconductance saturates. This effect is explainable in terms of carrier heating in the ohmic contact, which decreases the effect of overshoot at the gate.

The high-frequency response of the devices was tested to 20 GHz. The MESFETs yielded a maximum cutoff frequency f_t of 167 GHz, with f_t following a linear relation with the inverse gate length for gate lengths down to 55 nm. The HEMTs had a maximum f_t of 18 GHz, a number limited by the extra gate capacitance needed for the measurement configuration.

For HEMTs with gate lengths shorter than 30nm, the current is found to vary exponentially with the channel voltage. Simple calculations indicate that this would result when tunneling is the dominant current mechanism. Similar behavior has been observed in short-gate devices of other groups. For gate lengths larger than 30nm, we find the usual voltage dependence. For MESFETs with gate lengths shorter than 30nm, no evidence of an exponential relationship is found. This is probably due to weaker confinement of carriers in the MESFETs as compared to the HEMTs.

1-2 Quantitative Determinationof the Magnetic-Field Density of States as a Function of Fermi Energy in the Two-Dimensional Electron Gas, Raymond Ashoori, *Cornell University, U.S.A.* Using capacitive techniques, we have studied both the tunneling[1] density of states (DOS) and the thermodynamic DOS in a two-dimensional (2D) electron gas of variable density (0 to $6\times10^{11}cm^{-2}$) in the presence of magnetic field applied perpendicular to the plane of the electron gas. Our sample design allows for calculation of the thermodynamic DOS as a function of Fermi energy of electrons in the 2D gas from capacitive data using no smaple parameters other than the sample area. This type of measurement is distinct from other DOS measurements which have typically measured the DOS as a function of magnetic field at fixed density. The DOS determination presented here allows for direct comparison with models of the DOS as a function of energy. In contrast with fixed-density DOS results[2] which are fit well by Landau levels taken to be Gaussian (as a functionof electronic energy) plus a background DOS, our results for the DOS as a function of Fermi energy are fit best (at 4T and lower) with Lorentzian lineshapes. Further, at high fields (8.5T) the exchange-enhanced spin splitting is observed, and the exchange energy for electrons is determined. Finally, methods and results of application of these capacitive techniques to quantum dot arrays patterned on our samples will also be presented briefly.

[1] R. C. Ashoori, J. A. Lebens, N. P. Bigelow, and R. H. Silsbee, *Phys. Rev. Lett.* 64, 681 (1990).

[2] See for example E. Gornik, R. Lassnig, G. Strasser, H. L. Störmer, A. C. Gossard, and W. Weigmann, *Phys. Rev. Lett.* 54, 1820 (1985).

1-3 Study of Lower-Dimensional Transport by Electroluminescence, Hans P. Zappe, *Fraunhofer Institut für Angewandte Festkörperphysik, West Germany.* It is well known that many semiconducting materials and devices emit visible light when subject to high electric fields. This electroluminescence is becoming apparent at quite low operating voltages in modern, small-sized structures and may be used to study the transport behavior in transistors as well as 2- and 1-dimensional conducting layers. We have used this effect to study the scattering, real-space transfer and energy-loss mechanisms of conduction electrons in some of these systems.

The voltage-dependent intensity, lateral distribution and energy spectrum of the electroluminescence has been measured in a variety of structures. Thee nature of the sub-bandgap emission, coupled with bandgap peaks, indicates thatradiative recombination combines with indirect intraband transitions (Bremsstrahlung) to produce the observed photon spectrum. The observed polarization of the light from GaAs/AlGaAs field-effect devices (MESFETs) may be attributed to a dominance of forward-directed electron scattering, an effect predicted theoretically. The spectrum of radiation fromthe 2DEG of GaAs/AlGaAs HEMTs explicitly shows the energy-dependent real space transfer effect, and the "vertical" transfer of 2D conduction electrons to other device layers may be studied in detail.

These measurements, presently used to study transport in 2D systems, are currently being extended to examination of high-field transport in 1D devices. The electroluminescence spectra will likely provide a useful means to investigate the high-field and non-equilibrium transport in such lower-dimensional structures.

1-4 High Injection Effects in GaAs/AlGaAs Quantum Wells: Spontaneous Recombination and Band-Gap Renormalization, N. Kirstaedter, E. H. Böttcher, M. Grundmann, and D. Bimberg, *Institut für Festkörperphysik der Technischen Universität, Berlin, West Germany.* A comprehensive investigation of the charge carrier recombination and the band-gap renormalization (BGR) in GaAs/GaAlAs single quantum well (SQW) structures is presented. The results are of particular importance for the modeling of QW-based semiconductor devices. By measuring the differential lifetime, the spontaneous emission rate and the chemical potential as functions of the carrier injection rate in GaAs/AlGaAs SQW GRIN-SCH lasers, the carrier density dependence of the recombination rate and the BGR is determined in a sheet density range of 3×10^{11} cm^{-2} up to 7×10^{12} cm^{-2}. The radiative and non-radiative components contributing to the total emission rate are identified. It is shown that the radiative band-to-band recombination coefficient decreases significantly with increasing density. At low densities, the non-radiative contribution is dominated by interface recombination. At high densities, Auger-recombination becomes an important factor.

For a typical 7 nm GaAs/$Ga_{82}Al_{18}As$ QW structure at room temperature, the BGR increases from –20 meV to –70 meV. In contrast to previous investigations, which are based upon the fitting of photoluminescence or gain spectra, the number of fitting parameters for the extraction of BGR data is significantly reduced. In addition, it is shown that an appropriate band model for the QW, which takes into account band mixing effects of the valence bands, has to be used for the calculation of the electron and hole quasi-Fermi potentials.

1-5 Novel I-V Characteristic of Superlattices, Ping Ao, *University of Illinois, Urbana, IL, USA* and Jørgen Rammer, *Institutt for Fysikk, Norges Tekniske Høgskole, Universitetet i Trondheim, Trondheim, Norway.* In artificial structures, such as superlattices, arbitrary relationships between bandstructure parameters can be achieved, enforcing the need for a reinvestigation of interband transitions of a crystal electron originally studied by Zener. We report the existence of non-exponential corrections to the Zener tunneling rate and demonstrate their importance in parameter regimes relevant for quantum transport properties of superlattices such as Bloch oscillations. Novel oscillations in the I-V characteristic of a superlattice are predicted with periods determined by the band structure parameters.

1-6 Monte Carlo Simulation of Lateral Surface Superlattices in a Magnetic Field, T. Yamada, A. M. Kriman, and D. K. Ferry, *Arizona State University, Tempe, AZ, USA.* When a low magnetic field is applied to an electron gas in a crystal, the spectrum consists of sharp, closely spaced Landau levels. With increasing field, the finiteness of the lattice constant leads to nonparabolicity corrections, which broaden and shift these levels. When the cyclotron radius r_C becomes comparable to the lattice constant a, a perturbative approach fails. Hofstadter[1] and others have studied the full range of behaviors that occurs for general values of r_C/a. For a square lattice in the strong lattice-potential limit, the energy spectrum plotted as a function of the number of flux quanta per a unit cell is fractal, with a period of unity. A single period of this spectrum takes the form of a "butterfly." There are well-defined energy gaps, but the energy bands are extremely sensitive, nonanalytic functions of the field.

In real crystals, the lattice constants are so small that the magnetic fields corresponding to the above condition are unachievably large - on the order of 10^8G. However, this condition can now be realized in two-dimensional (2D) electron gases with artificially engineered lattices. In these lateral surface superlattices (LSSLs), 2D superlattice potentials can be achieved with periods on the order of 0.1μm, leading to the 10^5 reduction in the field required. LSSLs constructed on MESFET and HEMT structures with meshed gate electrodes have confirmed the periodicity in magnetic field, and displayed a variety of other transport properties.[2,3]

In order to study these structures, we perform a semiclassical Monte Carlo simulation of quasi-2D electrons described by a Stern-Howard wave function,[4] including electron-electron interaction, and scattering by acoustic phonons, optical phonons, and impurities. The electron-electron interaction is treated using a molecular dynamics technique.[5] We model GaAs-based LSSLs at 4.2K, with electron concentrations of 1.4×10^{10} cm^{-2}, and a superlattice potential described by $V_0[\cos(2\pi x/a)+\cos(2\pi y/a)]$ with V_0= 10meV and superlattice period a = 0.16μm.

The diffusivity is evaluated as a function of the magnetic field. It changes abruptly with small changes in the magnetic field, suggesting a fractal structure. The fractal structure found in the quantum case can be understood in terms of commensurability of two length scales - the superlattice potential period and the cyclotron radius. In the classical case studied

here, a commensurability issue is still present which concerns the cyclotron frequency and the frequency of oscillation in the superlattice-potential minimum.

[1] D. Hofstadter, *Phys. Rev. B* 14, 2239 (1978).
[2] G. Bernstein and D. K. Ferry, *Z. Phys. B* 67, 449 (1987).
[3] R. A. Puechner, J. Ma, R. Mezenner, W.-P. Liu, A. M. Kriman, G. N. Maracas and D. K. Ferry, *Surf. Sci.* 228, 520 (1990).
[4] F. Stern and W. E. Howard, *Phys. Rev.* 163, 816 (1967).
[5] P. Lugli and D. K. Ferry, *Phys. Rev. Lett.* 46, 594 (1985).

1-7 Far-Infrared Photoconductive Response of a One-Dimensional GaAs-(Ga,Al)As Channel in Zero and High Magnetic Fields, T. J. B. M. Janssen,[2] N. K. Patel,[1] J. Singleton,[2] M. Pepper,[1] H. Ahmed,[1] D. G. Hasko,[1] R. J. Brown,[1] J. A. A. J. Perenboom,[2] G. A. C. Jones,[1] J. E. F. Frost,[1] D. C. Peacock[1] and D. A. Ritchie[1], ([1]*University of Cambridge, Cambridge, United Kingdom;* [2]*University of Nijmegen, The Netherlands*). We have studied the photoconductive response of a one dimensional (1D) channel to monochromatic far-infrared (FIR) radiation from a laser (wavelengths 100-1200 μm) as a function of applied magnetic field at temperatures ~ 0.4 K. The channel is defined by a negative voltage applied to two gates 0.3 μm wide and 0.3 μm apart on top of a high mobility GaAs-(Ga,Al)As heterojunction: the transport through the channel is ballistic, and in zero magnetic field the resistance is quantised to a value $h/2ie^2$, where i is the number of occupied 1D subbands. In a magnetic fielld applied perpendicular to the plane of the heterojunction, the resistance as a function of gate voltage exhibits plateaux due to conductionthrough the channel by a finite number of edge states.

With a constant current through the channel, and in arbitrary magnetic field, the photovoltage across it due to the FIR exhibits oscillations as the gate voltage, and hence the number of 1D subbands or edge states within the channel, is varied: the oscillations are at positions related to the resistance plateaux, with the peakvalue of the photovoltage occurring to the low conductivity side of each plateau. It appears this non-resonant response results from an alteration of the electron energy distribution on either side of the channel due to the heating effect of the FIR.

At certain values of the magnetic field and gate voltage, however, the channel shows a strong **resonant** photoresponse, which is thought to consist of two components:

i) when the FIR is resonant with the Landau level spacing outside the channel, the first empty Landau level above the Fermi energy may be populated by photoexcited electrons: these photoexcited electrons will have access to extra states for conduction through the channel;

ii) when the FIR is resonant with the edge state spacing within the channel, electrons may be excited to states above the Fermi energy in the channel.

These results may be treated within the Büttiker formalism for phase coherent conductance between different electron reservoirs: within the reservoirs the electron energy distribution is altered by the absorption of FIR radiation. The implications of these observations for the lifetime of electrons in the Landau levels and channel states above the Fermi energy will be discussed.

1-8 Contribution of Ballistic Electrons in Vertically Integrated Resonant Tunneling Diodes, R. E. Carnahan, J. J. L. Rascol, K. P. Martin, and R. J. Higgins, *Georgia Institute of Technology, Atlanta, Georgia, U.S.A.*, L. A. Cury and J. C. Portal, *LPS-Institut National des Sciences Appliquees, Toulouse cedex and Service National des Champs Intenses-Centre National de la Recherche Scientifique, Grenoble cedex, France*, B. G. Park, E. Wolak, K. L. Lear, and J. S. Harris, Jr., *Stanford University, Stanford, California.* We have made a systematic study of the current-voltage (I-V) characteristics of vertically integrated resonant tunneling diodes (RTDs) in the presence of a transverse magnetic field ($\mathbf{B} \perp \mathbf{J}$) up to 18 T and at temperatures below 4.2 K. The unit RTD consisted of 17 Å AlAs barriers and a 45 Å GaAs well, surrounded on both sides by a 100 Å undoped GaAs spacer layer. By varying the n-doped GaAs separation layer between two identical RTDs, we were ahle to systematically study the effect of ballistic electrons on the device. In the decoupled case (1000 Å doped separation layer), relaxation of the electrons resulted in two negative differential resistance (NDR) peaks which both behave as single RTD peaks under the influence of **B**. However, the non-negligible minority ballistic electrons in the intermediate doped spacer layer (500 Å) sample had their incident kinetic energy magnetically tuned to be resonant with the second quantized level in the collector RTD, causing a repartition of the potential across the RTD. Thus the second NDR transition (by the majority thermalized electrons in the current) was shifted to a higher bias by 400 mV. The third sample (with no doped separation layer) was dominated by ballistic transport and showed interference features for biases below the first peak voltage. The transverse magnetic field enabled the first peak current to dramatically increase with a slight increase in bias. An anomalous strong decrease in both the second peak voltage and current was ohserved. Additionally, a feature we associate with tunneling through the X valley in the AlAs

collector.RTD barriers is observed for B > 10 T. We have shown that, even if the major portion of the current is thermalized, strong effects due to ballistic electrons are observed in integrated heterostructures.

1-9 Elastic and Inelastic Resonant Tunneling in an Imperfect Superlattice, P. Hyldgaard and A. P. Jauho, *H. C. Ørsted Institute, Copenhagen, Denmark.* A model calculation of vertical transport through an imperfect superlattice is presented. A tight-binding model is used to describe transport in the superlattice (SL). The imperfection we study can be modeled by a double barrier. We include effects of (dispersionless) longitudinal optical phonon scattering in the quantum well.

The model calculation includes the miniband that arises in an SL, and resonant tunneling within the miniband is found. Phonon-assisted resonant tunneling is found to be operative, and a phonon satellite is identified.

The calculations are done for experimentally realizable parameters, and constitute predictions for the experimental situation of two ideal superlattices interrupted by a set of neighboring higher/thicker barriers, with the entire imperfect superlattice situated in the base of a transistor.

1-10 Magnetic Field Effect on Resonant Tunneling in Short-Period Superlattices, A. M. Vasilev and I. N. Uraltsev, *A. F. Ioffe Physico-Technical Institute, Leningrad, USSR.* There has recently been a dramatic increase in research activities in the transport of carriers perpendicular to the layers of the strongly coupled superlattices (SLs), which has been shown to occur through extended, Bloch-type states.[1,2] The perpendicular transport is expected to be suppressed in the presence of magnetic field parallel to the SL layers due to the wave function squeezing resulting in carrier localization when the magnetic length becomes less than the SL period. Destroying of the SL extended states has been predicted to occur at remarkably low magnetic fields as a result of the untuning of the resonance between electron states when cyclotron energy exceeds the miniband halfwidth.[3]

We report the observation of the magnetic-field-induced suppression of the heavy-hole and electron transport through the extended states in $GaAs/Al_xGa_{1-x}As$ SLs. A magnetic field applied along the SL layers is found to produce a remarkable effect on the SL luminescence used for study of ambipolar transport of photocarriers in SLs with an enlarged QW. The efect is shown to depend dramatically on the SL period, d, for d<50Å to result from the magnetic-field-induced transition from coherent Bloch transport to the transport through localized states of the heavy holes. To demonstrate destroying of the resonance between electron states in the presence of magnetic field we have measured the diamagnetic shift of the heavy-hole exciton in the SLs with narrow electronic minibands. In SLs with d<100Å we have observed the strong quenching of the excitonic shift associated with the magnetic-field-induced localization of electrons.

[1] B. Deveaud, J. Shah, T. C. Damen. B. Lambert, A. Chomette, and A. Regreny, *IEEE J. Quant. Elec.* 24, 1641 (1988).
[2] P. S. Kop'ev, R. A. Suris, I. N. Uraltsev, and A. M. Vasiliev, *Solid State Commun.* 72, 401 (1989).
[3] A. M. Berezhkovskii, and R. A. Suris, *Sov. Phys. JETP* 86, 109 (1984).

Session 2
Wednesday, August 1

2-1 Thermopower in Scanning Tunneling Microscope Experiments, J. A. Støvneng, *Institutt for fysikk, Norges Tekniske Høgskole, Universitetet i Trondheim, Trondheim, Norway,* and P. Lipavský, *Ohio State University, Columbia, Ohio, U.S.A.* We present a theory for the thermopower observed in scanning tunneling microscopy (STM). The lateral variation of the thermopower is found to depend on the logarithmic derivative of the local sample density of states at the Fermi level. We also derive a relation between the thermopower and the nonlinear conductance. The heat transfer dueto the tunneling electrons obeys the Weidemann-Franz law.

2-2 Effective-mass Boundary Conditions for Strained Heterostructures, G. T. Einevoll and P. C. Hemmer, *Institutt for Fysikk, Norges Tekniske Høgskole, Universitetet i Trondheim, Trondheim, Norway.* The question of the appropriate form and validity of the effective-mass approximation when applied to strained heterostructures is addressed. Complications arise from the position-dependence of the effective mass and lattice constant, and the form of the appropriate kinetic operator in the effective-mass Hamiltonian is not *a priori* given. For the one-band theory we propose the following two-parameter family of hermitian kinetic operators

$$H_{kin} = -\frac{\hbar^2}{2} m^{\alpha} a^{\delta} \nabla m^{\beta} a^{-2\delta} \nabla m^{\alpha} a^{\delta}, \qquad (1)$$

where m and a are the, in general, position-dependent effective mass and lattice constant respectively, and one has the obvious requirement $2\alpha+\beta = -1$. The corresponding effecive-mass boundary conditions to use at abrupt heterointerfaces are uniquely determined by the form of H_{kin}, and it follows that the quantities $m^{\alpha}a^{\delta}\phi$ and $m^{\beta+\alpha}a^{-\delta}\nabla\phi$, where ϕ is the effective mass wave function, must be conserved at. material interfaces. In our approach the parameters α, β, and γ

are determined by comparing effective-mass results with exact results for simple microscopic models for abrupt heterostructures. Within our calculational scheme we determine uniquely $\alpha = 0$ and $\beta = -1$, while two distinct values, 0 and -1, are observed for δ. Based on qualitative similarities between the test models and realistic systems, we propose for calculations on conduction-band states the boundary conditions listed above with $\alpha = 0$, $\beta = -1$ and $\delta = -1$. When a one-band effective-mass theory is applicable for valence-band states, however, the parameter values $\alpha = 0$, $\beta = -1$ and $\delta = 0$ should be used.

2-3 Berry's Phase and Persistent Charge and Spin Currents in Textured Mesoscopic Rings, Daniel Loss, Paul Goldbart, and A. V. Balatsky, *University of Illinois, Urbana, Illinois, U.S.A.* The quantum orbital motion of electrons in mesoscopic normal metal rings threaded by a magnetic flux produces striking interference phenomena such as persistent currents and the Aharonov-Bohm effect. Similarly, when a quantum spin adiabatically follows a magnetic field which rotates slowly in time, the phase of its state vector acquires an additional contribution known as the Berry phase.

The purpose of this paper is to explore the combination of these two quantum phenomena by examining the interplay between orbital and spin degrees of freedom for an electron moving in a mesoscopic ring. To this end, we consider a ring which is placed in a classical, static, inhomogeneous magnetic field, *i.e.*, a texture. Zeeman coupling between the electron spin and this texture results in a Berry phase and, as a consequence, the system supports persistent equilibrium spin and charge currents, even in the absence of conventional electromagnetic flux through the ring. We mention the possibility of analogous persistent mass and spin currents in normal ^{3}He and spin-polarized H.

2-4 A Numerical Method for the Calculation of Transient Response in Mesoscopic Devices, Leonard F. Register and Umberto Ravaioli, *University of Illinois, Urbana, Illinois, U.S.A.* We present a numerical method for the solution of the 2-D time-dependent Schroedinger equation, suitable for the investigation of transients in mesoscopic devices under ballistic conditions. The features of the method are: (1) a tight-binding formulation of the quantum mechanical Hamiltonian; (2) 2-D absorbing boundary conditions to simulate the discretized open system; (3) Crank-Nicholson scheme for the evaluation of the space dependent operator, solved with an Alternate Direction Implicit (ADI) strategy. We believe that this is the first application of 2-D absorbing boundary conditions for the Schroedinger equation. The ADI scheme decouples the discretized linear system of equations into subsets requiring only the solution of tridiagonal systems. In addition, since all dependencies of data are eliminated, the code can be successfully vectorized [1]. The method allows us to investigate separately the various energy components, instead of having to resort to wavepackets, and this results in better flexibility for the future application and comparison of different formulations for the injecting conditions at the contacts. To illustrate the features of the algorithm, we present results of switching calculations for an ideal T-structure proposed earlier as a quantum modulated transistor. Although the structure is modelled as an ideal electron waveguide, the essential features of the quantum interference effects during the turn-on transient can be appreciated and understood.

[1] J. R. Barker, J. Pepin, M. Finch and M. Laughton, Solid State Elec., 32, 1155 (1989).
[2] F. Sols, M. Macucci, U. Ravaioli and K. Hess, J. Appl. Phys., 66, 3892 (1989).

2-5 Dephasing by a Dynamic Environment, D. Loss and K. Mullen, *University of Illinois, Urbana, Illinois, U.S.A.* We investigate the manner in which quantum interference is suppressed whena particle interacts with a localized, dynamical environment. To do so we examine a model with two classical paths along which an electron can travel, and allow it to interact with a bath of harmonic oscillators on one path, and travel freely on the other. In particular we show that the quantum fluctuations of the path of the particle can couple to the environment and thus lead to dephasing, and calculate the dephasing time in the high temperature limit. We compare this result to other views of how propagating electrons lose phase coherence.

2-6 Scattering in Nearly-Clean Mesoscopic Structures, A. M. Kriman, B. S. Haukness and D. K. Ferry, *Arizona State University, U.S.A.* We study theoretically the effect of small numbers of elastic scatterers on ballistic transport in low-dimensional microstructures. Both numerical and analytical methods are used, within a transfer matrix formulation. With lateral confinement, delta-function and other extremely sharp models of a single defect lead to sharp resonances when such defects are well isolated, occurring when total electron energy equals any miniband energy. Unit transmission probability is approached when the scattering defect is small, and far from other scatterers, even if the scatterer is strong enough to decrease significantly the conductance away from resonance. Resonances occur for all shapes of confining potential.

At low energies, single-barrier structures with a single nearby defect exhibit resonant transmission similar to that of double-barrier resonant tunneling

diodes (DBRTDs), with an approximate scaling behavior that relates transmission for defects at different distances to that at a fixed distance with different energy scales. In ordinary DBRTDs, the position of the transmission peak is strongly affected only by defects lying within the quantum well region. The height of the transmission peak is very sensitive to the positions of defects within that region, which act essentially as a probe of the resonance wave function. Defects in front of a DBRTD also affect the valley current by modifying the longitudinal component of the incident momentum.

2-7 The Effect of Dissipation on Phase Periodicity and the Quantum Dynamics of Josephson Junctions, D. Loss and K. Mullen, *University of Illinois, Urbana, Illinois, U.S.A.* We consider systems described by compact, periodic variables, such as Josephson junctions and 5mall normal metal rings. We examine how they can be coupled to a heat bath, via either momentum or position coupling, and how the two descriptions can be related by a unitary transformation. We show how it is critical to transform not only the Hamiltonian, but also the initial conditions. We then demonstrate that for certain types of initial conditions, paths of different winding number can interfere. Still other, ostensibly reasonable initial conditions, have no such interference. We conclude with a discussion of appropriate models to describe periodic systems.

2-8 The I-V Characteristic of a Resistively Shunted, Small Capacitance Josephson Junction, V. Bubanja, A. Maassen van den Brink, D. V. Averin* and G. Schön, *Delft University of Technology, Delft, The Netherlands.* We consider the low temperature behaviour of a small capacitance superconducting junction which is shunted by an ohmic shunt. The dynamics of such junctions was studied in ref. [1] within the framework of a deterministic single-band model. We allow finitie values of Zener tunneling probability and include transitions bewtween different quantum states caused by the shunt. A similar situation, but where the dissipation is due to quasiparticle tunneling, has been discussed in refs. [2] and [3]. Here we concentrate on the limit of negligible quasiparticle tunneling. For currents less than e/R_sC (R_s denotes shunt resistance, C junction capacitance) all the current is going through the shunt, so that the I-V curve is linear. Beyond it follows a branch with negative differential resistance due to Bragg reflection and due to the Cooper pair tunneling induced by the shunt. For large currents Zener tunneling prevails and the I-V curve bends over to the ohmic asymptote.

**permanent address: Moscow State University, Moscow, U.S.S.R.*

[1] K. K. Likharev and A. B. Zorin, J. Low Temp. Phys. 59, 347 (1985).
[2] D. V. Averin and K. K. Likharev, J. Low Temp. Phys. 62, 345 (1986) and to be published
[3] U. Geigenmüller and G. Schön, Physica B 152, 186 (1988).

2-9 Single-Electron Tunneling in Point-Contact Tunnel Junctions, R. T. M. Smokers, P. J. M. van Bentum and H. van.Kempen, *University of Nijmegen, Toernooiveld, Nijmegen, The Netherlands.* Charging effects like Single-Electron Tunneling or the Coulomb staircase can be observed in point-contact tunnel junctions on a wide range of materials. Thsi indicates that the charging energy $Q^2/2e$ can be the dominant term in the Hamiltonian describing the tunneling in these junctions. Recent theoretical developments have elucidated that this can only be the case if the high-frequency impedance of the junction environment is large enough to effectively decouple the junction from its long-range electromagnetic environment (e.g. stray capacitances). For a series connection of tunnel junctions, as in the case of tunneling through small isolated paticles, this condition is obviously satisfied. The observations in single junctions however cannot be readily understood. To investigate the influence of the electromagnetic environment we have performed tunneling measurements on surface-doped Si, at temperatures close to the metal-insulator transition. We observe clear S.E.T.-behavior, with capacitances of order 10^{-17}F, but find no appreciable effect when changing the series resistance of a point contact from $\approx 50k\Omega$ to$\approx 80M\Omega$. Decoupling from stray capacitances may be due to a relatively slow hopping conduction through impurity states in the direct vicinity of the point contact.

2-10 A Generalized Impact-Ionization Model for High-Energy Electron Transport in Si with Monte Carlo Simulation, Rossella Brunetti, *Università di Modena, Italy.* A new model for impact ionization in Si is presented, which goes beyond the limitations of the Keldysh formula and is based on a more realistic scheme developed starting from a first-order perturbation theory. This scattering mechanism is modeled by an extended band structure which includes many bands for electrons and one band for holes in a finite Brillouin zone. Some processes have been identified to bring the dominant contribution to the scattering probability, in the present approach, foir electron energies up to 3eV. Expressions for the differential and integrated scattering probabilities have been obtained which are consistent with the band model and can be included in a Monte Carlo simulation of the electron gas. Results for transport quantities are presented for a bulk material in presence of

homogeneous and static electric fields under physical conditions where impact ionization influences the carrier dynamics. A comparison with theoretical and experimental data from the literature is also given.

2-11 Ambipolar Perpendicular Transport in a Semiconductor Slab: Influence of Optical Phonons, K. Scheller, T. Held, and G. Mahler, *Universität Stuttgart, West Germany.* We investigate the perpendicular transport of an optically excited carrie plasma in a thin semiconductor slab. The excitation is done by a monochromatic laser, homogeneously illuminating one surface. A kinetic description by means of the Boltzmann equation is used. On length scales greater than the absorption length and on time scales greater than the carrier-carrier scattering time the distribution function is approximated by heated displaced Maxwellians. We focus on the role of optical phonons for the transport and relaxation properties of the plasma.

2-12 Phonon scattering and energy relaxation in two, one and zero dimensional electron gases, U. Bockelmann and G. Bastard, *Laboratoire de Physique de la Matière Condensée de l'ENS Paris, France.* We report on calculations of intra- and inter-subband phonon scattering in quantum confined electron gases based on lattice matched InGaAs/InP quantum wells. The emission of longitudinal acoustic (LA) phonons from two, one and zero dimensional electron gases are compared by studying the scattering times as a function of the width of the lateral confinement. Longitudinal optical (LO) phonon scattering in quantum wells and wires are discussed using a phenomenological broadening of the 1 D density of states. The energy relaxation rates of heated electron gases due to phonon emission and absorption are calculated for lattice temperatures T_l between 0.3 and 20K as a function of the electron density. In the 1 D systems the characteristic quantities (scattering rates τ^{-1}, mean LA phonon energies $\langle E_{ph} \rangle$, energy relaxation rates P) exhibit oscillations around their corresponding 2D values, which reflect the 1 D density of states. On the other hand, the LA phonon scattering in 0D systems becomes increasingly quenched with increasing quantization energies.

homogeneous and static electric fields under physical conditions where impact ionization influences the carrier dynamics. A comparison with theoretical and experimental data from the literature is also given.

2-11 Ambipolar Perpendicular Transport in a Semiconductor Slab: Influence of Optical Phonons, K. Scheller, T. Held, and G. Mahler, *Universität Stuttgart, West Germany.* We investigate the perpendicular transport of an optically excited carrie plasma in a thin semiconductor slab. The excitation is done by a monochromatic laser, homogeneously illuminating one surface. A kinetic description by means of the Boltzmann equation is used. On length scales greater than the absorption length and on time scales greater than the carrier-carrier scattering time the distribution function is approximated by heated displaced Maxwellians. We focus on the role of optical phonons for the transport and relaxation properties of the plasma.

2-12 Phonon scattering and energy relaxation in two, one and zero dimensional electron gases, U. Bockelmann and G. Bastard, *Laboratoire de Physique de la Matière Condensée de l'ENS Paris, France.* We report on calculations of intra- and inter-subband phonon scattering in quantum confined electron gases based on lattice matched InGaAs/InP quantum wells. The emission of longitudinal acoustic (LA) phonons from two, one and zero dimensional electron gases are compared by studying the scattering times as a function of the width of the lateral confinement. Longitudinal optical (LO) phonon scattering in quantum wells and wires are discussed using a phenomenological broadening of the 1 D density of states. The energy relaxation rates of heated electron gases due to phonon emission and absorption are calculated for lattice temperatures T_l between 0.3 and 20K as a function of the electron density. In the 1 D systems the characteristic quantities (scattering rates τ^{-1}, mean LA phonon energies $\langle E_{ph} \rangle$, energy relaxation rates P) exhibit oscillations around their corresponding 2D values, which reflect the 1 D density of states. On the other hand, the LA phonon scattering in 0D systems becomes increasingly quenched with increasing quantization energies.

2-13 Reduction of the conductance fluctuation amplitude in mesoscopic GaAs/GaAlAs heterojunctions due to spin-orbit scattering, M. W. Keller, O. Millo, S. J. Klepper, S. Xiong, A. D. Stone, and D. E. Prober, *Department of Applied Physics, Yale University, New Haven, CN 06520,* and R. N. Sacks, *United Technologies Research Center, East Harford, CN 06108.* We have studied weak-localization and conductance fluctuations in mesoscopic two-dimensional GaAs/GaAlAs heterojunctions for T=7 K to 60 mK. The weak localization data show that the spin-orbit scattering rate exceeds the inelastic rate below ~ 2K. At the same time, we find a significant reduction in the conductance fluctuation amplitude below ~ 2K, as compared to that extrapolated from the high temperature region, where the expected $T^{-1/2}$ behavior is observed. Our results agree well with calculations for the effect of spin-orbit scattering on the universal conductance fluctuation amplitude. The effect of spin-orbit scattering on the magnetic correlation range is also discussed.

2-14 Sensitivity of conductance fluctuations in mesoscopic devices to individual elastic scatterers, S. J. Klepper, O. Millo+, M. W. Keller, and D. E. Prober, *Department of Applied Physics, Yale University, New Haven, CN 06520**, and R. N. Sacks, *United Technologies Research Center, East Harford, CN 06108.* We have added elastic scatterers in a controlled fashion to mesoscopic wires fabricated from an AlGaAs/GaAs heterojunction, through the photoionization of DX centers in the AlGaAs. These devices are in the diffusive regime. The attached leads are optically masked, so that the Fermi parameters of the 2D electron gas do not change during infrared (IR) illumination. We are able to resolve switching events in the device conductance due to the addition of single scatterers. By studying the evolution of the conductance fluctuation characterisitic "magneto-fingerprint" of a sample between successive IR illuminations, we are also able to observe the statistically-averaged effect on device conductance of adding a number of scattering sites. From this we can determine the rms contribution of a single elastic scatterer. Our results for the amplitude and termperature dependence of these conductance changes are consistent with theoretical predictions.[1]

* Supported by NSF DMR 8505539

+ Weizmann Fellow

1. S. Feng et al., Phys. Rev. Lett. 56, 1960 (1986)

NATO Participation in Advanced Study Institute

Granular Nanoelectronics – ASI 890650

Lecturers

Denmark

Antti-Pekka Jauho
H. C. Oersted Institute
University of Copenhagen
DK-2100 Copenhagen, Denmark

France

Prof. J. P. Launay
Centre d'Elaboration de Materiaux et d'Etudes Structurales
Laboratoire d'Optique Electronique
CNRS
29, Rue Jeanne Marvig
31055 Toulouse Cedex, France

West Germany

Prof. Gerhard Abstreiter
Walter Schottky Institut
Technische Universität München
Am Coulombwall
D-8046 Garching, West Germany

Prof. Erich Gornik
Walter Schottky Institut
Technische Universität München
Am Coulombwall
D-8046 Garching, West Germany

Prof. Jörg Kotthaus
Sektion Physik
der Ludwig–Maximilians–Universität München
Geschwister–Scholl–Platz 1
D-8000 München 22, West Germany

Italy

Prof. Carlo Jacoboni
Dipartimento di Fisica degli Universita Modena
Via Campi 213/A
41100 Modena, Italy

Prof. Lino Reggiani
Dipartimento di Fisica degli Universita Modena
Via Campi 213/A
41100 Modena, Italy

The Netherlands

Dr. Carlo W. J. Beenakker
Philips Research Laboratories
P. O. Box 80000
5600 Ja Eindhoven, The Netherlands

Prof. L. J. Geerligs
Faculty of Applied Physics
T. U. Delft
P. O. Box 5046
2600 GA Delft, The Netherlands

United Kingdom

Prof. John R. Barker
Department of Electronics and Electrical Engineering
University of Glasgow
Glasgow G12 8QQ, United Kingdom

Prof. Steve Beaumont
Department of Electronics and Electrical Engineering
University of Glasgow
Glasgow G12 8QQ, United Kingdom

Dr. Trevor Thornton
Department of Electrical Engineering
Imperial College
Exhibition Road
London SW7 2BT, United Kingdom

United States

Prof. Lex A. Akers
Center for Solid State Electronics Research
Arizona State University
Tempe, AZ 85287-6206

Dr. Markus Büttiker
IBM Watson Research Laboratory
P. O. Box 218
Yorktown Heights, NY 10598

Prof. David K. Ferry
Department of Electrical Engineering
Arizona State University
Tempe, AZ 85287-5706

Dr. Alan Fowler
IBM Watson Research Laboratory
P. O. Box 218
Yorktown Heights, NY 10598

Prof. Karl Hess
The Beckman Institute
University of Illinois
Urbana, Illinois 61801

Dr. Gerald J. Iafrate
Electronics Technology and Devices Laboratory
Army Electronics Command
Fort Monmouth, N. J. 07703-5302

Prof. Tony Leggett
Department of Physics
University of Illinois
Urbana, Illinois 61801

Prof. Robert O'Connell
Department of Physics
Lousiana State University
Baton Rouge, LA 70803-4001

Dr. Mark Reed
Department of Electrical Engineering
Yale University
P. O. Box 2157, Yale Station
New Haven, CT 06520-2157

Prof. Aaron Szafer
Deparment of Physics
Yale University
P. O. Box 2157
New Haven, CN 06520

Students

Belgium

F. Geerinckx
Department of Physics
Universitaire Instelling Antwerpen
Universiteitsplein 1
B-2610 Antwerpen (Wilrijk), Belgium

Denmark

Dr. Rita Bertoncini
H. C. Oersted Insitute
University of Copenhagen
DK-2199 Copenhagen, Denmark

Mr. Thomas Fiig
H. C. Oersted Insitute
University of Copenhagen
DK-2199 Copenhagen, Denmark

Mr. Per Hyldgaard
H. C. Oersted Insitute
University of Copenhagen
DK-2199 Copenhagen, Denmark

France

Mr. Ulrich Bockelmann
Laboratoire de Physique
de la matiere condense
de l'Ecole Normale Superieure
24 rue Lhomond
F-75005 Paris, France

Mr. Pierre Tabourier
Centre Hyperfrequences et Semiconducteurs
Universite des Sciences et Techniques, Bt. P3
59655 Villeneuve d'Ascq Cedex, France

Dr. Francis Therez
Laboratoire d'Automatique et d'Analyse des Systemes
CNRS
7, Avenue du Colonel–Roche
31077 Toulouse Cedex, France

West Germany

Dipl.–Phys. Gunther Berthold
Walter Schottky Institut
Technische Universität München
Am Coulombwall
D-8046 Garching, West Germany

Dr. Ernst H. Böttcher
Institut für Festkorperphysik
Fachbereich 4, Physik
Technische Universität Berlin
Hardenbergstrasse 36
D-1000 Berlin 12, West Germany

Dipl.–Phys. Wolfgang Demmerle
Walter Schottky Institut
Technische Universität München
Am Coulombwall
D-8046 Garching, West Germany

Dipl.–Phys. Axel Emunds
Institut für Physik
Universität Aachen
Aachen, West Germany

Dipl.-Phys. P. Grambow
Max–Planck–Institut für Festkörperforschung
Heisenberstrasse 1
D-7000 Stuttgart 80, West Germany

Dipl.-Phys. Albert P. Haberle
Max–Planck–Institut für Festkörperforschung
Heisenberstrasse 1
D-7000 Stuttgart 80, West Germany

Dipl.–Phys. Thomas Held
Institut für Theoretische Physik
Universität Stuttgart
7000 Stuttgart 80, West Germany

Dr. Guido Hüpper
Institut für Theoretische Physik
Technical Universität Berlin, Sekt. PN7-1
Hardenbergstrasse 36
D-1000 Berlin 12, West Germany

Dipl.–Phys. Thomas Schäpers
Institute of Thin Film and Ion Technology
Forschungscentrum Jülich GmbH
Postfach 1913
D-5170 Jülich, West Germany

Dipl.–Phys. Klaus Scheller
Institut für Theoretische Physik
Universität Stuttgart
7000 Stuttgart 80, West Germany

Dipl. Ing. Andreas Schüppen
Kernforschungsanlage Jülich
Postfach 1913
D-5170 Jülich, West Germany

Dipl.-Phys. Jürgen Scriba
Sektion Physik
der Ludwig–Maximilians–Universität München
Geschwister–Scholl–Platz 1
D-8000 München 22, West Germany

Hans Zappe
Fraunhofer IAF
Tullastrasse 72
7800 Freiburg, West Germany

Italy

Dr. Paolo Bordone
Dipartimento di Fisica
Univsersita degli Studi di Modena
Via Campi 213/A
41100 Modena, Italy

Dr. Rosella Brunetti
Dipartimento di Fisica
Univsersita degli Studi di Modena
Via Campi 213/A
41100 Modena, Italy

Dr. Adriano Cola
Dipartimento di Scienza dei Materiali
Universita degli studi di Lecce
73100 Lecce, Italy

Mr. Paolo Colpani
Dipartimento di Fisica
Univsersita degli Studi di Modena
Via Campi 213/A
41100 Modena, Italy

Dr. Tilmann Kuhn
Dipartimento di Fisica
Univsersita degli Studi di Modena
Via Campi 213/A
41100 Modena, Italy

Dr. Gabriella Leo
Dipartimento di Scienza dei Materiali
Universita degli studi di Lecce
73100 Lecce, Italy

Dr. Massimo Macucci
S.S.S.U.P.S. Anna
Via Carducci, 40
L-56127 Pisa, Italy

Dr. Patrizia Poli
Dipartimento di Fisica
Univsersita degli Studi di Modena
Via Campi 213/A
41100 Modena, Italy

Dr. Wolfgang Quade
Dipartimento di Fisica
Univsersita degli Studi di Modena
Via Campi 213/A
41100 Modena, Italy

Dr. Fausto Rossi
Dipartimento di Fisica
Univsersita degli Studi di Modena
Via Campi 213/A
41100 Modena, Italy

Dr. Lucio Rota
Dipartimento di Fisica
Univsersita degli Studi di Modena
Via Campi 213/A
41100 Modena, Italy

The Netherlands

Mr. Vladimir Bubanja
Faculteit Technische Natuurkunde
Technische Universiteit Delft
2968 CJ Delft, The Netherlands

Mr. Jan-Theodoor Janssen
High Field Magnet Laboratory, Faculty of Science
University of Nijmegen
Toernooiveld
6525 ED Nijmegen, The Netherlands

Mr. Richard T. M. Smokers
Research Institute for Materials
Faculty of Science
University of Nijmegen
Toernooiveld
6525 ED Nijmegen, The Netherlands

Norway

Mr. Gaute Einevoll
Institutt for Fysikk
Universitetet i Trondheim
Sem Sælandsvei 9
N-7034 Trondheim, Norway

Prof. E. H. Hauge
Institutt for Fysikk
Universitetet i Trondheim
Sem Sælandsvei 9
N 7034 Trondheim, Norway

Mr. Jon Andreas Støvneng
Institutt for Fysikk
Universitetet i Trondheim
Sem Sælandsvei 9
N-7034 Trondheim, Norway

Spain

Dr. Fernando Sols
Departamento de Fisica de la Materia Condensada
Universidad Autonoma de Madrid
Cantoblanco
E-28049 Madrid, Spain

Turkey

Mr. Erkman Tekman
Department of Physics
Bilkent University
Bilkent 06533, Ankara, Turkey

United Kingdom

Mr. David Cobden
Semiconductor Physics Laboratory
Cavindish Laboratory
Madingley Road
Cambridge OB3 0HE, United Kingdom

Mr. Geoff Foote
Semiconductor Physics Laboratory
Cavindish Laboratory
Madingley Road
Cambridge OB3 0HE, United Kingdom

Mr. Hiroshi Mizuta
Hitachi Cambridge Research Laboratory
Cavindish Laboratory
Madingley Road
Cambridge OB3 0HE, United Kingdom

Prof. Gwynne James Morgan
Department of Physics
The University of Leeds
Leeds LS2 9JT, United Kingdom

Dr. Yukinori Ochiai
Microelectronics Research Laboratory
Cavindish Laboratory
Madingley Road
Cambridge OB3 0HE, United Kingdom

Dr. Peter E. Selbmann
Department of Electronics and Electrical Engineering
University of Glasgow
Glasgow G12 8QQ, United Kingdom

Dr. David Williams
Hitachi Cambridge Research Laboratory
Cavindish Laboratory
Madingley Road
Cambridge OB3 0HE, United Kingdom

United States

Mr. Ping Ao
Department of Physics
University of Illinois
Urbana, IL 61801

Mr. Raymond Ashoori
Department of Physics, Clark Hall
Cornell University
Ithaca, N.Y. 14853

Mr. Robert E. Carnahan
School of Electrical Engineering
Georgia Institute of Technology
Atlanta, GA 30332

Dr. Larry R. Cooper
Office of Naval Research
800 N. Quincy
Arlington, VA 22217

Prof. Yaotian Fu
Department of Physics
Washington University
One Brookings Drive
St. Louis, MO 63130-4899

Prof. Stephen M. Goodnick
Department of Electrical Engineering
Oregon State University
Corvallis, OR 97331-3211

Prof. Robert O. Grondin
Department of Electrical Engineering
Arizona State University
Tempe, AZ 85287-5706

Dr. Harold L. Grubin
Scientific Research Associates, Inc.
P. O. Box 1058
Glastonbury, CN 06033

Dr. Anne Guerrero
U. S. Army E.T.&D. L. Laboratory
Fort Monmouth, N. J. 07703

Dr. James Harvey
U. S. Army E.T.&D.L. Laboratory
Fort Monmouth, N. J. 07703

Mr. Mark Keller
Applied Physics, Becton Center
P. O. Box 2157, Yale Station
New Haven, CN 06520-2157

Mr. Steven Klepper
Applied Physics, Becton Center
P. O. Box 2157, Yale Station
New Haven, CN 06520-2157

Dr. Al Kriman
Center for Solid State Electronics Research
Arizona State University
Tempe, AZ 85287-6206

Mr. David Loss
Department of Physics
University of Illinois
Urbana, Illinois 61801

Mr. Kieran Mullen
Department of Physics
University of Illinois
Urbana, Illinois 61801

Mr. Joe Nucci
Center for Solid State Electronics Research
Arizona State University
Tempe, AZ 85287-6206

Mr. Changsoo Park
Department of Physics
Washington University, One Brookings Drive
St. Louis, MO 63130-4899

Mr. Vadim Pevzner
Computational Electronics Group
The Beckman Institute
University of Illinois
Urbana, IL 61801

Prof. Umberto Ravaioli
Computational Electronics Group
The Beckman Institute
University of Illinois
Urbana, IL 61801

Mr. Curt A. Richter
Applied Physics, Becton Center
P. O. Box 2157, Yale Station
New Haven, CN 06520-2157

Mr. Joseph Ryan
Center for Solid State Electronics Research
Arizona State University
Tempe, AZ 85287-6206

Prof. Sergio E. Ulloa
Department of Physics and Astronomy
Ohio University
Athens, Ohio 45701-2979

Mr. Toshishige Yamada
Center for Solid State Electronics Research
Arizona State University
Tempe, AZ 85287-6206

Non–NATO Participation in Advanced Study Institute

Granular Nanoelectronics – ASI 890650

Lecturers

Austria

Dr. Friedemar Kuchar
Ludwig Boltzmann Institut für Festkorperphysik
Kopernikusgasse 15
A-1060 Wien, Austria

Japan

Prof. Chihiro Hamaguchi
Department of Electronic Engineering
Osaka University
Suita City, Osaka 565, Japan

Prof. Toshiaki Ikoma
Institute of Industrial Science
University of Toky
7-22-1 Roppongi
Tokyo 106, Japan

Soviet Union

Prof. Konstantin K. Likharev
Laboratory of Cryoelectronics
Department of Physics
Moscow State University
Moscow 119899 GSP, USSR

Students

Austria

Prof. Peter Kocevar
Institut für Theoretische Physik
Karl–Franzens–Universität Graz
A-8010 Graz, Austria

East Germany

Dr. M. Suhrke
Department of Physics
Humboldt University
Berlin, East Germany

Israel

Dr. Joseph Salzman
Technion–Israel Institute of Technology
Haifa 32000, Israel

Soviet Union

Prof. K. Chaplik
Novosibirsk Ac. Sciences
Novosibirsk, USSR

Dr. G. Gusev
Institute of Semiconductor Physics
Academy of Sciences of the USSR
630090 Novosibirsk, USSR

Dr. G. Kuon
Institute of Semiconductor Physics
Academy of Sciences of the USSR
630090 Novosibirsk, USSR

Prof. A. M. Vasil'ev
A.F. Ioffe Physical Technical Institute
Academy of Sciences of the USSR
Polytechnichescaja 26
Leningrad 194021 USSR

Sweden

Mr. Magnus Persson
Department of Physics 4
Chalmers University of Technology
S-41296 Göteborg, Sweden

(1)

AN INTRODUCTION TO CHARGE QUANTUM TRANSPORT IN SEMICONDUCTORS AND NUMERICAL APPROACHES

Fausto Rossi, Rossella Brunetti, and Carlo Jacoboni
Dipartimento di Fisica dell' Università, Via Campi 213/A, I-41100 Modena, Italy

INTRODUCTION TO QUANTUM TRANSPORT

Since several years it has been recognized that the features of the new physical systems provided by present-day technology require a quantum treatment of electron transport. For recent reviews on the subject see Grubin et al., 1988 and Ferry et al., 1990.

We shall not repeat here the reasons for such a belief. We shall rather concentrate on an introduction to the basic ideas of quantum transport theory and to a brief account of some of the efforts that are presently being made to obtain numerical results for an estimate of the main quantum effects or, at least, for a test of new numerical approaches. In fact, even though the main theoretical concepts and methods of quantum transport have been developed long ago, they hardly yielded any numerical result to compare with experimental data, owing to the complexity of the mathematics involved and to the lack of clear experimental evidences.

In the theory of quantum transport, many concepts and techniques are used, such as density matrix, Green functions, path integrals and Wigner function, which may be not familiar to the non-experts, and for this reason we shall try to give in this section a quick summary of such concepts. In particular, we shall enphasize their differences, similarities, and mutual relationships. In the following sections these concepts will be somewhat developed and some numerical applications will be presented, without any attempt of completeness.

The quantum description of a physical system is given, in the Schrödinger picture, by the state vector

$$|\Phi(t)\rangle \tag{1}$$

as a function of time t.

If the problem is not completely specified, we have to use the concepts of statistical physics to deal with our incomplete knowledge, and the mathematical instrument in this case is the density-matrix operator (see, for example Ter Haar,

1961):

$$\rho(t) = \overline{|\Phi(t)\rangle\langle\Phi(t)|} \tag{2}$$

where the overbar indicates an average to be performed over a suitable statistical ensemble that accounts for our partial knowledge of the system.

In a given basis $|\phi_n\rangle$ the density-matrix operator has the following matrix elements

$$\rho_{nm} = \langle\phi_n|\overline{\Phi}\rangle\langle\overline{\Phi}|\phi_m\rangle = \overline{c_n c_m^*} \tag{3}$$

where c_n are the coefficients of $|\Phi\rangle$ in the given basis.

The state vector or the density matrix can be used, for a pure state or an ensemble, respectively, to obtain the average result of a measurement of a quantity A:

$$\langle A\rangle = \langle\Phi|A|\Phi\rangle \tag{4}$$

or

$$\langle A\rangle = \overline{\langle\Phi|A|\Phi\rangle} = \sum_{n,m} \overline{c_n^* A_{nm} c_m} = Tr(\rho A)\ . \tag{5}$$

If the system under examination can be devided into a "small" part of interest with coordinates x and a "large" part formed by an external interacting system with coordinates X, and if an observable $A^{(x)}$ acts only on the variables x, the mean value of $A^{(x)}$ is given by

$$\langle A^{(x)}\rangle = Tr(\rho A^{(x)}) = \sum_{x,X} \rho(x, X, x', X) A^{(x)}(x', x)$$

or

$$\langle A^{(x)}\rangle = Tr(\rho^{(x)} A^{(x)}) \tag{6}$$

where

$$\rho^{(x)}(x, x') = \sum_X \rho(x, X, x', X) = Tr_X(\rho) \tag{7}$$

is the reduced density matrix.

The Wigner function (Wigner, 1932) is a special transform of the density matrix in the coordinate representation. In a one-dimensional case, for example,

$$fw(x, k, t) = \frac{1}{2\pi} \int dy e^{iky} \rho(x - \frac{1}{2}y, x + \frac{1}{2}y, t) \tag{8}$$

Therefore the Wigner function carries the same information as the density matrix.

In order to solve a quantum problem with given initial conditions for the state vector or the density matrix, it is necessary to find the evolution operator $U(t, t_i)$ from the initial time t_i to the actual time t. In the Schrödinger picture, it evolves the state vector according to

$$|\Phi(t)\rangle = U(t, t_i)|\Phi(t_i)\rangle \tag{9}$$

and, therefore, the density matrix according to

$$\rho(t) = U(t,t_i)\rho(t_i)U^{\dagger}(t,t_i). \tag{10}$$

If $\{q\}$ represents a complete set of compatible dynamical variables for our system, Eq.(9) can be written in the $\{q\}$ representation as

$$\Phi(q,t) = \int U(q,q',t,t_i)\Phi(q',t_i)dq', \tag{11}$$

where $U(q,q',t,t_i)$ is the matrix element of the evolution operator $U(t,t_i)$ in the basis $\{q\}$.

The fundamental dynamical equation is then a differential equation for the evolution operator:

$$i\hbar\frac{d}{dt}U(t,t_i) = HU(t,t_i), \tag{12}$$

with the initial condition

$$U(t_i,t_i) = 1, \tag{13}$$

where H is the Hamiltonian of the system.

The Feynman path-integral theory (see Feynman and Hibbs, 1965) is an alternative approach to quantum mechanics. It starts from the idea that all possible paths of a system from an initial state to a final one are to be considered as simultaneously realized, and their amplitudes add up, rather than their probabilities as it would be in classical concepts, to give the probability amplitude of finding the system in the final state. The fundamental equation in this approach is an explicit expression for the evolution operator as an integral over all possible paths of the exponential of the classical action. If q represents a set of classical Lagrangian variables of the system, and $q(\tau)$ one given trajectory, or path, from the initial values $q_i = q(t_i)$ to the final values $q_f = q(t)$, the evolution operator in Eq.(11) can be written as

$$U(q_f,q_i,t,t_i) = \int_{q_i,t_i}^{q_f,t} \mathcal{D}q(\tau)\; e^{\frac{i}{\hbar}S[q(\tau)]}. \tag{14}$$

Here $\mathcal{D}q(\tau)$ indicates the integral over all paths that connect the initial state $\{q_i,t_i\}$ to the final state $\{q_f,t\}$ and $S[q(\tau)]$ is the classical action evaluated over each given trajectory in the integral:

$$S[q(\tau)] = \int_{t_i}^{t} L(q(\tau),\dot{q}(\tau),\tau)d\tau, \tag{15}$$

where $L(q,\dot{q},\tau)$ is the classical Lagrangian of the system.

If a path-integral approach is followed and the integral in Eq.(14) evaluated, the resulting evolution operator can then be used to evaluate the evolution of a state wavefunction (as in Eq.(9) or (11)), or of the density matrix (as in Eq.(10)).

As it regards the Green functions, they carry information on both the state of the system and its dynamical evolution (see, for example, Kadanoff and Baym,

1962; Rickayzen, 1980; Mahan, 1981). Several quantities are often indicated with this name, and they are somewhat different, even though strictly related to each other. Some of them are given by an average value of a commutator of field operators times a step function θ. More precisely, the single-particle retarded and advanced Green functions $G^{(r)}$ and $G^{(a)}$ are defined as

$$G^{(r)}(\mathrm{r},t,\mathrm{r}',t') = \frac{-i}{\hbar}\langle\Phi_{\mathcal{H}}|[\Psi_{\mathcal{H}}(\mathrm{r},t),\Psi^{\dagger}_{\mathcal{H}}(\mathrm{r}',t')]_{\mp}|\Phi_{\mathcal{H}}\rangle\theta(t-t') \tag{16}$$

$$G^{(a)}(\mathrm{r},t,\mathrm{r}',t') = \frac{i}{\hbar}\langle\Phi_{\mathcal{H}}|[\Psi_{\mathcal{H}}(\mathrm{r},t),\Psi^{\dagger}_{\mathcal{H}}(\mathrm{r}',t')]_{\mp}|\Phi_{\mathcal{H}}\rangle\theta(t'-t) \tag{17}$$

where $|\Phi_{\mathcal{H}}\rangle$ is the state under consideration in the Heisenberg picture, $\Psi_{\mathcal{H}}(\mathrm{r},t)$ is the field operator in the same picture, and θ is the Heaviside step function. The upper sign is to be used for Bosons and the lower sign for Fermions. An ensemble average must be included in case of statistical problems.

In order to grasp the physical meaning of such functions let us recall that the expectation value of the product $\Psi^{\dagger}_{\mathcal{H}}(\mathrm{r},t)\Psi_{\mathcal{H}}(\mathrm{r},t)$ at equal positions and times gives the intensity of the field in the state under consideration. The same product at different arguments gives, in the same way, the correlation between the two amplitudes at different positions and times. If we look for the dynamical correlation, that is the propagator, without the information on the field intensity, we must subtract the product in reverse order. We may compare this result with the more familiar relations for harmonic-oscillator creation, annihilation, and number operators:

$$a^{\dagger}a = N, \qquad aa^{\dagger} = N-1, \qquad \textit{so that} \qquad [a,a^{\dagger}] = 1 \tag{18}$$

Thus, the two Green functions above carry the information on which probability amplitude for the presence of a particle in r at t corresponds to a unit amplitude in r' at t'. They are, therefore, the equivalent of a propagator, or of an evolution operator. They include the fact, however, that they are single-particle operators defined in a many-particle system, and therefore they contain a reduction of many degrees of freedom and include the effect, on each particle, of the interaction with all the other ones in the system.

The average values of the single products of field operators discussed just now are themselves defined as other Green functions. More precisely, $G^{>}$ and $G^{<}$ are

$$G^{>}(\mathrm{r},t,\mathrm{r}',t') = \frac{-i}{\hbar}\langle\Psi_{\mathcal{H}}(\mathrm{r},t)\Psi^{\dagger}_{\mathcal{H}}(\mathrm{r}',t')\rangle \tag{19}$$

$$G^{<}(\mathrm{r},t,\mathrm{r}',t') = \frac{\mp i}{\hbar}\langle\Psi^{\dagger}_{\mathcal{H}}(\mathrm{r}',t')\Psi_{\mathcal{H}}(\mathrm{r},t)\rangle \tag{20}$$

Clearly, the following relations hold:

$$G^{(r)} = \theta(t-t')(G^{>}-G^{<}) \tag{21}$$

$$G^{(a)} = -\theta(t'-t)(G^{>}-G^{<}) \tag{22}$$

From what we have seen above it is understandable that $G^{<}$ is a correlation function of the field amplitudes at (r,t) and (r',t'); thus, for equal times it corresponds to a reduced density matrix for a single particle.

A Wigner function can therefore be defined starting from $G^<$ (Mahan, 1981) which generalizes the definition given in Eq.(8) to a single-particle Wigner function for a many-body system.

Other Green functions can be defined for two or more particles. For such Green functions, including the single-particle ones, there are no simple equations of motion since the propagation of a single particle depends on the entire many-body system. In fact it is possible to define a set of hierarchical equations of motion, where the equation for the single-particle Green function contains also the two-particle Green function; the equation for the two-particle Green function contains the three-particle Green function, and so on.

THE MONTE CARLO DENSITY-MATRIX APPROACH

The time evolution of the density matrix in Eq.(10) and the fundamental equation of motion (12) yield immediately the Liouville-von Neumann differential equation for the evolution of the density matrix

$$i\hbar\frac{d\rho}{dt} = [H, \rho]. \tag{23}$$

As an example of application of the density-matrix approach we briefly summarize the work performed by the Modena group (Brunetti et al., 1989; Menziani et al., 1989; Rossi and Jacoboni, 1989) for the solution of the Liouville–von Neumann equation for the electronic density matrix in semiconductors. In principle, the method allows to evaluate the electronic density matrix as a function of time without any assumptions on the intensity and the duration of the electron-phonon interaction, as well as on the strength of the applied field.

As starting point, let us consider a noninteracting electron gas in a semiconductor crystal, coupled to the phonon gas and to a constant and uniform electric field $\mathbf{E}$. The system is assumed to be homogeneous, and its Hamiltonian is given by

$$H = H_e + H_E + H_p + H_{ep} \tag{24}$$

where H_e is the term corresponding to an electron in a perfect crystal, $H_E = e\mathbf{E}\cdot\mathbf{r}$ is the term due to the electric field, and H_p is the Hamiltonian of the free phonons; the electron-phonon interaction Hamiltonian H_{ep} has the general form

$$H_{ep} = \sum_q i\hbar F(\mathbf{q})\{a_q e^{i\mathbf{qr}} - a_q^\dagger e^{-i\mathbf{qr}}\} = H_{ab} + H_{em} \tag{25}$$

where H_{ab} and H_{em} refer to phonon absorption and emission respectively, and $F(\mathbf{q})$ is a function of the phonon momentum $\mathbf{q}$ whose explicit form depends on the particular interaction mechanism.

Let us consider the set of time-dependent basis vectors $\mid \mathbf{k}_o, \{n_\mathbf{q}\}, t\rangle$ represented by

$$\frac{1}{\sqrt{V}} e^{i\mathbf{k}(t)\cdot\mathbf{r}} e^{-i\int_0^t d\tau\omega[\mathbf{k}(\tau)]} \mid \{n_\mathbf{q}\}, t\rangle, \tag{26}$$

where $\mathrm{k}(t) = \mathrm{k}_\circ - \frac{e\mathbf{E}}{\hbar}t$, and $\omega(\mathrm{k})$ represents the electronic band structure. They are direct products of electronic accelerated plane waves, or Houston waves (Houston, 1940), normalized to 1 over the crystal volume V, and the phonon states $|\{n_\mathbf{q}\}, t\rangle$.

The Liouville-von Neumann equation that describes the time evolution of the density matrix ρ of the system in the representation of the set in Eq.(26) contains only the perturbation Hamiltonian:

$$i\hbar\frac{\partial}{\partial t}\rho(x,x',t) = [H_{ep}, \rho(t)]_{x,x'}, \tag{27}$$

where the symbolic compact notation $x = (\mathrm{k}_\circ, \{n_\mathbf{q}\})$ has been used. Eq.(27) has the same form we would have obtained in an interaction picture, even though the basis functions (26) are not eigenstates of the unperturbed hamiltonian. This is so because the Houston waves are solution of the time-dependent Schrödinger equation with the unperturbed Hamiltonian that includes the electric field.

If one is interested in the evaluation of expectation values of electron quantities which are diagonal in the electronic part of the states in Eq.(26) only the diagonal elements $\rho(x,t) \equiv \rho(x,x,t)$ of ρ must be determined. Furthermore, a diagonal initial condition for ρ decoupled in electron and phonon coordinates has been assumed.

After a formal integration of Eq.(27), a perturbative expansion for ρ is easily obtained by iterative substitutions:

$$\begin{aligned}\rho(x,t) &= \rho(x,0) + \int_0^t dt_1 [\mathcal{H}_{ep}, \rho(0)]_{t_1;x,x} \\ &+ \int_0^t dt_1 \int_0^{t_1} dt_2 [\mathcal{H}_{ep}, [\mathcal{H}_{ep}, \rho(0)]_{t_2}]_{t_1;x,x} + \cdots \\ &= \rho^{(\circ)}(x,t) + \Delta\rho^{(1)}(x,t) + \Delta\rho^{(2)}(x,t) + \ldots ,\end{aligned} \tag{28}$$

where $\mathcal{H}_{ep} = \frac{1}{i\hbar}H_{ep}$.

The trace over phonon coordinates can be performed exactly under the assumption that each phonon mode is involved only once during the transport process, which is equivalent to neglect hot-phonon effects.

A diagrammatic representation of Eq.(28) allows us to regard each term of the perturbative expansion as a sequence of quantum processes which correspond to single scattering events in semiclassical transport. This constitutes the starting point of the numerical quantum Monte Carlo (QMC) algorithm devised for the solution of the Liouville-von Neumann equation, which is based on random generations of all possible sequences of processes associated to the different perturbative corrections, in the same way as the usual classical Monte Carlo (CMC) generates semi-classical scattering events.

Fig.1 shows an example of a seventh order diagram, which includes also the electron-impurity interaction discussed below. It is worthwhile stressing that this QMC technique is based on a perturbative expansion of the density matrix exactly as the standard CMC technique is based on a perturbative expansion of the distribution function: a path with n scattering events in CMC is a term of the

n-th order in a perturbative expansion of the Boltzmann equation in powers of the scattering rates (Poli et al., 1989).

In what follows we present some typical results obtained with this QMC procedure for different materials and physical conditions. The analysis has been devoted to very short times after the initial conditions, when quantum features are expected to be more relevant. This choice allows to include only few terms of the perturbation expansion given in Eq.(28), which, in turn, limits the computer time to affordable values.

Special emphasis has been given to the investigation of typical quantum effects, such as "intracollisional field effect" (ICFE) and the "collisional broadening", by comparison with the semiclassical case.

A. Analysis of the electronic density matrix in presence of electric fields

Numerical results have been obtained for the diagonal part of the electronic density matrix in presence of an arbitrary high electric field, starting from equilibrium conditions for the electron and phonon systems. Both electric field and electron-phonon coupling due to nonpolar optical phonons, are turned on at t=0. A simple-model semiconductor modeled on silicon with a single spherical and parabolic band, and carriers interacting with one optical-phonon mode has been used. The working conditions are T=20 K, E=150 kV/cm, t=50fs.

Fig. 2 shows the quantum distribution function as a function of k along the field direction compared with the corresponding classical result obtained from the perturbative expansion of the Boltzmann equation by means of a "backward Monte Carlo" (BMC) procedure (Jacoboni et al., 1988). The distribution functions are peaked around the ballistic value. The particles that are scattered out of the ballistic trajectories are spread in a large volume of k space and cannot be seen in the figure. The quantum distribution is lower than the classical one. It can be shown that this correction is the net result of two larger quantum effects of opposite signs. In fact, if the effect of the field during the collision is neglected, we may obtain a big effect of non energy-conserving transitions, due to the short time considered. When the ICFE is added, it reduces the scattering efficiency by reducing the time of positive interference which occurs when the energy difference between initial and final states is equal to the phonon energy.

B. Transient analysis of electron–impurity interaction

The theoretical approach discussed before has been also extended to the analysis of electron–impurity interaction, adding to the Hamiltonian in Eq.(24) an interaction term given by

$$H_{ei} = \sum_{i=1}^{N} \varphi(\mathbf{r} - \mathbf{r}_i), \tag{29}$$

where $\mathbf{r}_1, \mathbf{r}_2 \ldots \mathbf{r}_N$ describe a given impurity configuration within the crystal volume V, and $\varphi(\mathbf{r} - \mathbf{r}_i)$ is the interaction potential due to the i-th impurity.

The matrix elements between $\mathbf{k}_0$ and $\mathbf{k}'_0$ of the electron-impurity Hamiltonian depend on the impurity coordinates as

$$\sum_{i=1}^{N} e^{-i(\mathbf{k}_0-\mathbf{k}'_0)\cdot\mathbf{r}_i}, \tag{30}$$

and in order to proceed in the calculation an ensemble average over all the impurity configurations $\{\mathbf{r}_1, \mathbf{r}_2 \ldots \mathbf{r}_N\}$ is to be taken. These matrix elements have been incorporated into the QMC procedure described above. The diagonal part of the electronic density matrix has been evaluated for the same simplified silicon model used previously, and the electron-impurity interaction is taken as a screened Coulomb potential. In Fig. 3 the quantum distribution function up to the fourth order is shown as a function of k along the field direction, compared with the corresponding classical result obtained from a BMC simulation (Jacoboni et al., 1988). The distribution functions are peaked around the ballistic value. Here the quantum curve is higher than the classical one and this fact is again the net result of two larger quantum effects of opposite signs: on one hand, ICFE for a given transition (from k to k') in momentum space tends to decrease the scattering efficiency; on the other hand, energy-non conserving transitions, allowed at so short times, may increase the number of the available final states. In this case, the prevailing effect is the first one since the particular form of the electron-impurity interaction itself favours low momentum-transfer and, therefore, energy-conserving transitions.

Higher-order terms contain processes involving more than one scattering center (corresponding to several semiclassical scattering events), and processes involving more than two vertices related to the same scattering center (corresponding to higher-order corrections to the Born approximation).

C. Quantum energy relaxation of photoexcited carriers

The present QMC method has been also applied to the case of photoexcited electrons in bulk GaAs. The semiconductor model has been simplified to a single spherical and parabolic band similar to the one used before for Si and the interaction Hamiltonian includes only polar coupling to optical phonons.

Electrons are generated at $t = 0$ according to a distribution proportional to $exp\{-\alpha|\epsilon - \epsilon_o|\}$, where ϵ is the electron energy, and α is an appropriate constant.

For the sake of clarity we may consider first results of the classical theory. Fig. 4 shows the results obtained with CMC at $t = 100fs$ after excitation. The highest peak at 1000 K represents what is left of the initial distribution at $t = 0$. Two secondary peaks are clearly seen, corresponding to electrons having emitted one or two optical phonons. Fig. 5 shows the corresponding result obtained with the present QMC (note the scale change). The initial distribution is diminished of a quantity very similar to that of the classical case. However electrons can be found, at $t = 100fs$, in a very wide range of energies, since energy needs not be conserved. The secondary peaks are not yet well formed. If we go towards longer times the secondary peaks appear also in the quantum result.

D. Quantum analysis of drift velocity overshoot in GaAs

Finally, the QMC procedure has been applied to study the drift velocity overshoot in GaAs.

A simplified semiconductor model has been used, consisting of a Γ valley and four equivalent L valleys. The electron-phonon interaction is given by two scattering mechanisms: intra-valley polar optical and intervalley nonpolar optical phonons. The value of the electric field has been taken as 40 kV/cm, and this choice is based on the analysis of the overshoot in classical terms.

In Fig. 6 the drift velocity, as given by a traditional CMC simulation, is been compared with its corresponding perturbative solution up to the second order in the scattering rates (i.e. the ensemble of trajectories with up to two scattering events). We can see that this perturbative solution gives us a good approximation of the drift velocity peak. Therefore we expect to be able to describe velocity overshoot in quantum terms with the perturbative expansion described above up to the fourth order in the hamiltonian.

In fact, in the same figure the corresponding quantum drift velocity up to the forth order has been shown: it can be seen that under these conditions, quantum features are not very relevant. Because of energy-non-conserving transitions one could expect that the drift-velocity overshoot would be reduced in quantum theory, owing to an anticipation of the intervalley transitions. This does not occur however since the coupling constant is too weak.

THE WIGNER-FUNCTION APPROACH

In a sense the density matrix can be considered the quantum analogue of the classical distribution function, since its diagonal values in a fixed basis give the probabilities of finding a system of the ensemble in the corresponding eigenstate. However, the distribution function $f(q,p)$ for a classical statistical ensemble gives us the probability density to find the system in the position (q,p) of the phase-space. Owing to the incompatibility of q and p, it is not possible to find a function with the same property in quantum mechanics.

In this connection, we may rise the more general problem of defining a numerical function $\mathcal{A}(q,p)$, where q and p are c-numbers, which is the classical analogue of a general quantum operator $A(q,p)$, function of the operators q and p. A requisite could be that the sum of all possible values is the same for both quantities:

$$Tr\{A\} = \frac{1}{2\pi\hbar}\int\int \mathcal{A}(q,p)dqdp. \tag{31}$$

This, however, is not enough to determine $\mathcal{A}$ since only an integrated property of it is given. On the other hand, if we require that the property expressed by Eq.(31) holds true also giving different weights to the different points in the phase-space, then the equation can be inverted. Here we have an arbitrariety in how to

choose these different weights. If Fourier functions are chosen, we obtain the Weyl correspondence rule.

Let us then require that, for any real ξ and η,

$$Tr\{Ae^{i[\xi q/\hbar+\eta p/\hbar]}\} = \frac{1}{2\pi\hbar}\int\int \mathcal{A}(\alpha,\beta)e^{i\xi\alpha/\hbar}e^{i\eta\beta/\hbar}d\alpha d\beta. \tag{32}$$

Now the Fourier transform in the right hand side can be inverted:

$$\mathcal{A}(\alpha,\beta) = \frac{1}{2\pi\hbar}\int e^{-i\xi\alpha/\hbar}d\xi\int e^{-i\eta\beta/\hbar}d\eta\int dx\int dx'\langle x|A|x'\rangle\langle x'|e^{i[\xi q/\hbar+\eta p/\hbar]}|x\rangle. \tag{33}$$

By using

$$e^{i[\xi q+\eta p]} = e^{i\eta p/2}e^{i\xi q}e^{i\eta p/2}, \tag{34}$$

and

$$e^{i\eta p/\hbar}|x\rangle = |x-\eta\rangle, \tag{35}$$

we obtain, after streightforward calculations,

$$\mathcal{A}(\alpha,\beta) = \int e^{i\eta\beta/\hbar}\langle\alpha-\eta/2|A(q,p)|\alpha+\eta/2\rangle d\eta. \tag{36}$$

which is the Weyl correspondence rule.

The Wigner function is now defined as the Weyl transform of the density matrix operator (a factor$(1/2\pi\hbar)$ is added in order to use simple integration over the phase space):

$$f_W(q,p) = \frac{1}{2\pi\hbar}\int e^{i\eta p/\hbar}\langle q-\eta/2|\overline{\Phi}\rangle\langle\overline{\Phi}|q+\eta/2\rangle d\eta,$$

or

$$f_W(q,p) = \frac{1}{2\pi\hbar}\int e^{i\eta p/\hbar}\overline{\phi^*(q+\eta/2)\phi(q-\eta/2)}\ d\eta. \tag{37}$$

This is equivalent to the definition given in Eq.(8)

Several properties of f_W echo its origin of classical analogue of the density matrix. In particular

$$\int f_W(q,p)dp = \overline{|\psi(q)|^2}, \tag{38}$$

$$\int f_W(q,p)dq = \overline{|\varphi(p)|^2}, \tag{39}$$

where $\varphi(p)$ is the wavefunction in p representation. Therefore, if $F(q)$ and $G(p)$ are functions only of q and p, respectively, then

$$\int\int F(q)f_W(q,p)dqdp = \langle F\rangle \tag{40}$$

$$\int\int G(p)f_W(q,p)dqdp = \langle G\rangle \tag{41}$$

More generally, if A is an observable function of q and p, it can be shown that

$$\langle A \rangle = Tr(\rho A) = \int \int \mathcal{A}(q,p) f_W(q,p) dq dp, \tag{42}$$

where $\mathcal{A}$ is the Weyl transform of A.

On the other hand, it is not possible to give to f_W a simple probabilistic interpretation for the reasons indicated at the beginning of this section, and this is confirmed by the fact that f_W takes in general negative, as well as positive, values in phase space.

A physical insight of the meaning of the Wigner function can be obtained by observing that it has large values in the regions of q-space where there is a strong autocorrelation of the wavefunction, and for p values corresponding to the Fourier components present in this correlated parts of the wavefunction. Thus, f_W has large values in regions of phase-space where the presence of the particles can be "felt" within the uncertainty principle, and with some more information related to the quantum mechanical phases, that can result in negative values for f_W.

Concluding this brief introduction to the Wigner function, we note that if the interaction of one particle with the rest of the system can be described by a perturbation potential $V(q)$, an equation for the Wigner function can be written (Toda et al., 1983),

$$\frac{\partial f_W(q,p,t)}{\partial t} + \frac{p}{m}\frac{\partial f_W(q,p,t)}{\partial q} = \frac{1}{i\hbar}[V(q - \frac{\hbar}{2i}\frac{\partial}{\partial p}) - V(q + \frac{\hbar}{2i}\frac{\partial}{\partial p})] f_W(q,p,t) \tag{43}$$

which reduces to the classical Liouville equation for $\hbar \to 0$.

A more explicit form of the above equation can also be given, which is more suitable for application to real systems:

$$\frac{\partial f_W(q,p,t)}{\partial t} + \frac{p}{m}\frac{\partial f_W(q,p,t)}{\partial q}$$

$$= \frac{1}{2\pi\hbar^2}\int dP \int dy e^{\frac{iPy}{\hbar}} [V(x + \frac{1}{2}y) - V(x - \frac{1}{2}y)] f_W(x, p+P, t)\,. \tag{44}$$

In order to extend the above theory to the case where scattering is present, an "ad hoc" collision term $(\frac{\partial f_W}{\partial t})_{coll}$ is added to the rhs of Eq.(44) (Iafrate, 1988), but it may not necessarily express the same phenomenology as the corresponding term in the Boltzmann equation.

Some interesting results can be obtained by considering moments of Eq.(43) in relaxation-time approximation. Equations are obtained which reduce to the moments of the Boltzmann equation in the semiclassical limit, and contain quantum corrections which can be included in the classical picture (Iafrate, 1988).

In what follows we shall report some recent work performed with the WF approach on quantum systems of particular interest in modern microelectronic research.

A. Wigner function applied to the study of resonant-tunneling devices

Several theoretical works devoted to the development of the Wigner formalism for quantum electronic transport have appeared in the literature (Barker, 1980; Barker and Lowe, 1981; Barker and Murray, 1983; Lin and Chiu, 1984; Barker, 1985; Frensley, 1987; Frensley, 1988; Kluksdahl et al., 1987, 1989). As a general comment, however, it can be observed that most of the literature on the WF for transport problems is primarily concerned with the formulation of the problem, rather than its solution. In particular the complicated form of the collision term in the quantum kinetic equations requires either severe approximations or formidable numerical effords.

When quantum ballistic systems are considered collisions with phonons/impurities are substantially reduced by a combination of very high mobility materials and short channels (Barker and Murray, 1983). In this case the complexity of the collision term can be highly reduced by suitable approximations, and the discussion of the quantum features becomes easier.

The first attempt at calculating the WF for an actual physical system has been presented by Barker (1985) and it considers propagation of an incident electron gaussian wave packet on a single quantum well, with two bound states E_1 and E_2, at the center of a very wide barrier, thus forming two adiacent wide barriers.

Fig. 7 shows snapshots of the position probability distribution (a), momentum distribution (b), and Wigner distribution (c) for a gaussian wavepacket incident at resonance when the momentum width of the packet is greater than the resonance peak in the transmission coefficient. The complex central structure in the WF originates from correlation between the reflected and transmitted wave in phase-space.

The same quantum structure has extensively been studied, using the WF, as a resonant-tunneling diode (RTD), (Ravaioli et al., 1985; Kluksdahl et al., 1987; Frensley, 1987; Frensley, 1988; Kluksdahl et al., 1989). The quantum well constitutes a resonant-tunneling system, with a resonant energy marked by preferential tunneling. A common step forward of these works is the inclusion of the role of the contacts. They are described as "ideal" infinite reservoirs of thermally-distributed carriers which act like a source of injected randomly-distributed electrons into the device at one terminal, and a perfect sink absorbing all incident carriers randomizing their state without reflection upon the other terminal. These contacts serve as a boundary for quantum correlations , they remove size dependencies , and introduce time-irreversibility in the dynamical evolution of the carrier system.

Ferry and coworkers (Ravaioli et al., 1985; Kluksdahl et al., 1987; Kluksdahl et al., 1989) consider a self-consistent modelling of the RTD based upon the coupled solution of the WF equation of motion to the Poisson equation, and include electron scattering through a simple relaxation-time approximation. The equations are solved with a finite- difference approximation scheme.

A particular analysis has been devoted to the choice of the initial WF.

Fig.8 shows the WF for a gaussian wave packet interacting with the resonant quantum potential barriers at different times. Quantum interference and tunneling

effects are visualized.

Following this approach the I-V characteristics of the device is obtained by increasing the bias potential to its maximum, and then decrementing it towards zero, with the current being calculated on the way. The resultant curve shows an intrinsic bistability confirmed by the experiments.

For bias conditions near the peak of the I-V curve an overall depletion of electrons in the cathode occurs, as it appears from an analysis of the WF, reported in Fig.9-a for this case. As the bias increases further, the resonant charge becomes evident in the WF, illustrated in Fig. 9-b. In Fig.10 the difference between the steady-state WF for increasing and decreasing potentials in the bistable region is plotted. For decreasing potentials less current flows through the resonant structure, and the injected carriers accumulate on the cathode side of the structure, shown as the peak in Fig.10.

This analysis of the RTD with the use of the WF has been extended to transient regimes (Ravaioli et.al., 1985, Frensley, 1988), the frequency-responce of the RTD (Frensley, 1988), and of particular features, like the anomaly in the I-V curve for very low applied bias (Kluksdahl et al., 1989).

The difficulty in extending the formalism to 3-dimensional systems is mainly related to computer limitations in memory and CPU time...

THE PATH-INTEGRAL APPROACH

There are essentially two different ways of using the path-integral approach for the evaluation of the density matrix. One of them is based on the starting expression for the evolution operator given in Eq.(14). It is called the "real-time path-integral approach", and we shall discuss it some more at length later in this section, reviewing also one significant application.

The second way is called "immaginary-time path-integral approach", and is based on the formal analogy between the analytical form of the equilibrium density-matrix operator,

$$\rho_o \propto e^{-\beta H}, \tag{45}$$

where $\beta = (KT)^{-1}$, and the evolution operator for a time-independent hamiltonian

$$U(t,0) = e^{-\frac{i}{\hbar}Ht} \tag{46}$$

From the comparison of the last two equations it is easy to understand that there can be a formalism (see Feynman and Hibbs, 1965) where the path-integral approach to the evolution operator can be extended to the evaluation of the equilibrium density matrix with paths along an immaginary time proportional to the inverse temperature. Since this method, however, yields only equilibrium properties of the system, we shall not discuss it here in details. We simply mention the work performed by the North Carolina group (Register et. al., 1988b). They evaluate the equilibrium density matrix for some particular potential profiles related to tunnelling problems, adding a stochastic potential to model the effect of the

phonons. Their results suggest, as expected, that the presence of phonons may reduce the tunneling process by reducing the phase coherence of the electrons.

As indicated above, in the real-time path-integral approach the evolution operator as given by Eq.(14) is applied to the density matrix. The resulting expression

$$\rho(q,q',t) = \int dq_i \int dq_i' \int_{q_i,t_i}^{q,t} \mathcal{D}q(\tau) \int_{q_i',t_i}^{q',t} \mathcal{D}q'(\tau) e^{\frac{i}{\hbar}[S(q(\tau))-S(q'(\tau))]} \rho(q_i,q_i',t_i) \tag{47}$$

may be elaborated in a useful way by factorizing the effect of "external agents" with respect to the system of interest. In our case we call x the variables of the "small" system of interest (for example an electron) and X the variables of the interacting "large" system (for example the phonon bath). The exponential in Eq.(47) can then be factorized as follows:

$$\int ... e^{\frac{i}{\hbar}[S(x)-S(x')]} e^{\frac{i}{\hbar}[S(X)-S(X')+S(x,X)-S(x',X')]} \mathcal{D}x(\tau)\mathcal{D}x'(\tau)\mathcal{D}X(\tau)\mathcal{D}X'(\tau), \tag{48}$$

where $S(x)$ and $S(X)$ are the actions for the small and the interacting systems, respectively, and $S(x,X)$ the action of interaction between the two systems. The integrals over phonon paths involve only the second exponential, and an "influence functional" can be defined (Feynman and Vernon, 1963),

$$\mathcal{F}(x(\tau),x'(\tau)) = \int_{X_i,t_i}^{X,t} \int_{X_i',t_i}^{X',t} \mathcal{D}X(\tau)\mathcal{D}X'(\tau) e^{\frac{i}{\hbar}[S(X)-S(X')+S(x,X)-S(x',X')]}, \tag{49}$$

such that the evolved density matrix is written as

$$\rho(q,q') = \int dq_i \int dq_i' \int \mathcal{D}x(\tau) \int \mathcal{D}x'(\tau) \mathcal{F}(x(\tau),x'(\tau)) e^{\frac{i}{\hbar}[S(x)-S(x')]} . \tag{50}$$

Here for any given path of the system of interest, $\mathcal{F}$ carries the information of the influence on that path of the integral of all paths of the interacting system. The theoretical step performed by introducing the influence functional is not trivial since it includes all the effects of the interacting system influencing the behaviour of the system of interest.

However, the explicit evaluation of the influence functional is in general prohibitively difficult, and approximations must be made as in the more standard formulations of the problem. For systems where the coupling action is a linear function of the coordinates of the interacting part and for systems weakly coupled, an analytical evaluation of the influence functional is possible (Feynman and Vernon, 1963); for general systems however this is not true.

The real-time path-integral technique was applied, without recurring to the influence functional, in the pioneer work of Fischetti and DiMaria (1985). The effect of the phonons was included by introducing in the action a self energy obtained by solving iteratively the Dyson equations for the electron-phonon interaction. Electron paths were then sampled with a Monte Carlo algorithm.

As a more recent application of the real-time path-integral method to electron transport in semiconductors, we will refer to a work (Mason and Hess, 1989) that attacks the problem of the transient electron response in a homogeneous semiconductor after the application of a constant and uniform electric field. In order to be able to handle the influence functional analytically, the authors take a linear model for the electron-phonon interaction. In the resulting expression another approximation has to be made for the spectral density of the phonons. Finally, when the electron paths are sampled with a Monte Carlo technique, the space of possible paths had to be reduced to a manageable size, and for this purpose the time of integration had to be reduced to the order of one semiclassical collision time as for the applications of the QMC technique for the density matrix described above.

The problem seems to be the same in all such quantum-transport numerical calculations: multiple integrals with strongly oscillating kernels need to be evaluated. Any naive sampling technique is bound to fail because it samples very large values that should cancel with each other leaving a final result many orders of magnitude smaller. A smart technique must be found to handle such strongly oscillating integrals if numerical results are sought for times at which steady-state conditions are reached. Some attention may be deserved, in this connection, by the windowing technique proposed by Register et. al (1988).

An example of results obtained in (Mason and Hess, 1989) is shown in Fig. 11. In part (a) of the figure the normalization of the density matrix is shown as a function of the collected statistics. It does not reach unity even at the highest statistics because of the limitation in the path space sampled with the Monte Carlo algorithm. If a larger path space is sampled, the results for average physical quantities are in principle more exact but are affected by a bigger statistical error. In part (b) of the figure the average space covered by the electrons during the time t is given again as a function of the amount of collected statistics. From its limiting value an average drift velocity (averaged over the past) can be obtained. However, owing to the approximations indicated above no comparison can be made with equivalent semiclassical results.

GREEN-FUNCTION APPROACH

We shall not discuss here in further details the Green-function techniques since several speakers will deal with this subject. For numerical applications, in particular, we refer to Reggiani's lectures on "Monte Carlo algorithms for nonequilibrium quantum transport".

It is worth mentioning, however, the work of the Houston group (Lei and Ting, 1985, Xing and Ting, 1987). The old idea of momentum and energy balance equations with a Maxwellian distribution is resumed in their work and transferred to a Green functions quantum approach. The time evolution of the density matrix is written to first order in the electron-impurity and electron-phonon interaction, starting from an initial condition that already contains, as parameters, an electron

drift and an electron temperature different from the lattice temperature. By application of Green function tecniques balance equations are obtained whose solution yield the electron drift velocity and temperature.

ACKNOWLEDGEMENTS

This work was supported by the Italian C.N.R. under the "Progetto Finalizzato MADESS"

REFERENCES

- Barker, J.R., 1980, in "Physics of Non Linear Transport in Semiconductors", Urbino (Italy), pag. 127, Plenum.
- Barker, J.R. and Lowe, D., 1981: "Quantum theory of hot electron-phonon transport in inhomogeneous semiconductors", *Journal de Physique Coll.* **C7**, Suppl. to n.10 (42), 293.
- Barker J.R. and Murray, S., 1983: "A quasi-classical formulation of the Wigner function approach to quantum ballistic transport", *Phys. Lett.* **93** A(6), 271.
- Barker J.R., 1985: "Quantum theory of hot electron tunnelling in microstructures", *Physica* **134** B, 22.
- Brunetti, R., Jacoboni, C., and Rossi, F., 1989: "Quantum theory of transient transport in semiconductors: a Monte Carlo approach", *Phys.Rev.* **B39**, 10781.
- Ferry D.K., Barker J.R., and Jacoboni C.,1990:"Quantum transport in semiconductors", to be published.
- Feynman, R.P. and Hibbs, A.R., 1965, "Quantum Mechanics and Path Integrals", McGraw-Hill, New York.
- Feynman, R.P., and Vernon, F.L., 1963: "The theory of a general quantum system interacting with a linear dissipative system", *Annals of Phys.* **24**, 118.
- Fischetti, M.V., and DiMaria, D.J., 1985: "Quantum Monte Carlo simulation of high-field electron transport: an application to silicon dioxide", *Phys. Rev. Lett.*55(22), 2475.
- Frensley W.R., 1987: "Wigner-function model of a resonant-tunneling device", *Phys. Rev. B*36(3), 1570.
- Frensley W.R., 1988: "Quantum transport calculation of the frequency response of resdonant-tunneling heterostructure devices", *Superlatt. and Microstruct.* 4(4/5), 497.
- Grubin H.L., Ferry, D.K., and Jacoboni, C., 1988: " The Physics of Submicron Semiconductor Devices", NATO-ASI Series, Series B: Physics vol. 180.

- Houston, W.V., 1940: "Acceleration of electrons in a crystal lattice", *Phys. Rev.* **57**, 184.
- Iafrate, G.J., 1988: "Quantum Transport and The Wigner Function", in "The Physics of Submicron Semiconductor Devices", edited by H.L. Grubin, D.K. Ferry, and C. Jacoboni, NATO-ASI Series, Series B: Physics vol. 180, p. 521.
- Jacoboni, C., Poli, P., and Rota, L., 1988: "A New Monte Carlo Technique for the Solution of the Boltzmann Transport Equation", Proc. Int. Conference on Hot Carriers in semiconductors, edited by J. Shah and G. Iafrate, *Sol. State Elect.* **31** (314), 523.
- Kadanoff L. P. and Baym G., 1962, " Quantum Statistical Mechanics", Benjamin/Cummings, Reading, Mass. .
- Kluksdahl, N.C., Poetz, W., Ravaioli, U., and Ferry, D.K., 1987: "Wigner function study of a double quantum barrier resonant tunneling diode", *Superlatt. Microstruct.* **3**, 41.
- Kluksdahl, N.C., Kriman, A.M., Ferry, D.K., and Ringhofer, C., 1989: "Self-consistent study of the resonant-tunneling diode", *Phys. Rev.* B **39**(11), 7720.
- Lee, H.M. and Scully, M.O., 1982: "The Wigner phase-space description of collision processes", *J. Chem. Phys.* **77**, 4604.
- Lei, X.L., and Ting, C.S., 1985: "Green's-function approach to nonlinear electronic transport for an electron-impurity-phonon system in a strong electric field",*Phys. Rev. B***32**(2), 1112.
- Lin, J. and Chiu, L.C., 1984: "Quantum theory of electron transport in the Wigner formalism", *J. Appl. Phys.* **57**(4), 1373.
- Mahan, G.D., 1981: "Many-particle Physics", Plenum, New York.
- Mason, B.A., and Hess, K., 1985: "Quantum Monte Carlo calculations of electron dymanics in dissipative solid-state systems using real-time path integrals", *Phys. Rev. B* **39**(8), 5051.
- Menziani, P., Rossi, F., and Jacoboni, C., 1989: "Impurity scattering in quantum transport simulation", *Sol. State Electr.* **32**, 1807.
- Poli, P., Rota, L., and Jacoboni, C., 1989: "Weighted Monte Carlo for electron transport in semiconductors", *Appl. Phys. Lett.* **55**(10), 1026.
- Ravaioli, U., Osman, A.M., Poetz, W., Kluksdahl, N.C., and Ferry, D.K., 1985: "Investigation of ballistic transport through resonant-tunneling quantum wells using Wigner function approach", *Phisica* **134B**, 36.
- Register, L.F., Littlejohn, M.A., and Stroscio, M.A., 1988: "Feynman path integral study of confined carriers subject to a statistical potential", *Sol. State Electr.* **31**,563.
- Register, L.F., Stroscio, M.A., and Littlejohn, M.A., 1988b: "Numerical Evaluation of the Feynman integral-over-paths in real and imaginary-time", *Superlatt. and Microstr.***4**(1), 61.
- Rickayzen, G., 1980: "Green's functions and condensed matter", Academic Press, New York.
- Rossi, F. and Jacoboni, C., 1989: "A quantum description of drift velocity overshoot at high electric fields in semiconductors", *Sol.State Electr.* **32**, 1411.
- Ter Haar, D., 1961: "Theory and applications of the density matrix", *Reports on progress in Physics*, **24**, 304.

- Toda, M., Kubo, R., and Saito, N., 1983: "Statistical Physics I", Springer-Verlag.
- Wigner, E., 1932: "On the quantum correction for thermodynamic equilibrium", *Phys. Rev.* **40**, 749.
- Xing, D.Y., and Ting, C.S., 1987: "Green's-function approach to transient hot-electron transport in semiconductors under a uniform electric field", *Phys. Rev. B* **35**(8), 3971.

FIGURE CAPTIONS

Fig.1 : diagram of a seventh-order term of the perturbative expansion for the electronic density matrix including electron-phonon and electron-impurity interactions (see text).

Fig.2 : quantum distribution function as a function of k along the field direction containing terms up to the fourth order (continuous line), compared with the corresponding classical one (dashed line). The third curve (dotted line) represents the ballistic translation of the initial distribution function (Brunetti et. al., 1989).

Fig.3 : quantum electronic distribution function as a function of k along the field direction containing terms up to the fourth order (continuous line), after 50 fs from the initial conditions. The dashed curve is the classical distribution function at the same perturbative order obtained from the Boltzmann equation. The dot-dashed curve is the initial distribution after the ballistic shift produced by the electric field (Menziani et. al., 1989).

Fig.4 : classical electron distribution as a function of energy for a a simplified GaAs model at t=100 fs after excitation. The highest peak at 1000 K is the initial distribution at t=0 (Brunetti et. al., 1989).

Fig.5 : quantum electron distribution as a function of energy for the same simplified GaAs model as in Fig.4 at t=100 fs after excitation (Brunetti et. al., 1989).

Fig.6 : comparison among the CMC drift velocity (continuous curve), its perturbative expansion up to the fourth order (dot- dashed line), and the quantum fourth-order result (Rossi and Jacoboni, 1989).

Fig.7 : momentum (a), and position (b) distributions at a short time after the collision of a gaussian wave packet with a resonat-tunneling barrier. (c) shows the Wigner distribution of the system at the same observation time (Barker, 1985).

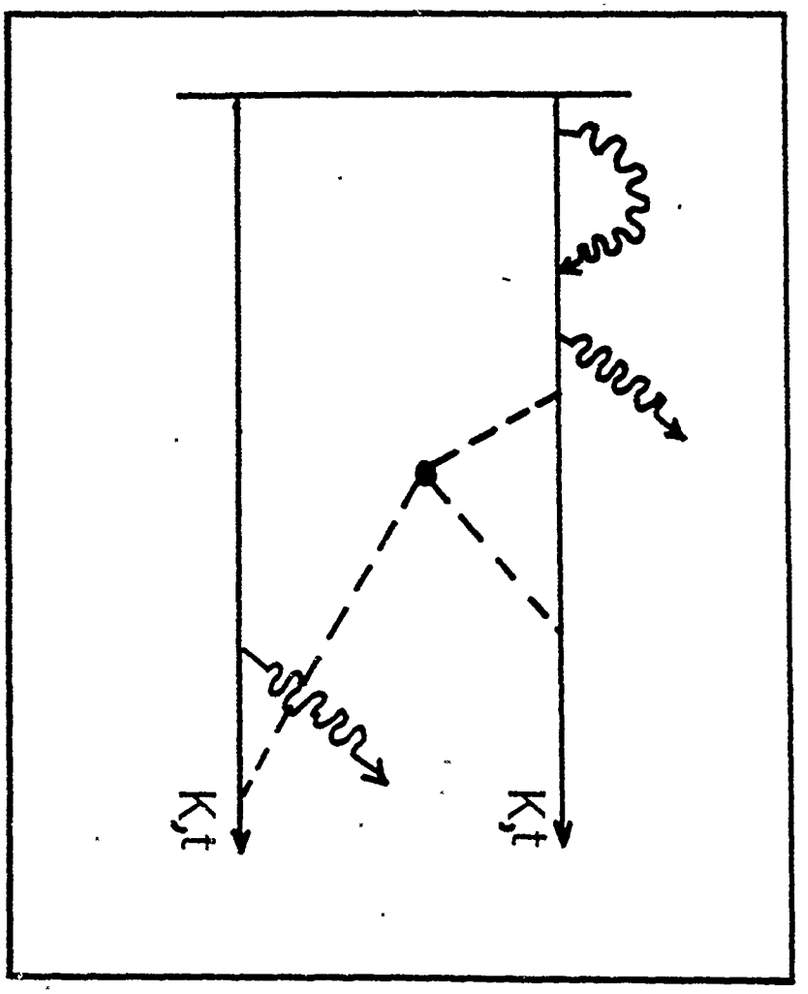
Fig.8 : gaussian wave packet interacting with resonant quantum potential barriers. The barriers are indicated by the dark band. (a) The incident wave packet, moving from left to right, is just beginning to interact with the barriers. (b) Gaussian wave packet during reflection. The incident and reflected components are visible, as is the correlation centered around k=0. Part of the packet is tunneling through the barrier. (c) Gaussian wave packet after reflection. Most of the wave packet has been reflected. The tunneling packet is visible to the right of the barriers (Kluksdahl et. al., 1989).

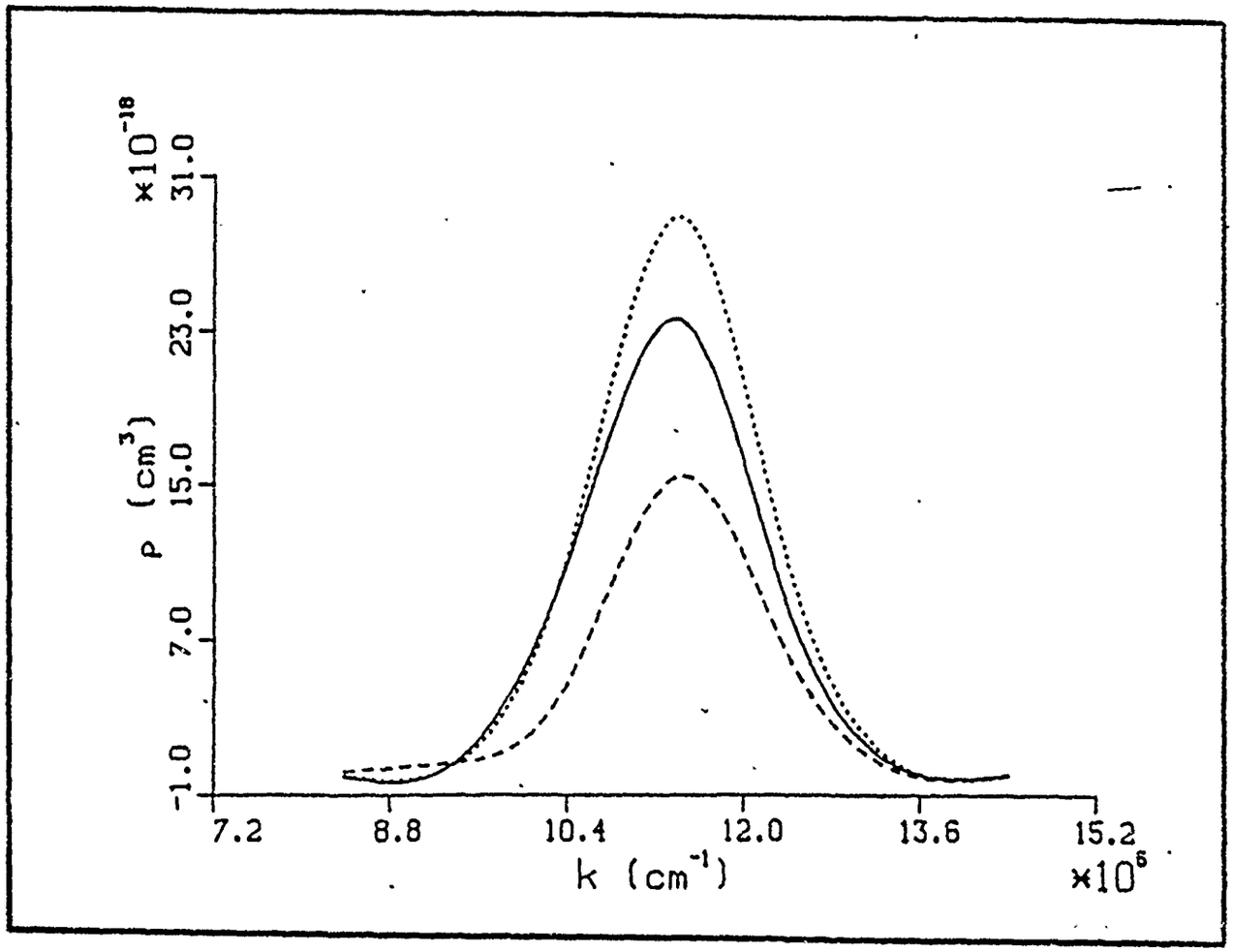
Fig.9 : (a) steady-state Wigner distribution at the peak of the I-V curve. Depletion in the cathode region is evident. At the left contact the incoming distribution appears as a shifted Fermi-Dirac distribution. (b) steady-state Wigner distribution at the valley of the I-V curve. Depletion is strongly evident in the cathode region. The distribution in the cathode -barrier-interface region forms a quantised state, the ring structure to the left of the barriers (Kluksdahl et. al., 1989).

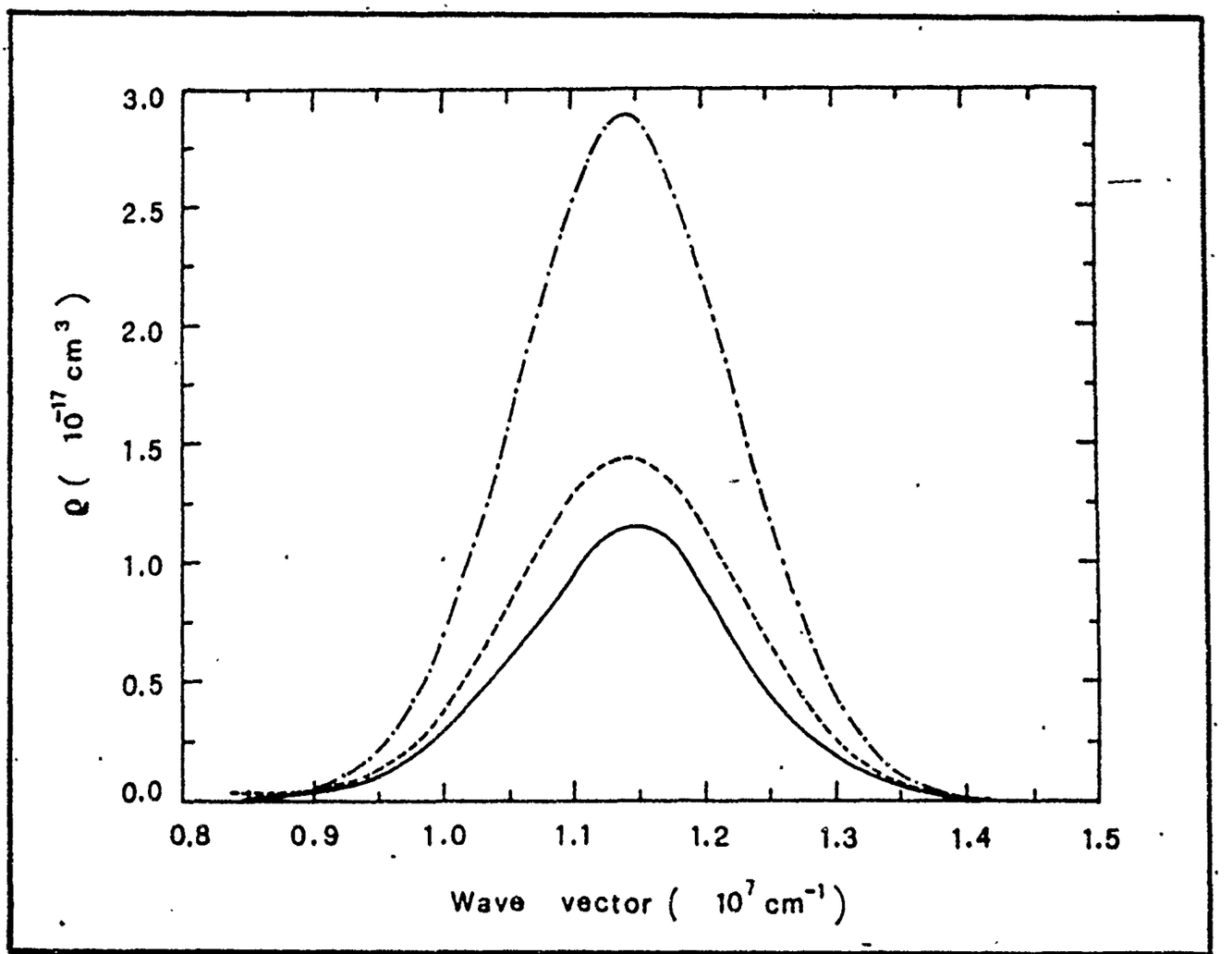
Fig.10 : difference between the bistable Wigner distributions at bias of 0.36 V. The quantised state in the cathode well has more carriers. More current is flowing, indicated by the "ridge" in the distribution (Kluksdahl et. al., 1989).

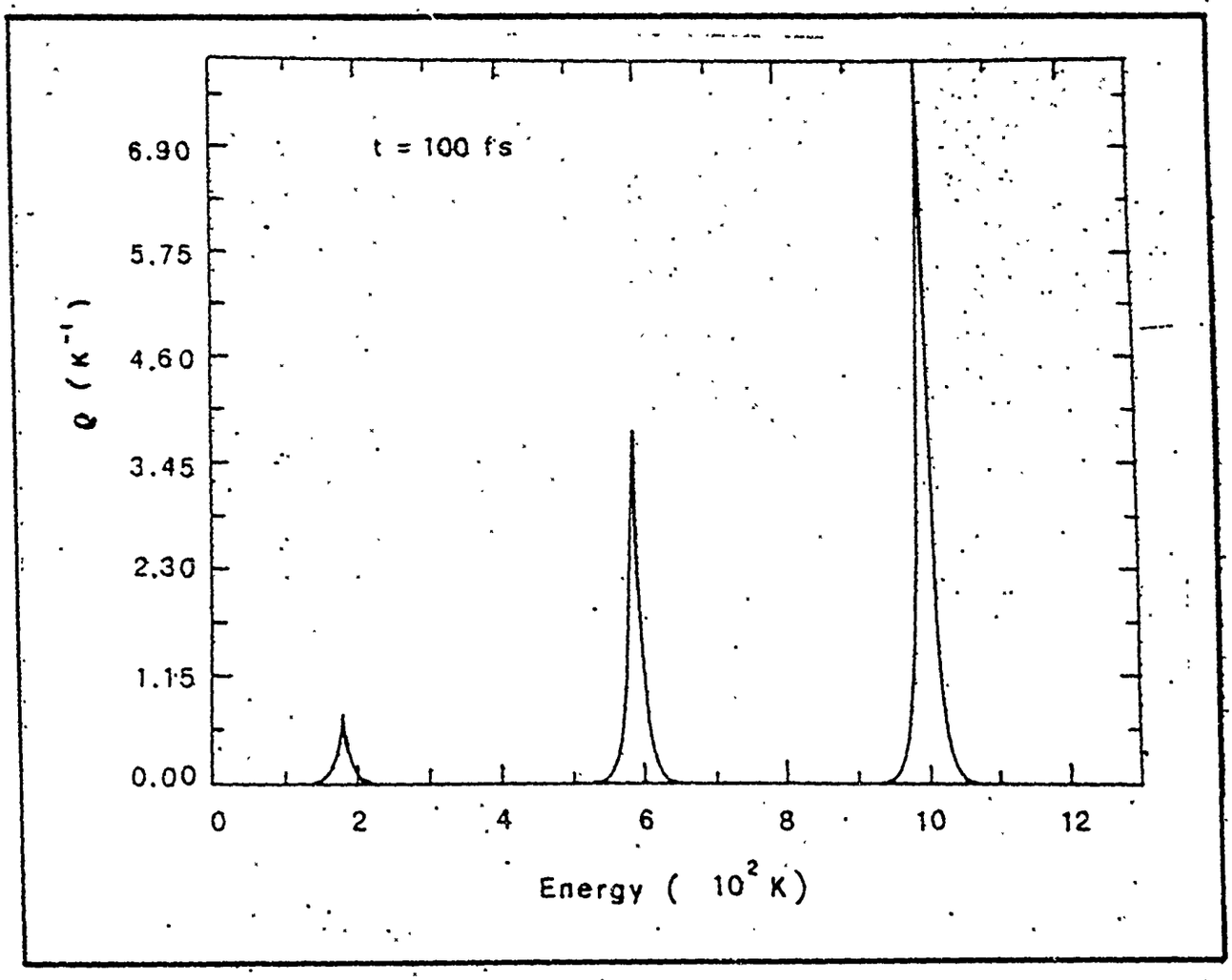
Fig.11: expectation values as functions of the number of paths sampled. Sampling

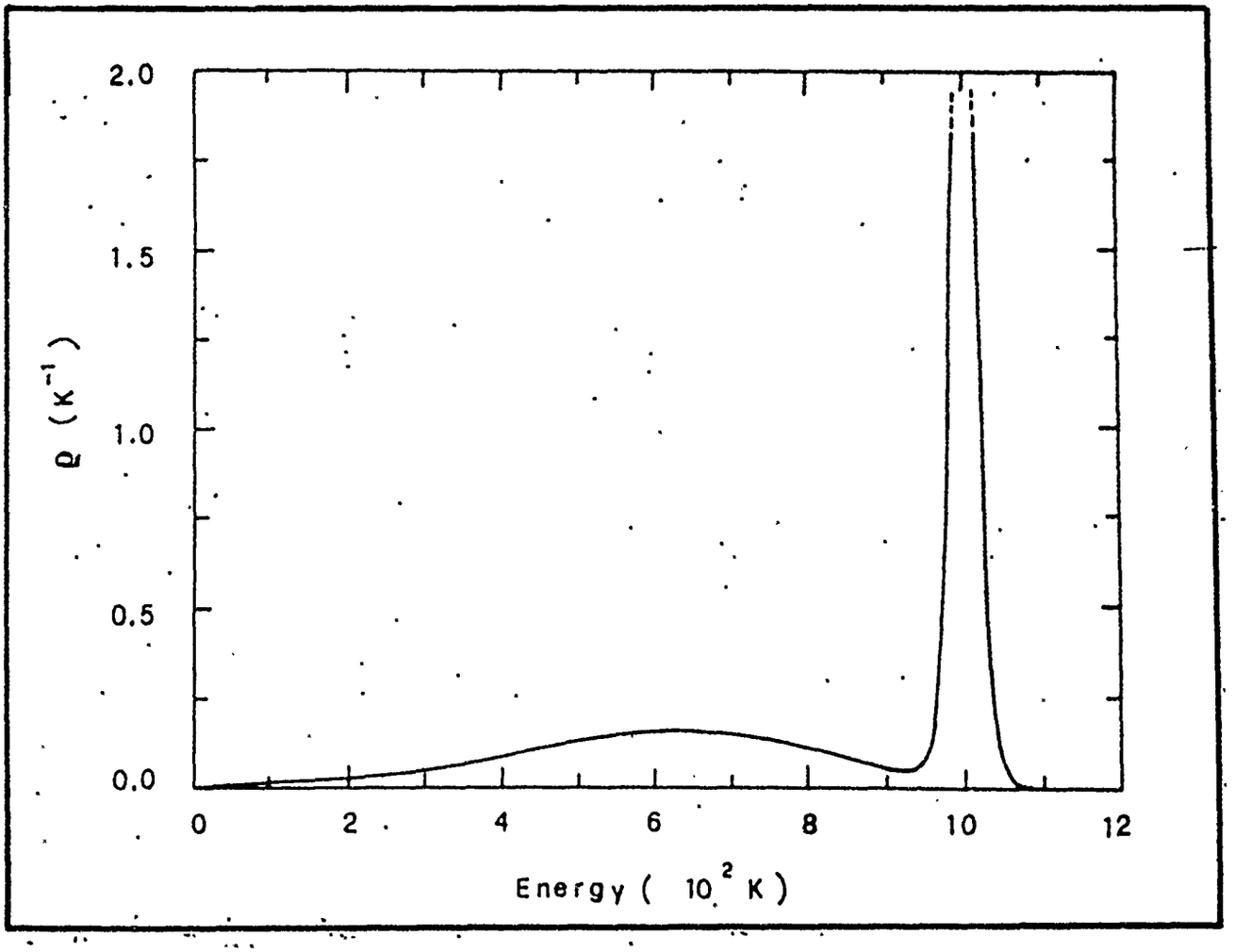
of each array includes 250.000 paths. (a) shows $Tr(\rho)$, (b) represents $\frac{\langle x \rangle}{\lambda}$. γ is an equivalent scattering rate that gives the high-temperature friction, and λ is a reference-length: $\lambda = \sqrt{\hbar/m\gamma}$. All are for $t = \frac{1}{\gamma}$ and $K_B T = 1.2\hbar\gamma$ (Mason and Hess, 1985).

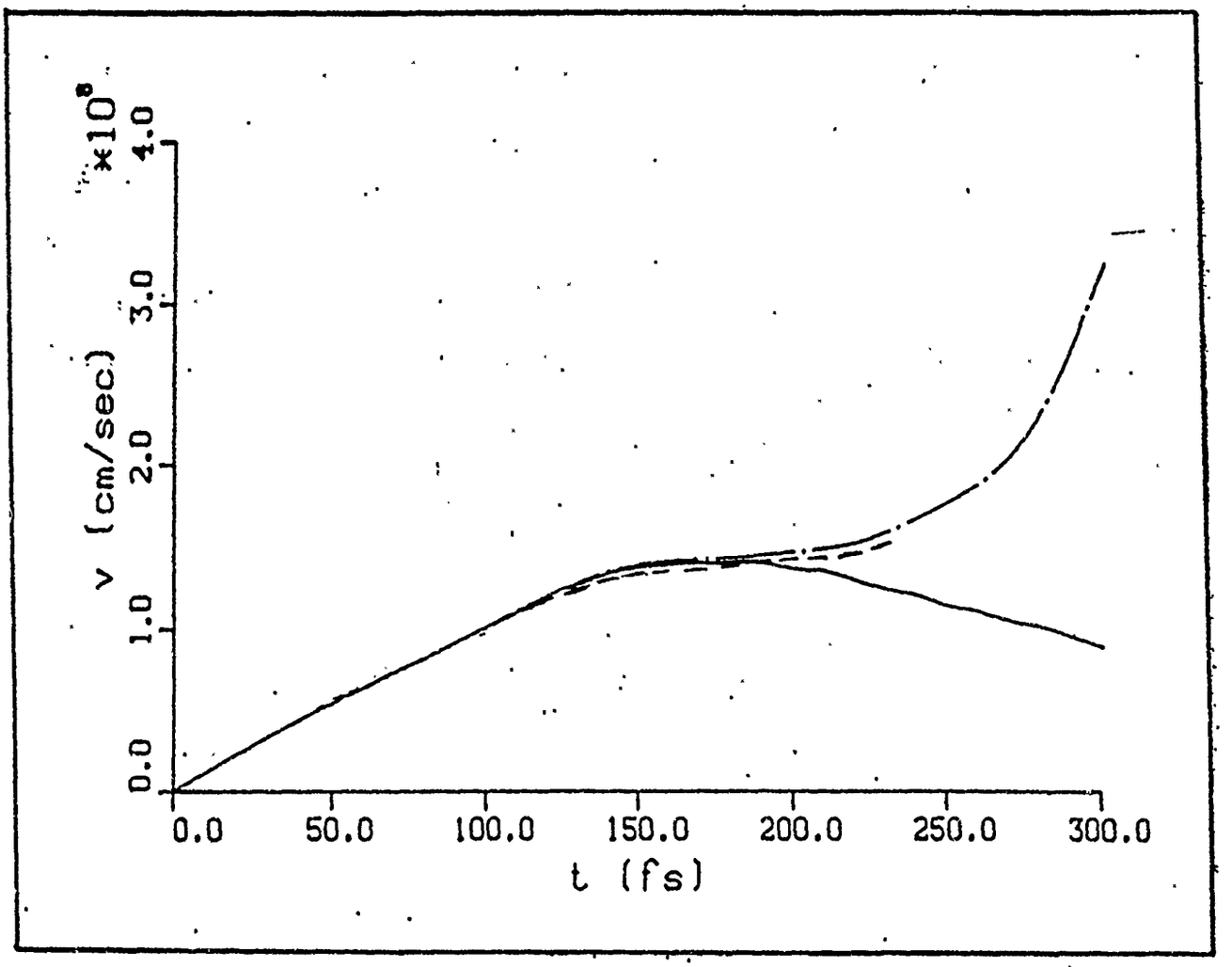

Kt
Kt

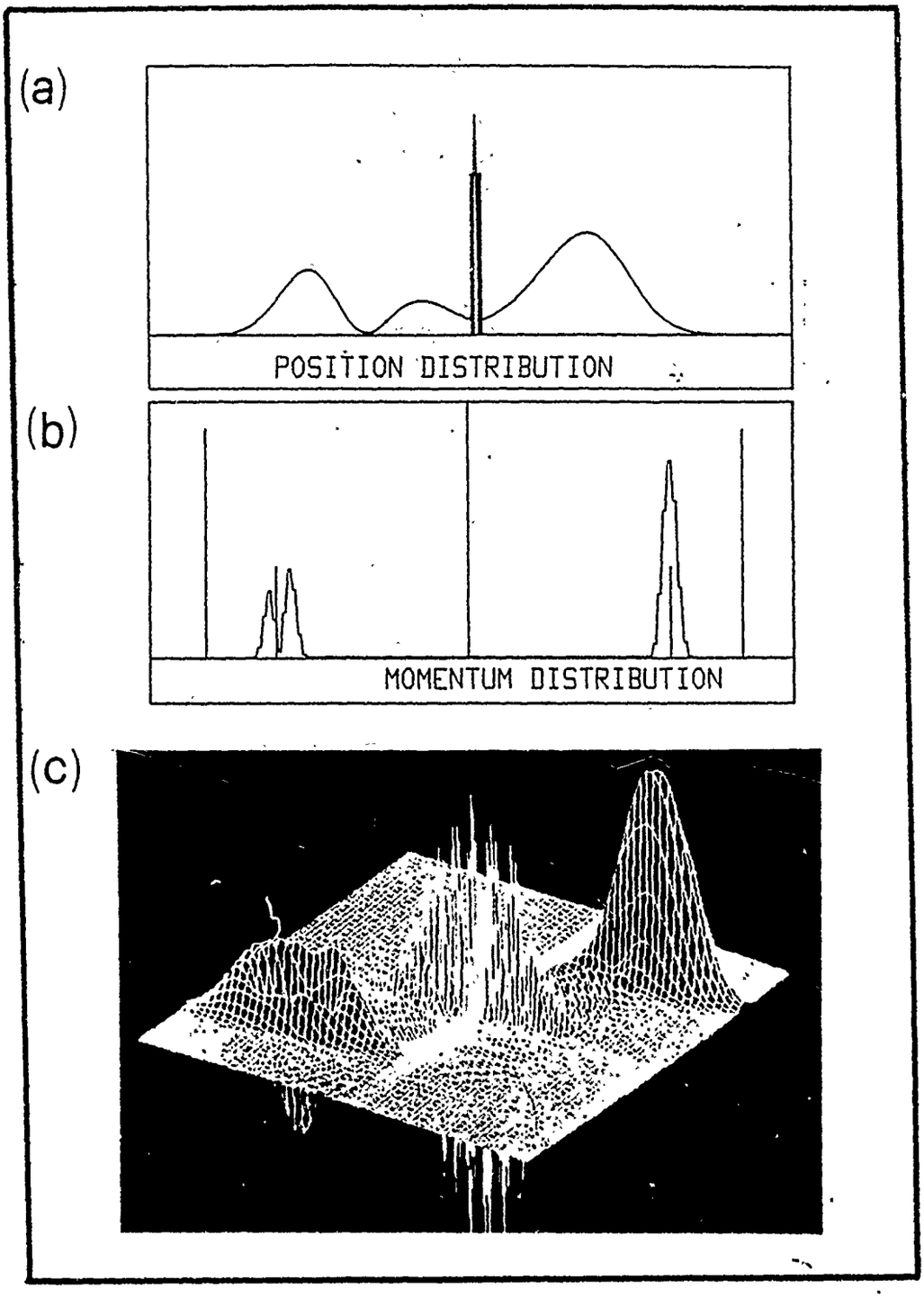

×10⁻¹⁸
31.0
23.0
15.0
7.0
-1.0
P (cm³)
7.2
8.8
10.4
12.0
13.6
15.2
k (cm⁻¹)
×10⁶

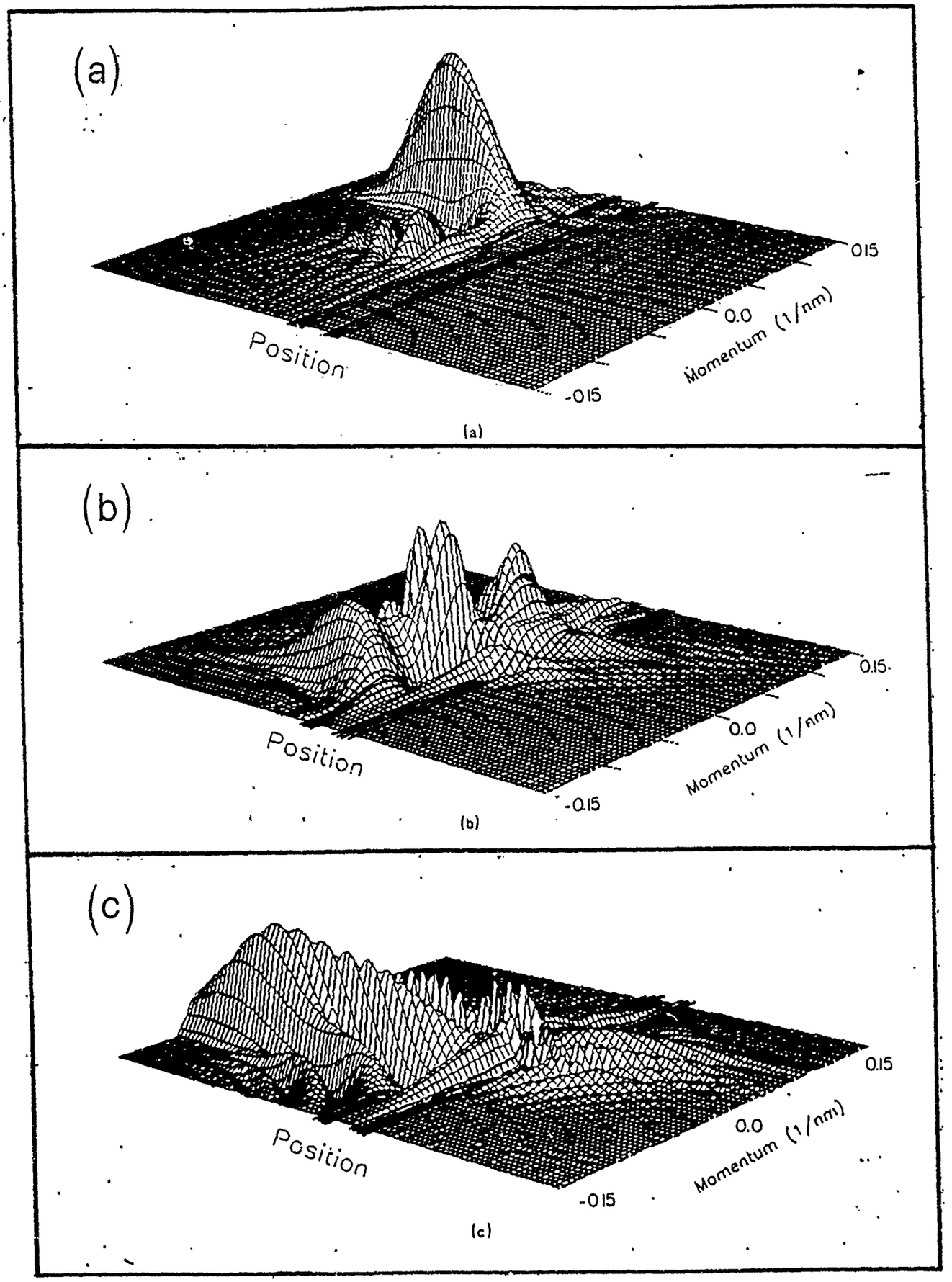

3.0
2.5
2.0
1.5
1.0
0.5
0.0
Q (10^{-17} cm^3)
0.8
0.9
1.0
1.1
1.2
1.3
1.4
1.5
Wave vector (10^7 cm^{-1})

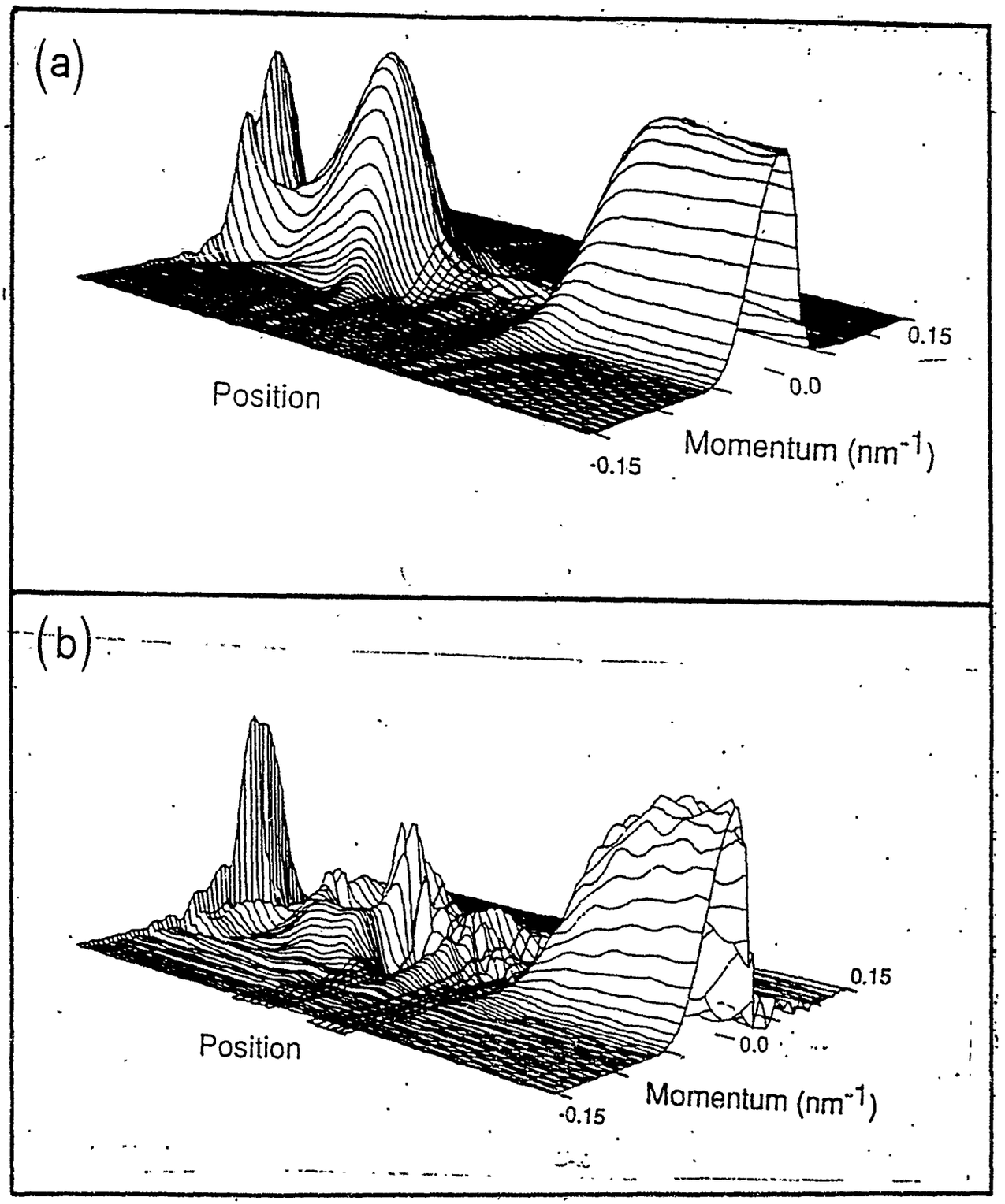

t = 100 fs
6.90
5.75
4.60
3.45
2.30
1.15
0.00
Q (K⁻¹)
0
2
4
6
8
10
12
Energy (10² K)

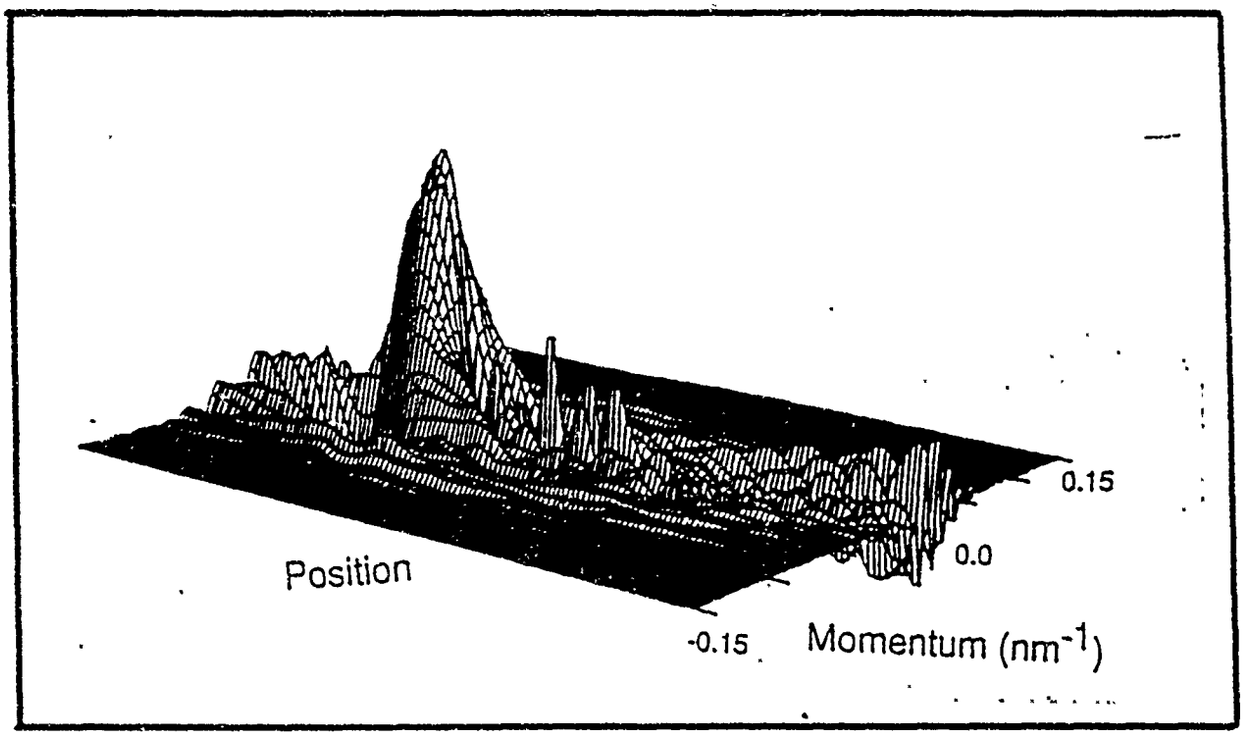

2.0
1.5
1.0
0.5
0.0
ρ (K⁻¹)
0
2
4
6
8
10
12
Energy (10² K)

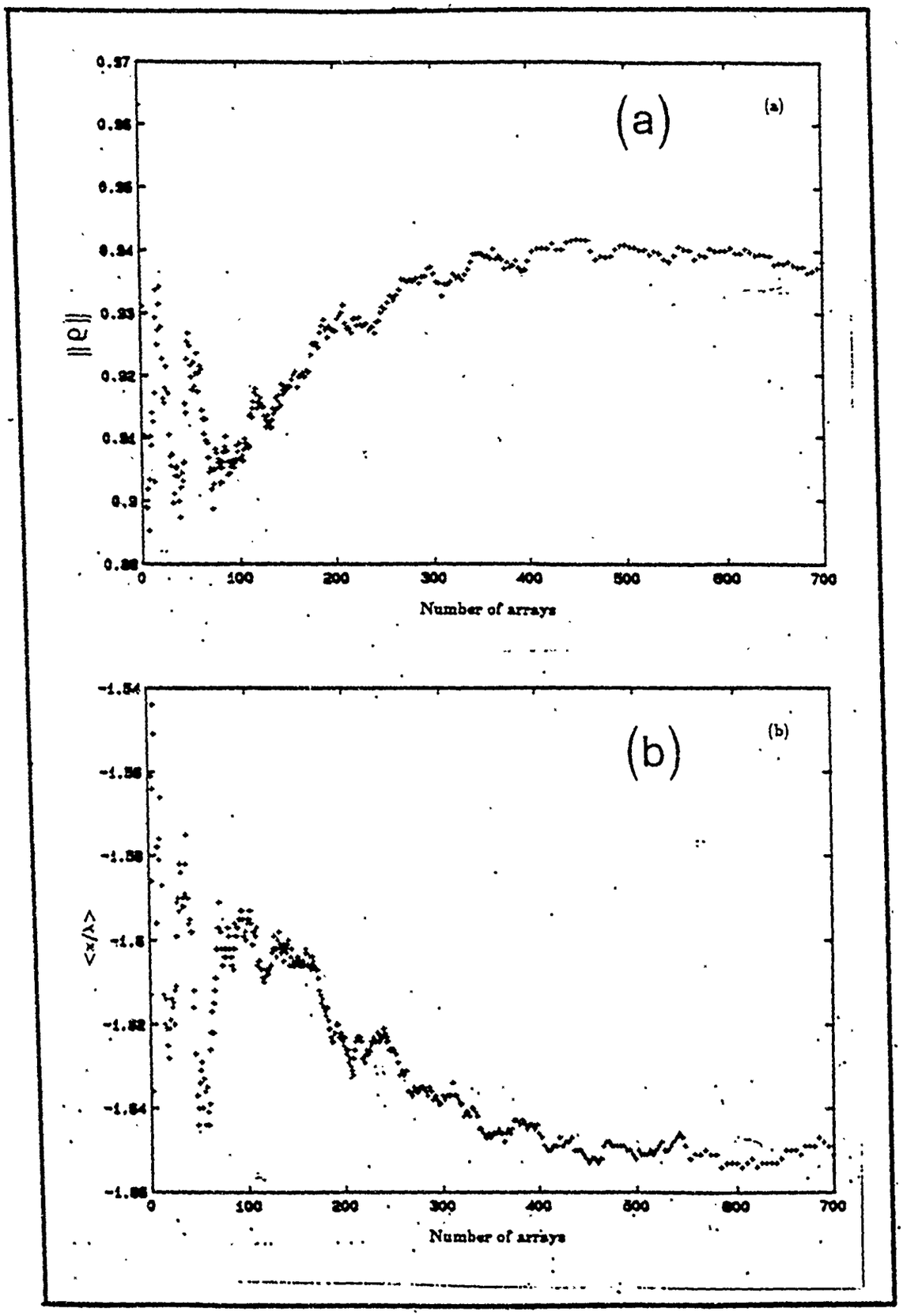

v (cm/sec)
*10^8
4.0
3.0
2.0
1.0
0.0
0.0
50.0
100.0
150.0
200.0
250.0
300.0
t (fs)


(a)
POSITION DISTRIBUTION
(b)
MOMENTUM DISTRIBUTION
(c)


(a)
Position
-0.15
0.0
0.15
Momentum (1/nm)
(a)
(b)
Position
-0.15
0.0
0.15
Momentum (1/nm)
(b)
(c)
Position
-0.15
0.0
0.15
Momentum (1/nm)
(c)


(a)
Position
-0.15
0.0
0.15
Momentum (nm^{-1})
(b)
Position
-0.15
0.0
0.15
Momentum (nm^{-1})


Position
0.15
0.0
-0.15
Momentum (nm^{-1})


(a)
(a)
0.97
0.96
0.95
0.94
0.93
0.92
0.91
0.9
0.89
||a||
0
100
200
300
400
500
600
700
Number of arrays
(b)
(b)
-1.54
-1.56
-1.58
-1.6
-1.62
-1.64
-1.66
<x/λ>
0
100
200
300
400
500
600
700
Number of arrays

LATERAL SURFACE SUPERLATTICE AND THE FUTURE OF ULSI MICROELECTRONICS

David K. Ferry

Department of Electrical Engineering
Arizona State University
Tempe, Arizona 85287-5706

INTRODUCTION

Since the introduction of integrated circuits in the late 1950's, the number of individual transistors that can be placed upon a single circuit has approximately doubled every three years. Today, even university design laboratories for the teaching of students can access chip foundries which produce 1.2 μm (and smaller) design rule circuits. Compared with this, many commercial companies are experimenting with the production of chips with critical dimensions of 0.1 μm, and university laboratories have produced individual devices with gate lengths much smaller than this (Patrick *et al.*, 1985; Bernstein and Ferry, 1986; Jin *et al.*, 1987; Sai-Halasz *et al.*, 1988; de la Haussaye *et al.*, 1988; Ishibashi *et al.*, 1988). The creation of devices whose spatially important scales may be only a few tens of nanometers opens the door to the study of many new and important physical effects, some of which have been described earlier (Barker and Ferry, 1981). Indeed, it can rightfully be said that it will be impossible to understand fully the operation of these devices without a full understanding of these newly appearing physical effects.

It is easy to understand the driving forces (and the need for further understanding of the physics). In the early 1980's, Hewlett-Packard produced a single-chip microprocessor containing approximately 0.5M devices in its 1 cm^2 area (Mikkelson *et al.*, 1981). This chip was fabricated with essentially 1.25 μm gate length transistors. Today, megabit memories and dense signal processing chips with devices of these same dimensions are being discussed. Yet, *we are also talking about reaching chip densities of 10^9 devices within a short period of time.* While the first question is what would one use so many devices to accomplish, even if we could reliably fabricate the chips with a meaningful yield, it is also to ask just what this does to the required device technology. In general, progress in the integrated circuit field has followed a complicated scaling relationship (Baccarani *et al.*, 1984). This scaling reduces feature sizes by an amount S. To reach a billion transistors, as envisaged, requires a scale-up of a factor of 2000 over the HP chip, which means S=45. Thus, if the scaling relationships are followed, one expects to see transistors with gate lengths of only 30 nm! Very few laboratories have produced research devices with gate lengths on this scale and little is understood about the limitations (from the physics) that will determine whether or not these devices are practical. The growth of the number of transistors per chip is illustrated in Fig. 1.

What is happening in the reduction of individual feature sizes of a transistor, used as the basic building block for ULSI, is that the critical length (e.g. the gate length or a depletion length) will become so small that it approaches the *coherence length* of the electrons that provide the operation. Over the past several years, it has become evident that this latter length is not the wavelength of the electron itself, but the *inelastic mean free path*, or the length over which the energy coherence is maintained by the electron. With modern modulation doping techniques in heterojunction device structures, this latter length can be more than 1 μm at low temperatures, but there is also evidence that it can be as much as 0.1 μm at room temperature. The consequence is that such small devices must now be treated as quantum mechanical objects, and many phenomena become important that have never been treated in the normal classical and semi-classical treatments of semiconductor devices. While this has served to invigorate studies of quantum behavior in device structures, we are limited in that many of these quantum phenomena are only poorly understood at best.

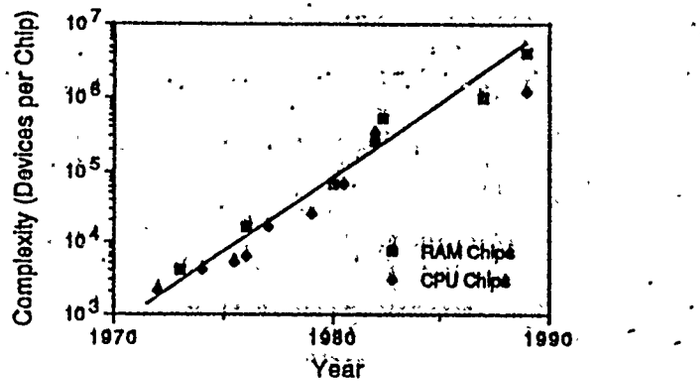


Fig. 1 The increase in the number of devices contained on a single integrated circuit chip has not slowed for more than two decades.

In this paper, a number of quantum effects that are important in devices are reviewed. No comprehensive treatment of each, nor a comprehensive treatment of all effects, is intended. Rather, the selection is governed by those effects which have been shown to already occur in devices. Here we try to establish the connection between a few such effects. Nearly all semiconductor devices operate on the principle of hindering the transport of carriers from the source (or emitter) to the drain (or collector) by the presence of a potential barrier, which is modulated by the gate (or base potenial). As the size of devices has been reduced, so-called second-order effects have introduced unintended modifications to this barrier through parasitic effects such as drain-induced barrier lowering (Troutman, 1979). In the ultra-submicron regime, we must begin to consider that many carriers will actually tunnel through the barrier, and encounter strong potential variations, further changing the basic operation of the device. In addition, the active channel length can be much less than 0.1 μm after the barriers have been surpassed. Carriers then have the possibility of transiting this region *ballistically*; e.g. without scattering, or perhaps suffering just a few elastic collisions. Then one can expect to see quantum effects and quantum resonances in this ballistic transport, such as that seen in transport through thin oxides in MOS devices (Lewicky and Maserjian, 1975; Fischetti *et al.*, 1987).

In the following section, the fabrication of small semiconductor FETs will be discussed and the effects which limit the normal scaling rules will be described. These include small aspect ratio, source impedances, velocity overshoot, and tunneling through the depletion barrier. However, it will also be shown that limitations to down-scaling can arise when tunneling becomes important as well. Then, attention will turn to a treatment of lateral surface

superlattices, in which a large two-dimensional array of "quantum boxes" will be defined by electron-beam lithography. The study of cooperative transport in this array will be described.

ULTRASMALL FETS

For devices with large gate lengths, the gradual channel approximation is valid. In the gradual channel approximation, the mobility is assumed constant. The density in the channel drops in going from the source to the drain, which leads to a rise in the electric field. As a consequence of the conservation of current, the velocity must rise along the channel. Thus the general behavior that results (the Shockley model) is the drain current is proportional to $(V_{GS} - V_{DS})^2$. In this region, the transconductance increases as the gate length decreases. This is indicated as region I in Fig. 2. The next region occurs when the velocity is no longer allowed to increase forever and saturates at some value, typically 10^7 cm/sec. The current density through the device is then limited to a value given by the product of the saturation velocity and the carrier density at the position in the device where the velocity saturation sets in. The current through the device can be written as

$$J = nev = C_0 v_s (V_{GS} - V_{sat}) , \qquad (1)$$

where v_S is the saturation velocity and V_{sat} is the voltage in the channel at the position where the velocity rises to v_S. Equation (1) is written for a MOSFET, but the equivalent behavior is found in a MESFET (Bernstein and Ferry, 1988). Equation (1) also shows that the transconductance is no longer a function of the gate length, and to first order, this region of the transconductance versus gate length plot remains flat (assuming that the thickness of the epitaxial layer is not reduced as the gate length is reduced). At the point where the aspect ratio (L_g/a or gate length to active layer thickness) drops below 5, the transconductance begins to decrease as the gate length is reduced. This is due to the depletion layer under the gate becoming dominated more by the fringing region rather than the flat "parallel plate" region directly under the gate metalization. The transconductance can be written as

$$g_m = \frac{\partial I_{DS}}{\partial V_{GS}} = W v_s C_{GS} . \qquad (2)$$

For the HEMTs, a purely parallel plate model for the capacitance yields a transconductance of 575 mS/mm and a reduction by the normalizing factor to nearly 200mS/mm. This illustrates the well-recognized need to recess gates to increase the aspect ratio for submicron devices to maintain high transconductance. The aspect ratio reduction (region III in Fig. 2) is shown in Fig. 3.

The reduction of the aspect ratio arises from the fact that the capacitance of the gate is no longer that corresponding to a pair of parallel plates. Rather, the capacitance is now dominated by the fringing capacitance of the gate, so that the latter looks more like a wire over a plane. Thus, the induced charge, corresponding to a given voltage is spread out over a much larger area of the inversion layer, and the charge density at any one point is smaller than expected from the long-channel theory (Hauser, 1967). Once this aspect ratio reduction begins, it is only possible to further gain increases in the transconductance by increasing the saturated velocity, as may be seen from (2). This occurs when the entire gate depletion length is smaller than the inelastic, or energy relaxation, mean free path. Then, the gate transit region is dominated by the non-stationary dynamics of the carriers (Ferry, 1986). This is observed in FETs when the gate length is below about 0.1–0.2 micron. In some cases, further reduction at still smaller gate lengths occurs due to heating in the source resistance (Ryan *et al.*, 1989a, 1989b). This latter is indicated as region IVb in Fig. 2. If such heating occurs, then a fraction of the non-stationary transport actually occurs in the source-resistance area, and a corresponding shorter fraction of it occurs under the gate. This reduces the effective velocity over the gate length.

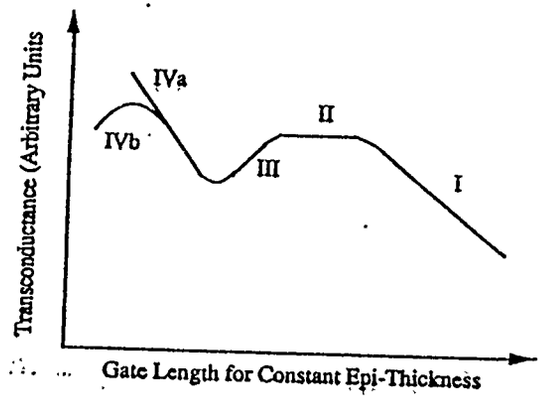


Fig. 2 The various operating regions for the transconductance as the gate length is reduced, for the assumption of a constant thickness of the active region.

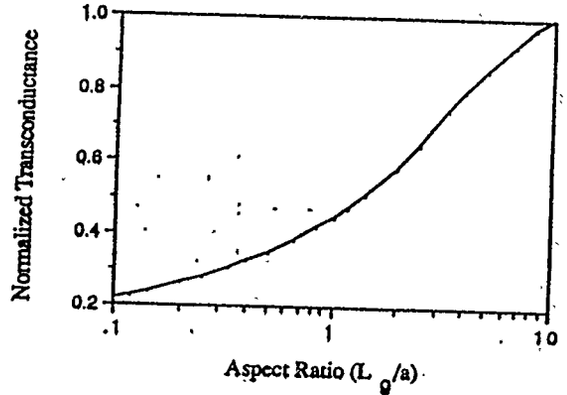


Fig. 3 The variation of the normalized transconductance with the aspect ratio, which provides the behavior of region III in Fig. 2.

HEMT devices were fabricated on molecular beam epitaxially (MBE) grown wafers, having layers of 5 nm thick, Si-doped 2×10^{18} cm^{-3} n^+-GaAs cap layer/ 35 nm thick, Si-doped 4×10^{18} cm^{-3} n^+-AlGaAs layer/ 300 nm undoped GaAs layer/ semi-insulating GaAs layer heterostructures (Han *et al.*, 1990). The wafer was designed to have a high carrier concentration in the active layer, which maximizes the carrier density in the channel. A typical structure is shown in Fig. 4, which is an electron micrograph of the mesa and gate/channel regions. It is evident from the figure that the gate lengths can be quite short. Transconductances (intrinsic) measured for a range of gate lengths are shown in Fig. 5. In this latter figure, it is quite evident that the behavior expected for regions III and IV of Fig. 2 are, in fact, observed in these devices. This is not always the case, depending upon the aspect ratio, of the fabricated devices, at which velocity overshoot begins to be observed. Nevertheless, the transconductances observed in Fig. 5 are of the order of magnitude expected for the epitaxial structure and aspect ratios in the actual devices. Effective velocities as high as 3×10^7 cm/sec

are observed for gate lengths in the 35-45 nm range (Han *et al.*, 1990). This is an increase of a factor of 3 in the effective velocity for these devices over the long-channel values.

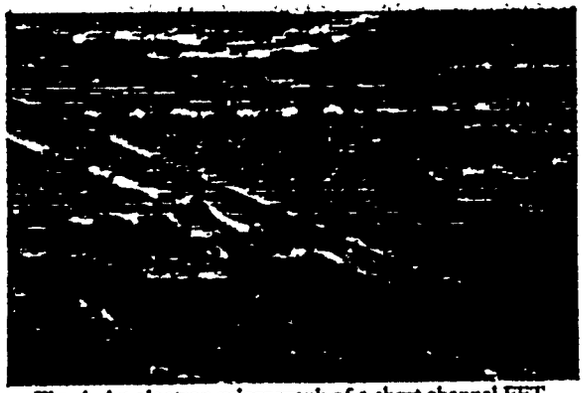


Fig. 4 An electron micrograph of a short channel FET.

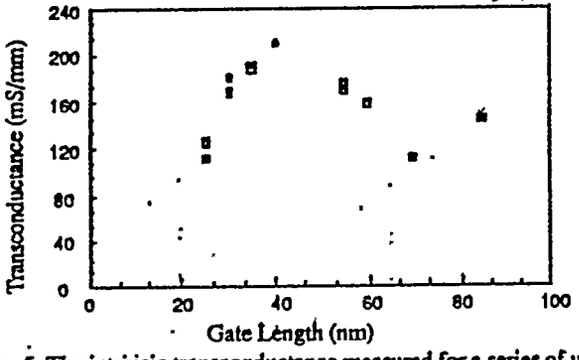


Fig. 5 The intrinisic transconductance measured for a series of ultra-short gate length HEMTs.

The two smallest devices (25 nm gate length) in Fig. 5 exhibit a totally different behavior, as it is quite difficult to pinch-off the channel in these devices. This can be understood by considering that up to this point in the discussion, all the reductions in transconductance as a result of gate length scaling can be overcome. The problem of a reduced aspect ratio is solved by recessing the gate. This just requires a slow, well controlled, uniform etch. Source ohmic contacts can be made with increasingly lower contact resistance to reduce overshoot saturation. Improved material growth techniques can solve substrate current and real-space transfer problems. The final scaling limit to FETs will be when the gate barrier seen by the carriers becomes narrow enough that tunneling through the gate is possible. The behavior of the 25 nm and 35 nm devices from Fig. 5 are shown in Fig. 6. It is clear that the 35 nm device still exhibits considerable rounding in the plot of the log of the drain current as a function of the total channel potential. This is expected for the behavior of (2). On the other hand, the 25 nm device exhibits a long region in which the log of the drain current varies linearly with the total channel potential. It is believed that this effect is in fact the onset of tunneling through the gate depletion region. This can be understood in the following manner.

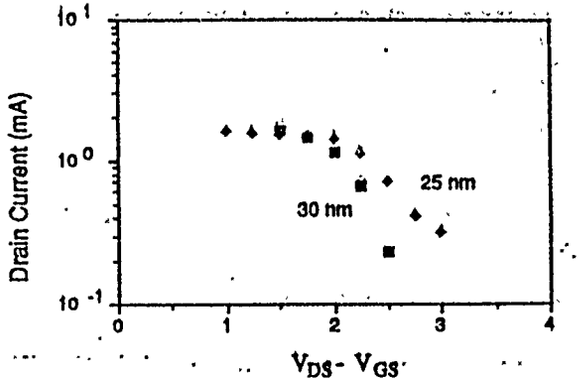


Fig. 6 Drain charactoristics for the 25 and 30nm gate length HEMTs which show the exponential dependence of the current, charactoristic of tunneling, for devices with gate lengths under 30nm (Ryan *et al.*, 1990).

For tunneling current, it is expected that we may use a simple WKB approximation in which the current decays exponentially with the integrated decay wave vector. Thus,

$$I_D \sim \exp\left[-\int_0^{L_{eff}} \gamma(x)dx\right], \tag{3}$$

where γ is the tunneling decay constant. Here, the kernel will be of the form

$$\gamma(x) = \sqrt{\frac{2m}{\hbar^2}V(x)}. \tag{4}$$

Without explicit knowledge of the shape of the tunneling potential V(x), we are free to assume it has the form

$$V(x) \cong V_{max}\left[1-\left(\frac{x}{L_{eff}}\right)^2\right], \tag{5}$$

Here, V_{max} is the barrier height, and L_{eff} is the width of the barrier at the Fermi level. This requires the assumption that the barrier height is a linear relationship to the channel voltage,

$$V_{max} \propto V_{DS} + |V_{GS}|. \tag{6}$$

At first, it might appear that the dependence should be on the square root of the voltage, but the effective length L_{eff} actually increases with V_{max}, and actually also varies with the square root of the potential. This leads to the log of the current varying linearly with the potential. Then, this implies that if $\ln(I_D)$ is plotted as a function of $(V_{DS} + |V_{GS}|)$ for the devices in Fig. 5, it will be possible to determine the effective barrier height at each current. For the HEMTs with a gate length less than 30nm, a linear relationship is observed, as shown in Fig. 6.

This behavior is not observed for longer gate lengths, and the onset of this behavior in the shortest channel length devices indicates the current is no longer a function of V^2 but has an exponential dependence. From this behavior, a barrier height of approximately 12mV is found

for the "cutoff" of the current. Ishibashi *et al.* (1988) have fabricated a 20nm gate length AlGaAs/GaAs HEMT. By again plotting the ln(I) as a function of $V_{DS}+|V_{GS}|$ for their 20nm and 50nm gate length devices (Fig. 7), a linear relationship is found for the device with a gate length less than 30nm, but not for the larger device. For MESFETs with gate lengths less than 30nm, no evidence of an exponential relationship is found. This is probably due to less confinement of the carriers to the channel in the MESFETs as compared to the HEMTs, as the current is pushed into the semi-insulating substrate, rather than being confined in the quantum well at the interface.

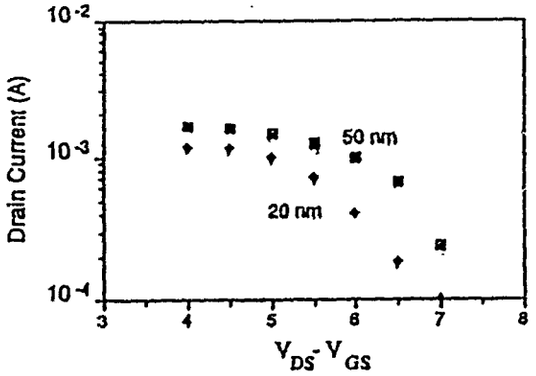


Fig. 7 Drain charactoristics for 20 and 50nm gate length devices from Ishibashi *et al.* (1988), which also exhibit exponential dependence of the current.

The fact that tunneling is being observed in the shortest gate length HEMTs (gate length less than 30 nm) indicates that the transfer behavior of the carriers is in fact starting to approach similar behavior to molecular structures. In molecular chains, it is thought that electrons move along the chain by a hopping, or tunneling, interaction. The gating mechanism for such charge transfer is not known at present, but the current observation once more brings home the point that traditional VLSI devices are becoming so small that the distinctions between molecular-level electronics and ULSI are rapidly becoming blurred. A further connection between these two regimes has been illustrated by Geerligs *et al.* (1990), in which charge transfer between two series tunneling junctions was controlled by capacitively gating the potential at the node between the series junctions. In the series junctions, the device capacitance was sufficiently small that each exhibited the phenomenon of single electron tunneling (SET), where the induced voltage arising from the transfer of a single charge across the small capacitance is large enough to self-bias the junction out of the tunneling regime (Averin and Likharev, 1990, and references therein). This latter effect is known as the Coulomb blockade (Averin and Likharev, 1988; Büttiker, 1987, and references therein). Both SET and the Coulomb blockade are discussed in other chapters in this proceedings by Büttiker and Likharev. In essence though, the transfer of charge, and the gating of this transfer, through tunneling junctions reinforces the connections between our current devices and those on the much smaller scale.

LATERAL SURFACE SUPERLATTICES

The two-dimensional motion of electrons that are subjected to both a two-dimensional periodic potential and a perpendicular magnetic field is a problem that has been studied for a great many years. In general, the solution of this problem is complicated by the fact that there are two characteristic lengths in the problem – one is the period a of the superlattice potential, while the second is the magnetic length $L_m=(\hbar/eB)^{1/2}$. The complete problem can be solved generally in an infinite domain only when these two lengths are related by the ratio of two integers. However, there are two distinct limits in which perturbation theory can be used to obtain solutions. In one limit, the Landau regime, the periodic potential is regarded as a weak perturbation on the usual magnetic Landau level structure. When the flux coupled through each unit cell of the periodic potential is given by $\phi=ea^2B/h=p/q$, each Landau level splits into p subbands of equal degeneracy. It has been known for some time that a series of oscillations, periodic in 1/B, should arise in magneto-transport measurements (Pippard, 1964; Azbel, 1964; Rauh *et al.*, 1974). These oscillations have an appearance similar to, but an origin different than, the Shubnikov-de Haas oscillations. It is easy to think of these oscillation in terms of commensurability of the cyclotron radius with the lattice periodicity, but the physics of the oscillations is related to the Fermi level moving through the split bands of the Landau level. In fact, these predicted oscillations have recently been observed (Weiss *et al.*, 1988; Gerhardts *et al.*, 1989; Winkler *et al.*, 1989) with periodic potentials macroscopically produced by a superlattice, and are now known as Weiss oscillations. When the strength of the superlattice periodic potential is increased, the Landau levels are broadened significantly, the subbands merge, and the Weiss oscillations, *as well as the normal Shubnikov-de Haas oscillations*, are heavily damped (Schellnhuber and Obermair, 1980), and eventually disappear. This has also been observed experimentally (Beeton *et al.*, 1989).

In the opposite limit to that above, the magnetic field is treated as a perturbation on the periodic potential. In this regime, known as the Onsager regime, the magnetic transport properties are expected to be periodic in the magnetic flux coupled through each unit cell, i.e., periodic in magnetic field (as opposed to 1/B in the previous case) (Azbel, 1963; Harper, 1955; Hofstadter, 1976). While the required magnetic field is unreasonably high in normal semiconductor lattices, it is an observable effect in lateral surface superlattice (LSSL) periodic potentials (Ferry *et al.*, 1988; Ma *et al.*, 1989). Clearly, the observation of these effects, which are linear in the magnetic field, requires the phase coherence length of the electrons to be larger than the superlattice period. On the other hand, the limit being taken here is the tight-binding limit of the superlattice potential, and the transport can be expected to have strong similarities to more localized types of transport.

The structure of the bands can be obtained by solving Schrödinger's equation with the magnetic field as a perturbation. In the absence of the latter field, the energy structure of the superlattice mini-bands are given by

$$E = E_o\left[\cos(k_x a) + \cos(k_y a)\right] , \qquad (7)$$

With a magnetic field, described in the Landau gauge A=(0,Bx,0), the Peierl's substitution leads to the equation

$$\left\{\cos(k_x a) + \cos\left[\left(k_y - \frac{eBx}{\hbar}\right)a\right]\right\}\Psi(x,y) = \frac{E}{2E_o}\Psi(x,y) . \qquad 98)$$

The introduction of the wave function *ansatz* $\Psi(x,y)=g(x)\exp(ik_y y)$, and the substitutions x=ma, y=na (Hofstadter, 1976), lead then to the iterative equation

$$g(m+1) + g(m-1) + 2\cos(2\pi m\alpha - \nu)g(m) = Eg(m) , \qquad (9)$$

where $\alpha=eBa^2/h$, $\eta=k_y a$, and $E(=E/E_o)$ is the reduced energy. This equation is the Harper equation (Harper, 1955), and a study of the solutions of this equation have been discussed by Hofstadfer in some detail. The energy structure is periodic in α, which means that it is periodic in magnetic field, as this quantity is the ratio of the flux coupled through a unit cell to the quantum unit of flux h/e.

The source of the periodicity in magnetic field in the tight-binding limit can be understood in one sense by its relationship to the Aharonov-Bohm effect. Consider the presence of

magnetic translation operators connected with the periodicity of the lattice. In a periodic lattice, it is known that $\psi(x+a,y)=\exp(ik_x a)\psi(x,y)$, where $\psi(x,y)$ is the Bloch function corresponding to the superlattice in the absence of the field. If the magnetic field is normal to the layer, and the vector potential is taken (as above) in the Landau form A=(0,Bx,0), the motion of successive translations about a rectangle of unit cells (returning to the original point) leads to

$$T(-naj)T(-mai)T(naj)T(mai) = T(-naj-mai+naj+mai)\exp\left[i\int_0^{2\pi}dy\int_0^{2\pi}dx\frac{eB}{\hbar}\right]$$

$$= \exp(2\pi inm\alpha)\ , \qquad (10)$$

where i and j are unit vectors in the directions of the LSSL. In fact, the group theoretical arguments for the magnetic translation group have been worked out in some detail (Zak, 1964). In fact, the source of the periodicity can be understood quite easily with a simple Fermi energy argument. Recall that the periodicity of 1/B arises from the Fermi energy being forced down through the Landau levels, and the split bands of the levels. The 1/B behavior arises from the increase in the degeneracy of each level with the magnetic field and the spreading apart of the levels in the magnetic field. In the present case, the superlattice potential breaks up the conduction band into a series of mini-bands of width ΔE. The number of states in each mini-band is constant, but as the magnetic field is increased these bands are *depopulated by the magnetic field*. The conductance oscillations arise as the Fermi energy passes through each mini-band. In the absence of the periodic potential, the Fermi level is given by $E_{Fo}=\pi\hbar^2 n_s/m^*$. This energy range is the range of *allowed states*, and will be the sum of the widths of the occupied energy mini-bands in the presence of the periodic potential. Each periodic potential can accomodate $2/a^2$ electrons, so that the number of full and fractionally occupied mini-bands is just $n_s a^2/2$. In a sense, this number is the *filling factor* for the mini-bands. Thus, the average mini-band width may be found from

$$\Delta E_{av} = \frac{E_{Fo}}{ff} = \frac{2\pi\hbar^2}{m^* a^2} = \frac{eh}{m^*}\left(\frac{h}{ea^2}\right) = \hbar\omega_{c,o}\ . \qquad (11)$$

Here, $\omega_{c,o}$ is the cyclotron frequency corresponding to the magnetic field periodicity. The term in the parantheses in (11) is the flux coupled through each unit cell of the superlattice, and an integer number of flux quanta is coupled through each cell when the Landau level has been swept through a mini-band.

High Mobility Structures

Samples were prepared by the molecular beam epitaxy of a pseudomorphic InGaAs single quantum well structure on an undoped semi-insulating GaAs substrate. The InGaAs layer, 13.5 nm thick with 20% In content, was grown on an undoped GaAs buffer layer (0.5μm thick on top of a GaAs substrate). An undoped GaAs layer 15.7 nm thick was then grown, followed by a Si doped (1×10^{18} cm^{-3}) GaAs layer 40 nm thick. The carrier density in the pseudomorphic quantum well was 2×10^{11} cm^{-2} at 5 K. The first step of the processing is mesa isolation, in which a cross structure is defined by photolithography and etched about 200nm deep. After that, 200nm-thick AuGe/Ni/Au contacts were placed down by electron-beam evaporation and lift-off. These were alloyed at 450 C for 5 minutes in forming gas to form the ohmic contacts. The grid gate was patterned by electron beam lithography and lift-off processing. Finally the bonding pads were made by evaporation and lift-off of 300nm-thick Cr/Au. Figure 8 shows the grid gate itself. The grids are composed of 40nm lines on a 160nm pitch. The active area of the device structure is 10μm ×20μm.

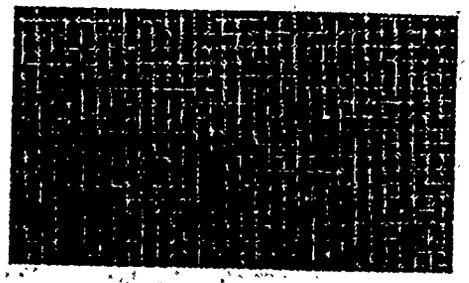


Fig. 8 The grid gate, pictured here, is placed at the cross of the mesa. It is composed of 40 nm lines on 160 nm centers.

In Fig. 9(a), the source conductance, in which the current is along the long axis of the sample, is shown at 5 K. It is apparent that there are significant fluctuations and a weak periodicity of the conductance that is present in the magnetoconductance. The structure is fully repeatable as long as the sample is maintained at low temperature, but does change somewhat upon heating and recooling of the sample. The applied longitudinal voltage on the sample was only 1 mV over the entire range, so that the amplitude of the fluctuations in conductance is about $0.1e^2/h$. We have Fourier transformed the conductance in order to bring out the underlying periodicity, and this is shown in Fig.2(b). The d.c. component has been removed to enhance the signal, but there is still a low frequency component that arises from the weak magnetoresistance variations in the sample. In this latter figure, we have marked the range expected from estimating the frequency that would arise from the fabricated superlattice periodicity, allowing for the possibility that the actual flux coupled to each well varies due to the finite width of the individual gate lines. A second set of weaker peaks is observed near the second harmonic. Whether these relate to *h/2e* oscillations seen in weak localization in rings or are simply the second harmonic is not discernible at this time.

In Fig. 9(b), the dominant peak is approximately 6.9 T^{-1}, which corresponds to a unit cell whose side is 176 nm, while actual scanning electron microscopy measurement of the sample suggests a number closer to 168 nm. Considering the quality of the data, this agreement is quite good. A secondary peak is also observed which lies very close to the first, and within the range of the spread expected from the fabricated grid. This secondary peak could arise from a slightly different spacing over part of the grid, which could arise from differences across the grid in the linearity of the electron beam sweep during e-beam lithography.

The source of the conductance fluctuations in the data, and the relatively large amplitude of these fluctuations compared to that expected for universal conductance fluctuations (UCF), is also quite interesting. The UCF is regarded as arising from quantum interference of different modes, or paths, of the electrons as the chemical potential or the magnetic field is varied, so that interference effects appear in the end-to-end conductance, and are related to the Aharonov-Bohm effect. In general, the observations of these effects in the past have been confined to quasi-one-dimensional conductors. The structure we are investigating is considerably larger than the estimate of the inelastic mean free path. UCF is generally found to decay faster than linearly in quasi-one dimensional wires and a sequence of rings. The inelastic mean free path inferred below is such that the amplitude observed for the fluctuations is of the order of magnitude expected from these studies.

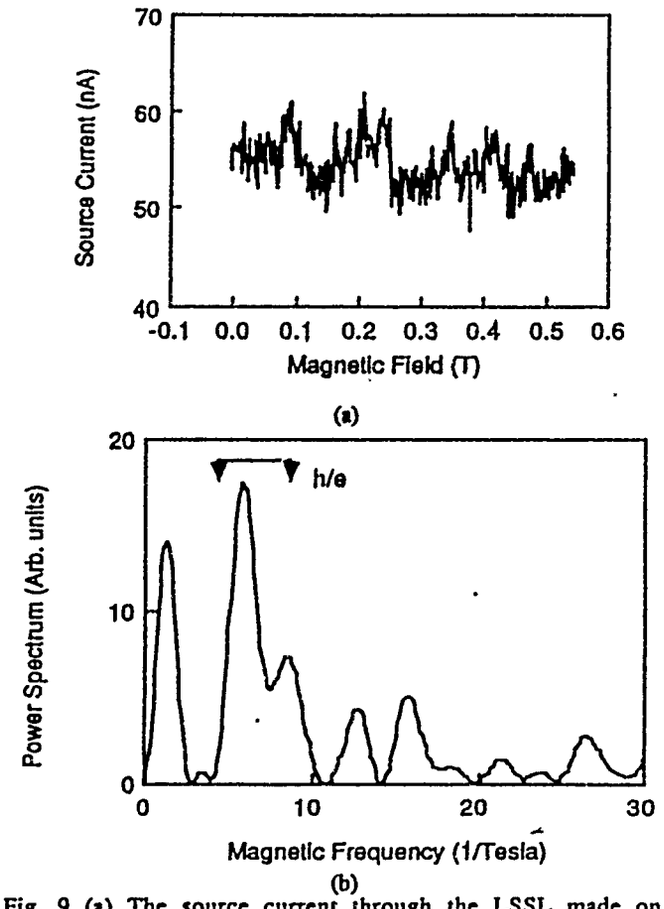


Fig. 9 (a) The source current through the LSSL made on a InGaAs/GaAs high mobility sample. (b) Fourier transform of the magnetoconductance, showing the peak at the value h/e flux per unit cell of the LSSL.

In Fig. 10, the correlation function calculated from the data of Fig. 9(a) is plotted as a function of the magnetic field separation. This clearly evidences the expected exponential behavior, and has a correlation "length" of 85 G. This translates to a fundamental active area which is described by a inelastic mean free path of 0.7 μm. The size of this area also fits well the amplitude of the oscillations in terms of the number of basic area that are being ensemble averaged. This averaging scales here exactly as that expected for UCF.

We suggest that the conductance fluctuation effect may be explained in terms of the fractal energy structures arising from application of a magnetic field to an electron gas in a two-dimensional periodic potential. As mentioned above,there are two fundamental lengths in the problem: the periodicity of the two-dimensional periodic potential, which here has n=m=*a*, and the magnetic length $L_m=(\hbar/eB)^{1/2}$. The Hamiltonian can be solved for its eigenvalues when these two lengths are rationally related as $a/l_m=p/q$, where p and q are integers. This property by itself leads to a magnetoconductance that exhibits fluctuations of the order of e^2/h, even in two-dimensional systems. This was probed with a theoretical calculations. For both hard-wall and periodic boundary conditions, it was found that the existence of magnetoconductance fluctuations and periodicity in the flux coupled through each unit cell. These results are consistent with the interpretation expected from Hofstadter's work that the fractal nature of the eigenvalues for this system imply that small changes in the magnetic field produce significant changes in the eigenfunctions and therefore produce significant changes in the quantum interference in the structure. While the effect is essentially the same as that of UCF, the source of the effect has a different origin, arising here from the superlattice potentials.

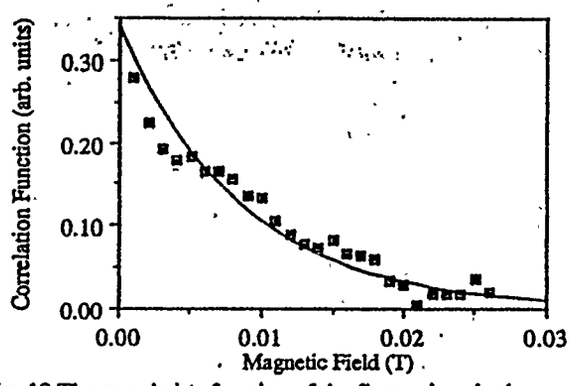


Fig. 10 The correlation function of the fluctuations in the conductance shown in Fig. 1(a). The expected exponential behavior leads to a correlation "length" of 85 G.

Low Mobility Structures

We have also incorporated the LSSL gates into structures fabricated on material normally used for MESFETs. Here, the active layer is typically a 60 nm thick, epitaxial layer grown by vapor-phase epitaxy on a lightly-doped substrate. The epitaxial layer is doped to 1.5×10^{18} cm^{-3}. It has been demonstrated previously that such layers will show quasi-two-dimensional behavior at low temperatures when the channel is biased near pinchoff. We estimate (see below) that the inelastic mean free path in these structures is of the order of 0.2–0.6 μm at 5 K, which is considerably less than either dimension in the plane of the sample.

In Fig. 11, the source conductance and Hall voltage are both shown so that the presence of the negative magnetoresistance at low magnetic fields can be seen, which is clear evidence of weak localization. The source current shows sharp drops at regular values of the magnetic field, and sharp changes in the Hall voltage are often correlated with these. The source current drops occur at integral multiples of a flux quantum coupled through each unit cell of the surface superlattice, which we interpret to be periodic replicas of the negative magnetoresistance at zero magnetic field, and hence periodic replicas of the weak localization of the electrons. The current drops at the first and third flux quanta are very weak, being a reduction of only about 1-2% in the current, and are not easily distinguished in the data of Fig. 11, but can be measured by separating the data from the background, which is done for the evaluation described below. On the other hand, we will also show data from another sample below in which the entire series of peaks is easily discerned. The basic periodicity in the

conductance is about 1500 (±100) Gauss, which corresponds to 165 nm periodicity, consistent with the grid.

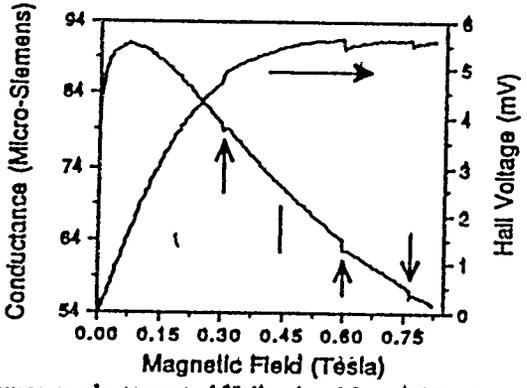


Fig. 11 Source conductance and Hall voltage for a LSSL made on a MESFET material at low temperatures. We show the jumps in the conductance and Hall voltage that are seen in these samples.

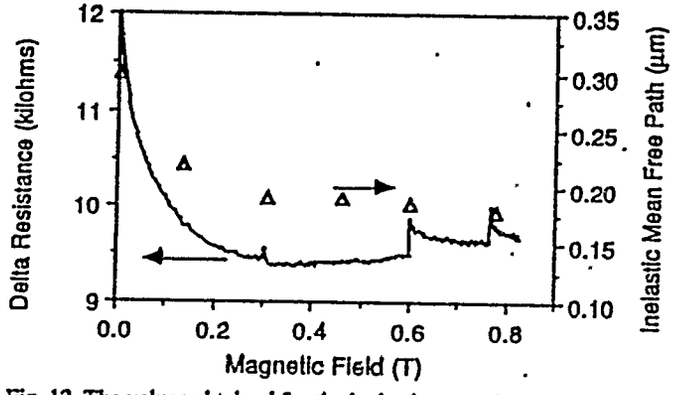


Fig. 12 The values obtained for the inelastic mean free path.

Using a weak localization theory (Altshuler *et al.*, 1986), we can estimate the inelastic mean free path near these jumps from the negative magnetoresistance in the source current following the jump. The value of the inelastic mean free path inferred from a data fit to several drops in the conductance at the various integer multiples of a flux quantum (coupled to each well) is nearly the same for a given sample and gate voltage. In Fig. 12, the value of the inelastic mean free path computed for each of the drops (six), found in the data of Fig. 11, are plotted on the same graph with the magnetoresistance data. The value found in this sample is about 0.2 μm and the uncertainty of the fit makes the data consistent with a constant value. However, it also appears that the value of the mean free path is tending to lock to the periodicity distance of the superlattice itself, reinforcing the idea that localization within the superlattice is important for the transport properties.

While we do not understand the jumps in the Hall voltage, if the complicated band structure mentioned above is invoked, Thouless (1984) has speculated on the possibilities of jumps in the Hall conductance due to mixing of Landau levels and/or subbands, and this should occur at preferential values of the number of flux quantum per well. Moreover, the coupling of various orbits by Bragg scattering from the superlattice should occur at field values such that $\Phi=2n\Phi_o$.

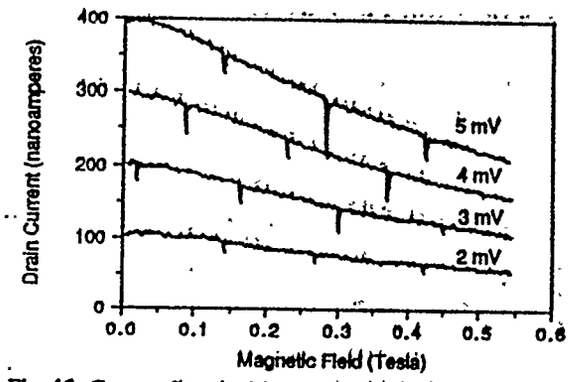


Fig. 13 Current flow in a structure with inelastic mean free path of about 0.55±0.1 μm. The parameter is the source-drain potential (1 mV corresponds to 5.7 μV per unit cell potential drop).

In Fig. 13, measurements made on a MESFET layer with a slightly thinner epitaxial layer (about 50 nm), and a significantly, higher mobility are shown. Here, negative magnetoresistance occurs only at higher values of the source-drain voltage (2 mV source-drain potential corresponds to approximately 11 μV drop across each cell of the superlattice, while the superlattice potential is estimated to be about 100 mV peak-to-peak amplitude). The onset of the magnetic field induced localization generally does not occur at integer values of the flux coupled through each unit cell (except for the curves with 2 mV and 5 mV potential drop). Rather, the localization dips occur approximately periodically through the magnetic field range. Their periodicity is found to be approximately 0.14 T for all of the curves (Fig. 14), which corresponds to a periodicity of about 170 nm in the superlattice. The inelastic mean free path, estimated from the negative magnetoresistance that is found at each local minimum in conductivity, is about 0.55±0.1 μm for this sample. The conductance, at zero magnetic field, varies from about 40 μS, at 2 mV, to 80 μS at 5 mV. The (spin degenerate) conductance expected from the Landauer formula for a single channel is 77.5 μS[22] so that it is clear that the structure is not conducting strongly in a free electron fashion. Rather, the conductance is quite likely to be localized. The behavior shown in Fig. 13 is for a temperature 5.7 K, and the effect is not observed for a temperature of 15 K. It is also interesting that the increase of the drain potential from 2 mV to 5 mV has shifted the spectrum almost exactly one cell, bringing the spectra back into commensurability.

The shifts introduced by the source-drain potential can be understood as follows. As it is well known, for a two-dimentional lattice with a magnetic field applied as a perturation, the tight-binding approximation and the Peierl's substitution leads to Harper's equation (9). With a source-drain voltage applied, the energy levels are not only shifted along the channel, in reference to their values at the source end, but also distorted slightly within each unit cell. It is the latter effects that can give rise to the shifts noted above. For this, we consider the electric field in the vector potential gauge, and oriented along the y-axis, so that $A=(0,Bx+eE\tau,0)$, where the time dependence has been averaged over the momentum relaxation process. For this vector potential, equation (9) is modified to

$$g(m+1) + g(m-1) + 2\cos(2\pi m\alpha - \omega_B\tau - \nu)g(m) = \varepsilon g(m) \qquad (12)$$

where $\omega_B = eEa/\hbar$ is the Bloch frequency. Here, the presence of the electric field, and the resulting drift velocity shifts the y-momentum contained in ν, and shifts the zero of the cosine function yielding the energy levels. This shift is reflected in the conductance itself. This can be checked by plotting the shift introduced for Fig. 14 as a function of the source-drain potential. This is shown in Fig. 15.

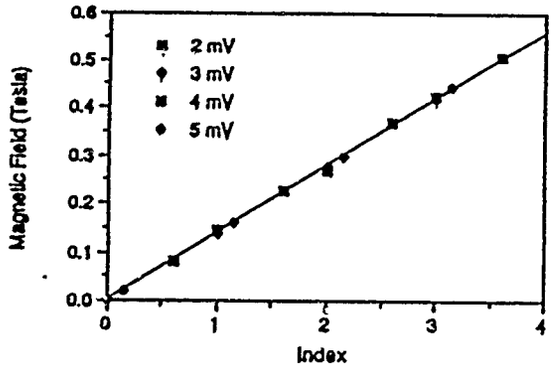


Fig. 14 The value of the magnetic field at which a conductance dip is observed is plotted as a function of its "index" – the rank order of the dips plus an offset in magnetic field. The offset is a constant for each value of the source-drain bias.

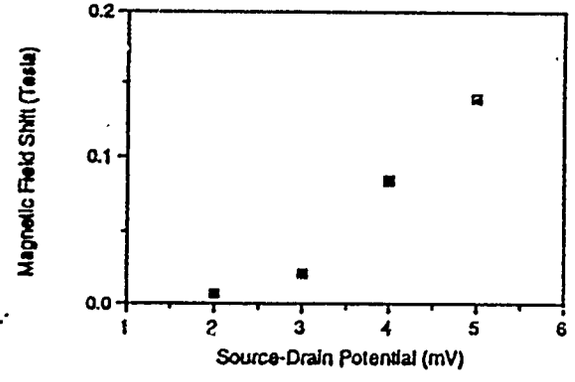


Fig. 15 The value of the magnetic field shift required for the data of Fig. 5 in order to align the curves in Fig. 14.

The shift in Fig. 15 becomes almost linear in source-drain potential for values of this latter quantity above 3 mV. It was remarked above that the conductance of the channel did not turn on linearly until this value as well, so that it is likely that the transport is strongly localized at the lower values of the source-drain potential. The linear portion of the curve in Fig. 15 can be used with the behavior of equation (12) to estimate the scattering time τ, and this gives a value of approximately 2.8×10^{-10} sec. If this value is taken together with the inelastic mean free path, a diffusion constant of 11 cm^2/sec is inferred, which corresponds to a mobility of about 2.5×10^4 cm^2/V-s. The observed mobility is about 25% larger than this, but the numbers give a consistent order-of-magnitude argument to suggest that the shifts observed in the weak localization effects (in magnetic field) are related to the increasing drift momentum caused by the source-drain potential.

REFERENCES

Altshuler, B. L., Khmel'nitzkii, D., Larkin, A. I., and Lee, P. A., 1986, Magnetoresistance and Hall Effect in a Disordered Two-Dimensional Electron Gas, Phys. Rev. B, 22:5142.

Averin, D. V., and Likharev, K. K., 1987, Coulomb Blockade of Single-Electron Tunneling, and Coherent Oscillations inSmall Tunnel Junctions, J. Low Temp. Phys., 62:345.

Averin, D. V., and Likharev, K. K., 1990, Single Electronics: A Correlated Transfer of Single Electrons and Cooper Pairs in Systems of Small Tunnel Junctions, in "Quantum Effects in Small Disordered Systems," B. L. Altshuler, P. A. Lee, and R. A. Webb, Eds., Academic Press, New York, in press.

Azbel, M. Ya., 1963, Quantization of Quasi-Particles with a Periodic DispersionLaw in a Strong Magnetic Field, J. Exptl. Theor. Phys., 44:980 [transl. in Sov. Phys. JETP, 17:665].

Azbel, M. Ya., 1964, Energy Spectrum of a Conduction Electron in a Magnetic Field, J. Exp. Theor. Phys., 46:929 [transl. in Sov. Phys. JETP, 19:634].

Baccarani, G., Wordeman, M. R., and Dennard, R. H., 1984, Generalized Scaling Theory and Its Application to a 1/4 Micrometer MOSFET Design, IEEE Trans. Electron Dev., 31:452.

Barker, J. R., and Ferry, D. K., 1981, On the Physics and Modeling of Small Semiconductor Devices, Sol.-State Electron., 23:519; 23:531.

Beeton, P., Alves, E. S., Hennini, M., Eaves, L., Main, P. C., Hughes, O. H., Toombs, G. A., Beaumont, S. P., and Wilkinson, C. D. W., 1989, Proc. Symposium on New Phenomena in Mesoscopic Systems, Kona, Hawaii, Japan Society for the Promotion of Science, unpublished.

Bernstein, G., and Ferry, D. K., 1986, Electron Beam Lithographic Fabrication of Ultra-Submicron Gate GaAs MESFETs, Superlatt. Microstruct., 2:147.

Bernstein, G., and Ferry, D. K., 1988, Velocity Overshoot in Ultra-Short-Gate-Length GaAs MESFETs, IEEE Electron Dev. Letters, 35:887.

Büttiker, M., Zero-Current Persistent Potential Drop Across Small-Capacitance Josephson Junctions, Phys. Rev. B, 36:3548.

de la Haussaye, P. R., Allee, D. R., Pao, Y. C., Schlom, D. G., Harris, J. S., and Pease, R. F. W., 1988, Electron Saturation Velocity Variation in InGaAs and GaAs Channel MODFETs for Gate Lengths to 550 Å, IEEE Electron Dev. Letters, 9:148.

Ferry, D. K., 1982, Material Considerations for Advances in Submicron Very Large Scale Integration, in "Advances in Electronics and Electron Physics," C. Marton, Ed., Vol. 58, Academic Press, New York, pp. 312-390.

Ferry, D. K., Bernstein, G., Puechner, R., Ma, J., Kriman, A. M., Mezenner, R., Liu, W.-P., Maracas, G. N., and Chamberlin, R., 1988, Magnetoconductance in Lateral Surface Superlattices, in "High Magnetic Fields in Semiconductor Physics II," G. Landwehr, Ed., Springer-Verlag, Heidelberg, pp. 344-352.

Fischetti, M. V., DiMaria, D. J., Dori, L., Batey, J. Tierney, E., and Stasiak, J., 1987, Phys. Rev. B 35:4404.

Geerligs, L. J., Anderegg, V. P., Holweg, P. A. M., Mooij, J. E., Pothier, H., Esteve, D., Urbina, C., and Devoret, M. H., 1990, Frequency-Locked Turnstile Device for Single Electrons, submitted for publication.

Gerhardts, R. R., Weiss, D., and von Klitzing, K., 1989, Novel Magnetoresistance Oscillations in a Periodically Modulated Two-Dimensional Electron Gas, Phys. Rev. Letters, 62:1173.

Han, J., Ferry, D. K., and Newman, P., Ultra-Submicron Gate AlGaAs/GaAs HEMTs, IEEE Electron Dev. Letters, in press.

Harper, P. G., 1955, Single Band Motion of Conduction Electrons in a Uniform Magnetic Field, Proc. Phys. Soc. (London), A68:874.

Hauser, J. R., 1967, Characteristics of Junction Field-Effect Devices with Small Channel Length-to-Width Ratios, Sol. State Electron. 10:577.

Hofstadter, D. R., 1976, Energy Levels and Wave Functions of Bloch Electrons in Rational and Irrational Magnetic Fields, Phys. Rev. B, 14:2239.

Ishibashi, A., Funato, K., and Mori, Y., 1988, Heterointerface Field Effect Transistor with 200 Å-Long Gate, Jpn. J. Appl. Phys., 27:L2382.

Jin, Y., Mailly, D., Carcenac, F., Ettiene, B., and Launois, H., Nanostructures in Gallium Arsenide TEGFET, Microelectron. Engr., 6:195.

Lewicky, G., and Maserjianl, J., 1975, J. Appl. Phys., 46:3032.

Mikkelson, J. M., Hall, L. A., Malhotra, A. K., Seccombe, S. D., and Wilson, M. S., 1981, IEEE J. Sol. State Circ., 16:542.

Patrick, W., Mackie, W. S., Beaumont, S. P., Wilkinson, C. D. W., and Oxley, C. H., 1985, Very Short Gate Length GaAs MESFETs, IEEE Electron Dev. Letters, 6:471.

Pippard, A. B., 1964, Quantization of Coupled Orbits in Metals: II. The Two-Dimensional Network, with Special Reference to the Properties of Zinc, Phil. Trans. Roy. Soc. (London), A68:317.

Rauh, A., Wannier, G. H., and Obermair, G., 1974, Bloch Electrons in Irrational Magnetic Fields, Phys. Stat. Sol. (b), 63:215.

Ryan, J. M., Han, J., Kriman, A. M., Ferry, D. K., and Newman, P., 1989a, Overshoot Saturation in Ultra-Short Channel FETs due to Minimum Acceleration Lengths, In "Nanostructure Physics and Fabrication," M. A. Reed and W. P. Kirk, Eds., Academic Press, New York, pp. 195-199.

Ryan, J. M., Han, J., Kriman, A. M., and Ferry, D. K., 1989b, Overshoot Saturation in Ultra-Submicron FETs due to Minimum Acceleration Lengths," Sol. State Electron. 32:1609.

Ryan, J. M., Han, J., Kriman, A. M., Ferry, D. K., and Newman, P., 1990, Scaling of Transconductance in Ultra-Submicron GaAs MESFETs and HEMTs, SPIE Conf. Proc. 1284:in press.

Sai-Halasz, G. A., Wordeman, M. R., Kern, D. P., Rishton, S., and Ganin, E., 1988, High Transconductance and Velocity Overshoot in NMOS Devices at the 0.1 μm Gate-Length Level, IEEE Electron Dev. Letters, 9:464.

Schellnhuber, H. J., and Obermair, G. M., 1980, First-Principles Calculation of Diagmagnetic Band Structure, Phys. Rev. Letters, 45:276.

Thouless, D., 1984, Quantized Hall Effect in Two-Dimensional Periodic Potentials, Phys. Repts., 110:279.

Troutman, R., 1979, Drain-Induced Barrier Lowering in Short Channel MOSFETs, IEEE Trans. Electron Dev., 26:461.

Weiss, D., von Klitzing, K., Ploog, K., and Weimann, G., 1988, New Magnetotransport Phenomenon in a Two-Dimensional Electron Gas in the Presence of a Weak Periodic Submicrometer Potential, In "High Magnetic Fields in Semiconductor Physics II," G. Landwehr, Ed., Springer-Verlag, Heidelberg, pp. 357-365.

Winkler, R. W., Kotthaus, J. P., and Ploog, K., 1989, Landau-Band Conductivity in a Two-Dimensional Electron System Modulated by an Artificial One-Dimensional Superlattice Potential, Phys. Rev. Letters, 62:1177.

Zak, J., 1964, Magnetic Translation Group, Phys. Rev., 134:A1602; 134:A1607.

4

APPROACHES TO TRANSPORT IN SEMICONDUCTOR NANOSTRUCTURES

V. Pevzner, P. Sols, and Karl Hess

Beckman Institute
University of Illinois at Urbana-Champaign
Urbana, IL 61801

INTRODUCTION

Our view of electronic transport in nanostructures is based on semiconductor technology and aims at the development of numerical methods that are applicable in virtually arbitrary geometrical structures. By semiconductor technology we refer to existing technology as it has been presented in this summer school but also to future technological possibilities that appear on the horizon. The conventional methods of pattern fabrication appear to be able to produce systems that show mesoscopic effects only at very low temperatures. However, it is conceivable that structures can be fabricated, even on silicon, that will exhibit waveguide like properties at 77 K and maybe even at room temperature. Pattern generation by tunneling microscopy techniques has been demonstrated on silicon surfaces with feature sizes of around 100 Å and below (Loenen *et al.*, 1989). Patterns of these sizes promise not only the information of the library of congress on square inch dimensions but also the smallness which is necessary to produce electron waveguides or general quantum interference phenomena at high temperatures. In fact, recent estimates show that silicon-silicon dioxide structures written with tunneling microscopy methods will enable us to investigate a wide range of quantum effects far above the temperatures where they are observed now (Lyding *et al.*, 1990).

The use of semiconductor technology also guarantees that imperfections can be largely avoided and impurities will play much less of a role compared to metallic systems. Of course there are unavoidable imperfections. Phonon scattering will determine in certain ranges the limiting size that still shows significant quantum interference effects determined by geometry. This leaves us with the following transport problem: Few electrons (of the order of one to hundreds) propagate in geometries that are small enough to lead to quantum interference and interact with phonons in the weak coupling limit as well with few impurities (of the order of one to hundreds). All band-structure effects are dealt with by the effective mass theorem and the effective masses are typically one tenth to one hundredth of the free electron mass. It is precisely this type of problem for which we attempt to develop numerical techniques which describe transport.

As a further simplification, we concentrate on cases in which transport can be described by the framework developed by Landauer (1957, 1970, 1988) and Büttiker (1987, 1988a,1988b), that is, cases where the transmission coefficient plays a crucial role. Since we want to find solutions for complex geometrical structures, we involve the use of numerical algorithms and large computational resources. These lecture notes have developed from our first attempts in this area and are therefore not entirely cohesive and systematic; they are rather illustrations of interesting topics.

TYPICAL TRANSPORT PROBLEMS

A T-shaped structure, exhibiting interference, can serve as a typical example of a mesoscopic transport problem as defined above. In this section, we will highlight an explicit approach to solve this problem for one dimensional chains. This approach can be generalized to more complicated geometries (Guinea and Verges, 1987; Sols *et al.*, 1989a, 1989b; Datta, 1989) and also can form the basis for a treatment of the electron-phonon interaction. Here it gives us an opportunity for some basic definitions and for introducing the well known and very useful tight-binding formalism. In addition, it forms an example of how powerful analytical methods can be combined with numerical approaches such as finite difference solutions.

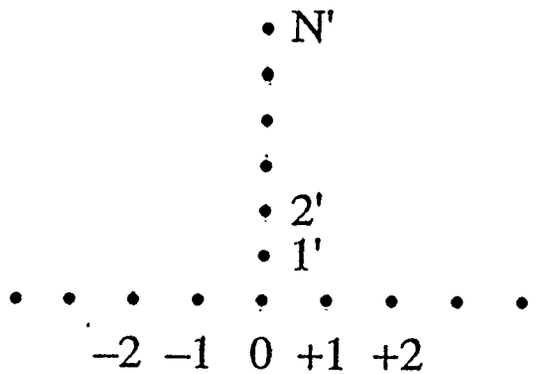


Fig. 1. Schematic of a one-dimensional chain T-structure.

Consider the T-structure of Fig. 1 with sites (atoms) infinitely extended on the main chain and a finite side stub. A tight-binding Hamiltonian for this structure can be written as (Büttiker, 1987)

$$H = \sum_{l=-\infty}^{\infty} \{ \varepsilon_l \, | \, l \rangle\langle l \, | + \Delta (| \, l \rangle\langle l+1 | + | \, l+1 \rangle\langle l \, |) \} + \sum_{l'=1}^{N'} \{ \varepsilon_{l'} \, | \, l' \rangle\langle l' \, | + \Delta (| \, l' \rangle\langle l'+1 | + | \, l'+1 \rangle\langle l' \, |) \} + \Delta (| \, 0 \rangle\langle 1' \, | + | \, 1' \rangle\langle 0 \, |) \quad , \tag{1}$$

where the last term couples the side-stub to the main arm. This coupling can be treated by perturbation theory to all orders. If one denotes the Green's function for the solution of the problem with $\Delta=0$ by G_0 then the Dyson equation

$$G = G_0 + G_0 \, V \, G \tag{2}$$

gives G in terms of G_o and $V = \Delta(|0\rangle\langle 1'| + |1'\rangle\langle 0|)$. The relevant Green's functions G_o (actually retarded, $G_o^{(+)}(E)$) are well known (Economou, 1983) and their matrix elements are analytically simple for the main chain

$$G_o(l,m;E) \equiv \langle l|G_o^{(+)}(E)|m\rangle = \frac{i}{2\Delta\sin\theta}\, e^{i\,|l-m|\theta} \tag{3}$$

and for the side arm

$$G_o^{(+)}(1',l';E) = G_o^{(+)}(l',1;E) = \frac{\sin(N'-l'+1)\theta}{\Delta\sin(N'+1)\theta} \tag{4}$$

where θ is given implicitly by

$$(E - E_l) = 2\,\Delta\cos\theta\ . \tag{5}$$

The derivation of these matrix elements (Guinea and Verges, 1987; Sols *et al.*, 1989a; Economou, 1983) is not entirely simple but algebraically straightforward. The matrix elements for G can then be easily obtained. With the coupling equal everywhere, we have

$$\langle l|G^{(+)}(E)|m\rangle = \langle l|G_o^{(+)}(E)|m\rangle + \langle l|G_o^{(+)}(E)\,\Delta(|0\rangle\langle 1'| + |1'\rangle\langle 0|)G^{(+)}(E)|m\rangle\ , \tag{6}$$

which is easily solved algebraically by using completeness of the site kets:

$$G^{(+)}(l,m;E) = \frac{G_o^{(+)}(l,m,E)}{1 - \Delta^2\, G_o^{(+)}(1',1',E)\, G_o^{(+)}(0,0,E)}. \tag{7}$$

This matrix element is directly related to the transmission coefficient and many other important quantities. The approach also can be generalized to include scattering by phonons. To show this, we derive the transmission coefficient for such a problem in detail.

SCATTERING IN THE PRESENCE OF DISSIPATION

General Relations

In this section we derive some general expressions that describe the effect of dissipation on the transport of electrons in mesoscopic systems. We introduce two main simplifications in our analysis. First, we employ a one-electron picture and, secondly, we assume that the dissipation takes place in a finite region of space. Although the first approximation prevents us from studying the role of the Pauli exclusion principle and Coulomb correlation in dissipation, we expect it to keep much of the essential physics and we note in this regard that the one-electron picture has historically proven to be quite useful in the study of transport in semiconductors (Mahan, 1981). The second approximation on the finite-spatial extent of the inelastic interaction allows us to adopt a simple scattering picture in which the channels for asymptotic propagation correspond to the various transverse modes of the different leads that are connected to the nanostructure under study. The resulting scattering problem can be described as follows: An electron comes from lead a in transverse mode m and can be transmitted with a certain probability into, e.g., mode n of lead b, after having interacted with the boundaries, impurities, and phonons in the central region ("sample") where scattering takes place. We are essentially following the Landauer approach in which the resistance of a given sample is viewed as a direct consequence of its scattering properties (Landauer, 1957). The main difference is that we are now including the possibility of inelastic processes within (or in the vicinity of) the sample. As in the standard Landauer picture, we assume that the electron reservoirs introduce additional randomization of the relative phase between the electron waves that enter and leave the reservoir. The mesoscopic regime of electron transport is characterized by the preservation of coherence in a given region where interference of the electron wave builds up. We wish to study how that interference is destroyed when inelastic interactions are present in the same region where elastic scattering by boundaries (and, possibly, impurities) takes place. The assumption that dissipation occurs in a finite region of space simplifies the scattering problem considerably, since the asymptotic scattering channels can be known exactly. A more realistic model in terms of extended phonons that in turn can be scattered by the structure would complicate the problem without necessarily adding too much new insight on the interplay between quantum interference and dissipation. In addition, we wish to point out that a model of localized phonons is quite adequate in some situations in which, due to the specifics of the structure, phonon modes develop that have most of their amplitude in the scattering region. These are cases in which the lattice vibrations are sensitive to the geometry, also because of their wave nature. In this sense, we can assert that the geometry affects the electron motion both by changing its wave function and by modifying the phonons that tend to destroy the coherence. Whether this modification of the phonon mode enhances or inhibits the loss of electron phase coherence remains to be seen and certainly constitutes a question of great interest.

Stone and Szafer (1988) have recently studied the electron elastic scattering in the general structure shown in Fig. 1 and rederived the Landauer formula as generalized by Büttiker (1988a,1988b) to the case of many leads and channels. They derived the following relation between the electron Green's function and the transmission and reflection coefficients:

$$G_{nm}^{(+)}(x_a,x_a') = \frac{-i}{\hbar v_{ma}}\{\delta_{ba}\exp[ik_{ma}(x_a'-x_a)] + \left(\frac{k_{ma}}{k_{na}}\right)^{1/2} r_{aa,nm}\exp(ik_{na}x_a+ik_{ma}x_a')\} \tag{8a}$$

$$G_{nm}^{(+)}(x_b,x_a) = \frac{-i}{\hbar v_{ma}}\, t_{nm,ba}\left(\frac{k_{ma}}{k_{nb}}\right)^{1/2}\exp(ik_{nb}x_b+ik_{ma}x_a)\ , \tag{8b}$$

where $t_{nm,\,ba}$ is the probability amplitude that an incident electron in transverse mode m of lead a is transmitted into mode n of lead b, $r_{nn,\,ba}$ is the reflection coefficient to go from mode a to mode b within lead n, and k_{ma} and v_{ma} are the electron wavevector and velocity in mode a of lead m at energy E [note that the index convention differs from that used by Stone and Szafer (1988)]. Equation (8) generalizes a relation previously obtained by Fisher and Lee (1981) for one-dimensional scattering. The Green's functions are defined as

$$G_{nm}^{(+)}(x_b,x_a;E) = \langle x_b,\chi_n|\, G_0^{(+)}(E)\,|x_a,\chi_m\rangle\ , \tag{9a}$$

$$G_0^{(+)}(E) \equiv (E - H_0 + i0^+)^{-1}\ , \tag{9b}$$

where $\chi_m(y_a)$ is the wave function for the transverse mode m in lead a, and $x_a=(x_a,y_a)$ is a point in lead a. In (8a), $x_a'>x_a$ must be taken, where, by convention, x_a grows in the outward direction. H_0 is the one-electron Hamiltonian that includes the effect of boundaries and impurities.

We argue at this point that dissipation can be included by introducing in (8) and (9) extra-indices associated to the states of the phonon bath $|\alpha\rangle$, $|\beta\rangle$, etc. The scattering states of our complete physical system require now the additional specification of the bath state. Thus,

for instance, $t_{nm,ba,\beta\alpha}$ is the probability amplitude that an electron coming from lead a in transverse mode m, with the bath initially in state $|\alpha>$, is transmitted into channel n of lead b, leaving the bath in state $|\beta>$, at a given total energy $E = E_i + e_\alpha = E_f + e_\beta$, where E_i and E_f are the initial and final electron energies. The total system composed by the electron plus the bath is a Hamiltonian system where energy is conserved. Dissipation will be a manifestation of the lack of control on what the bath does, which mathematically is described by tracing out the bath coordinates or, in path-integral language, by summing over all histories of the bath that are compatible with a given history of the electron. This is the approach to dissipation that has been emphasized by Leggett and coworkers (1983, 1987) and that follows the seminal work of Feynman and Vernon (1963). The work on dissipation in quantum mechanics has been mostly based on path-integral formulations. In this section, we show that similar ideas can be implemented in the framework of Schrodinger mechanics.

We focus for the moment on the transmission probability and rewrite (8b) with new bath indices.

$$t_{nm,ba,\beta a} = ih(v_{nb\beta}\, v_{ma\alpha})^{1/2}$$
$$\exp(-ik_{nb\beta}x_b - ik_{ma\alpha}x_a)\, <x_b,\chi_{n;\beta}|\, G^{(+)}(E)\, |x_a,\chi_{m;\alpha}> \quad , \tag{10}$$

where $G^{(+)}(E)$ is also given by (9b) with the difference that the one-electron Hamiltonian H_o must be replaced by the total Hamiltonian

$$H = H_o + H_B + V \ ,$$

where H_B describes the isolated phonon bath and V is the electron-phonon interaction. We note that $G^{(+)}(E)$ can be written

$$G^{(+)}(E) = \frac{-i}{\hbar} \int_{-\infty}^{\infty} dt\ e^{iEt/\hbar}\, \theta(t)\, e^{-iHt/\hbar}\, e^{-\eta|t|} \tag{11}$$

where $\eta \to 0^+$ and $\theta(t)$ is the step function. On the other hand, we introduce for convenience a field-theoretical description of the electron, in which

$$|x_a,\chi_m;\alpha> \ \equiv\ \psi_m^+(x_a)\, |0,\alpha>$$

where $|0>$ is the vacuum of electrons and the field operator $\psi_{n,}^+(x_a)$ creates an electron in the transverse mode m of lead m with longitudinal coordinate x_a. As a result, the transmission probability can be written:

$$T_{nm,ba,\beta a}(E) \equiv |t_{nm,ba,\beta a}(E)|^2$$
$$= v_{nb\beta}\, v_{ma\alpha} \int_{-\infty}^{\infty} ds \int_{-\infty}^{\infty} dt\ e^{iE(t-s)/\hbar}$$
$$\theta(s)\, \theta(t)\, e^{-ie_\beta(t_0-s_0+t)/\hbar}\, e^{ie_\alpha(t_0-s_0+s)/\hbar}$$
$$<0,\alpha|\psi_m(x_a,s_0-s)\psi_n^+(x_b,s_0)|0,\beta><0,\beta|\psi_n(x_b,t_0+t)\psi_m^+(x_a,t_0)|0,\alpha> \tag{12}$$

where the Heisenberg field operators have been introduced:

$$\psi_m(x_a,t) = e^{iHt/\hbar}\, \psi_m(x_a)\, e^{-iHt/\hbar} \tag{13}$$

In deriving (13) we have used the fact that $H|0,\alpha> = e_\alpha|0,\alpha>$. Due to time translational invariance, (13) is independent of the variables s_o and t_o, which have been introduced for convenience.

We are interested in quantities that describe the reduced dynamics of the electron, i.e., the dynamics that result from tracing out the bath coordinates. In particular, we would like to know the inelastic transmission probability $T_{nm,ba}(E_f,E_i)$ defined in such a way that the probability that an incident electron in mode m of lead a with energy E_i is transmitted to mode n of lead b with energy between E_f and $E_f + dE_f$ is $T_{nm,\,ba}(E_f,E_i)dE_f$. Such a probability distribution must be given by the relation

$$T_{nm,ba}(E_f,E_i) = Z^{-1} \sum_\alpha e^{-\beta e_\alpha} \sum_\beta T_{nm,ba,\beta\alpha}(E_i+e_\alpha)\, \delta(E_f-E_i+e_\beta-e_\alpha) \ , \tag{14}$$

where $Z=\sum_\alpha e^{-\beta e_\alpha}$ is the partition function of the phonon bath and the delta function guarantees the conservation of the total energy $E=E_i + e_\alpha = E_f + e_\beta$. In (14), it is assumed that the phonons are initially in thermal equilibrium and a sum is performed over all possible final bath states $|\beta>$. The phenomenological description that results from tracing out the phonon coordinates contains the essence of dissipation.

We now introduce (12) in (14) and take advantage of the independence in (12) of the variables t_o and s_o (and in particular on their difference) and write the delta function as

$$\delta(E_f-E_i+e_\beta-e_\alpha) = \frac{1}{2\pi\hbar} \int_{-\infty}^{\infty} d(t_0-s_0)\ e^{i(E_f-E_i+e_\beta-e_\alpha)(t_0-s_0)/\hbar} \ .$$

The terms in the phase that are proportional to $(e_\beta - e_\alpha)$ (note also the presence of the total energy E) cancel out and (12) becomes

$$T_{nm,ba}(E_f,E_i) = \frac{v_f v_i}{2\pi\hbar} Z^{-1} \sum_\alpha e^{-\beta e_\alpha} \sum_\beta \int \int_{-\infty}^{\infty} ds\, dt\, d(t_0-s_0)$$
$$e^{i[(E_f-E_i)(t_0-s_0)+E_f t-E_i s]/\hbar}\, \theta(t)\, \theta(s)$$
$$<0,\alpha|\psi_m(x_a,s_0-s)\psi_n^+(x_b,s_0)|0,\beta><0,\beta|\psi_n(x_b,t_0+t)\psi_m^+(x_a,t_0)|0,\alpha> \ . \tag{15}$$

By completeness, the sum over final bath states $|\beta>$ gives the identity and can be removed. In going from (12) to (15), the velocities $v_{ma\alpha}$ and $v_{nb\beta}$ have been replaced by v_i and v_f, respectively, since these are quantities that depend only on the initial and final electron states (v_i is a function of a, m and E_i). We rename $s_o - t_o \equiv \tau$ and make $t_o=0$ without loss of generality. The final result is

$$T_{nm,ba}(E_f,E_i) = \frac{v_f v_i}{2\pi\hbar} \sum_\alpha e^{-\beta e_\alpha} \sum_\beta \int \int_{-\infty}^{\infty} d\tau\, ds\, dt\ e^{i[(E_f-E_i)\tau+E_f t-E_i s]/\hbar}\, \theta(t)\, \theta(s)$$
$$<\psi_m(x_a,\tau-s)\psi_n^+(x_b,\tau)\psi_n(x_b,t)\psi_m^+(x_a,0)> \ , \tag{16}$$

where the expectation value of an operator A is taken as the equilibrium thermal average over the phonon states in the absence of the electron:

$$\langle A\rangle = Z^{-1} \sum_{\alpha} e^{-\beta\varepsilon_\alpha} \langle 0,\alpha|A|0,\alpha\rangle \, . \tag{17}$$

Equation (16) closely resembles the expression obtained by Wingreen *et al.* (1988) in their analysis of the role of phonons in resonant tunneling. Actually, a straightforward adaptation of (16) to a tight-binding formulation would exactly yield the formula given by Wingreen *et al.* (1988) in the particular case where phonons are located on one site and only one channel on each of the two "leads" is available for propagation (one orbital per chain site). We have thus generalized the equation for $T(E_f,E_i)$ derived by Wingreen *et al.* (1988) to an arbitrary mesoscopic structure with many leads and many channels on each lead and in which phonons are distributed in an extended finite region. Equation (16) can also be derived through an appropriate analysis of the reduced density matrix of the electron (Sols, 1990).

We note that the evaluation of (16) requires full knowledge of the dependence of a two particle Green's function (with four field operators) on its three significant time variables (a fourth one is meaningless due to time-translational invariance). In particular, the dependence on τ (former $s_o - t_o$) is directly linked to the irreducibility of the expectation value in (16), which, as we argue below, stems from the possibility of real excitation of phonons. In order to see this, one notices that, if one introduces the approximation

$$\langle\psi_m(x_a,\tau\text{-}s)\psi_n^+(x_b,\tau)\psi_n(x_b,t)\psi_m^+(x_a,0)\rangle \approx \langle\psi_m(x_a,\tau\text{-}s)\psi_n^+(x_b,\tau)\rangle\langle\psi_n(x_b,t)\psi_m^+(x_a,0)\rangle \, , \tag{18}$$

the dependence on τ is lost because of time translational invariance and the integration over τ yields $2\pi\hbar\delta(E_f - E_i)$, which corresponds to elastic scattering. However, the expression resulting from introducing (18) in (16) is different from that which would be obtained in the total absence of phonons. The reason is that the approximation (18) amounts to neglecting inelastic scattering while including the dressing by phonons of the elastic peak. In an extended solid, the same effect is known as polaron shift or mass renormalization. The connection between real phonon excitation (as opposed to virtual phonon dressing) and the irreducibility of the Green's function in (16) will be made clear in the perturbative analysis in Appendix A.

We conclude this section by noting that a similar analysis for the reflection probability yields the following result:

$$R_{nm,aa}(E_f,E_i) = T_{nm,aa}(E_f,E_i) + \delta_{nm}\delta(E_f\text{-}E_i)[1+2\hbar v_i \, \mathrm{Im}\, G_{nn}^{(+)}(x_a,x_a;E_i)] \, , \tag{19}$$

where the first term on the left is to be taken as short-hand notation for an expression that is formally equivalent to (16) with $m=n$. The second term on the left of (19) is a correction to the full propagation contained in the first term that removes the effect of direct propagation from $|\chi_m,x_a\rangle$ to $|\chi_n,x_a\rangle$ without reflection on the sample, much as in (8a). The dressed one-particle Green's function

$$G_{nn}^{(+)}(x_a,x_a;E_i) = \int_{-\infty}^{\infty} \frac{dt}{\hbar} \, e^{iE_i t/\hbar} \, [-i\,\theta(t) \langle\psi_n(x_a,t)\psi_n^+(x_a,0)\rangle] \tag{20}$$

stems from the interference between full and direct propagation (corresponding to the l.h.s. and first term of r.h.s. in (8a), respectively) that results from computing a probability.

Equations (16) and (19) are formally exact. Unfortunately, an exact evaluation of the one- and two-particle Green's functions that fully includes the effect of boundaries, impurities, and phonons is not possible in general. One has to resort to various types of approximations. One possibility is to treat the phonons exactly and include the elastic scattering approximately, but this is only possible in tight-binding formulations where the phonons only couple to one electron site (Mahan, 1983; Wingreen *et al.*, 1988). It is more common to assume that the electron motion in the presence of boundaries and impurities is known exactly and to include the effects of phonons approximately. We follow here the second approach because there is a wide range of situations where the one-electron problem can be solved exactly by analytical or numerical methods. We wish to develop a diagrammatic perturbation theory in the electron-phonon coupling that allows us to include the effect of phonons in a systematic fashion and eventually to introduce correlative approximations by summing a given class of diagrams. We will focus on (16) for the transmission probability because the two-particle Green's function requires some non-standard manipulation before field-theoretical techniques can be readily applied. Such a perturbative analysis is presented in Appendix A, where the Feynman rules for a diagrammatic representation of the transmission probability are derived.

In this section we follow a simpler (but more limited) approach and calculate the probability of phonon absorbtion or emmision by means of the Fermi golden rule in the Born approximation (one-phonon exchange). We consider the case of a free particle in one dimension interacting with a set of localized phonons. We assume that the coupling to phonons is of the rather general form

$$V = \lambda \sum_q \int dx \, M_q(x) \, \psi^+(x)\psi(x) \, (a_q + a^+_{-q}) \, , \tag{21}$$

where the only restriction is that $M_q(x)$ is nonzero in a finite region of space. The index q labels the type of phonon mode. The time-reversed of phonon mode q is $-q$, and both are identical, if the phonons are localized. The coupling constant λ is introduced for convenience.

The transition rate for a plane wave due to the presence of phonons is given in the Born approximation by

$$\tau^{-1}_{k_i\to k_f} = \frac{2\pi}{\hbar} \sum_{\alpha} \frac{e^{-\beta E\alpha}}{Z} \sum_{\beta} |\langle k_f,\beta|V|k_i,\alpha\rangle|^2 \, \delta(E_f - E_i + \varepsilon_\alpha - \varepsilon_\beta) \, , \tag{22}$$

where $\langle x|k\rangle = L^{-1/2}\exp(ikx)$. After some algebra, (22) becomes

$$\tau^{-1}_{k_i\to k_f} = \frac{2\pi}{\hbar L^2} \sum_q \iint dx\, dx' \, M_q^*(x) \, M_q(x') \, e^{i(k_i - k_f)(x-x')} \, [(N_q+1)\delta(E_f - E_i + \hbar\omega_q) + N_q\delta(E_f - E_i - \hbar\omega_q)] \, . \tag{23}$$

If we divide (23) by the flux of incoming particles, v_i/L, we obtain the probability that the particle is scattered from k_i to k_f or, treating k as a continuous variable, the probability of being scattered into the interval $(k_f, k_f + dk_f)$, where $dk_f \cong 2\pi/L$. Since we are interested in the probability $T(E_f,E_i)dE_f$ of scattering into $(E_f, E_f + dE_f)$, we multiply the former probability by $(dk_f/dE_f)/(2\pi/L) = L/2\pi\hbar v_f$ and obtain

$$T(E_f,E_i) = \frac{\lambda^2}{h^2 v_i v_f} \iint dx\, dx' \sum_q M_q(x) M_q(x') e^{i(k_i - k_f)(x-x')}$$
$$[(N_q+1)\delta(E_f - E_i + h\omega_q) + N_\alpha \delta(E_f - E_i - h\omega_q)] \,. \qquad (24)$$

The first term in square brackets corresponds to the phonon emission, while the second is due to phonon absorbtion.

Inelastic scattering in tight-binding chains. Phonons localized on a stub

The formalism developed in the last sections can be easily adapted to tight-binding chains, where the space is discrete instead of continuous. We provide below a list of the most important changes (a is the lattice spacing):

$$|x\rangle \to a^{-1/2} |l\rangle, \qquad \psi(x) \to a^{-1/2} c_l,$$
$$\int dx \to a \sum_l \qquad G(x,x';E) \to a^{-1}\, G(l,l';E)$$
$$v \to -(2\Delta a/h)\sin\theta \qquad kx \to \theta l \quad . \qquad (25)$$

As an illustration, one can easily see that (A20) of the appendix becomes

$$T(E_f,E_i) = 4\Delta^2\lambda^2 \sin\theta_i \sin\theta_f \sum_{(l,l')} \sum_q M_q^*(l)\, M_q(l)$$
$$[(N_q+1)\delta(E_f - E_i + h\omega_q) + N_q \delta(E_f - E_i - h\omega_q)]$$
$$G_0^{(-)}(l_a,l';E_i)\, G_0^{(-)}(l',l_a;E_f) G_0^{(+)}(l_b,l;E_f) G_0^{(+)}(l,l_a;E_i) \,. \qquad (26)$$

We can apply our perturbative method to the case where a collection of phonon modes is localized in the stub of the tight-binding structure presented in the first section. If we want to apply (26), we have to postulate (and motivate) a form of the electron-phonon interaction and then provide expressions for the Green's functions that are needed.

For the electron-phonon coupling we choose a deformation potential adapted to the case of a finite chain. This model will be more realistic in the case of long chains. For short chains, it can be taken as a qualitative *ansatz*. By applying the usual assumptions of the deformation potential (Kittel, 1983) to a chain of N atoms, we obtain

$$V_{e\text{-}ph} = \sum_{q>0} [M_q^{(c)}(l)\,(a_{qc} + a_{qc}^+) + M_q^{(s)}(l)\,(a_{qs} + a_{qs}^+)] \qquad (27a)$$

$$M_q^{(c)} = -\,C_1 \left(\frac{hq^2}{2MN\omega_q}\right)^{1/2} \sin(ql a) = -\gamma_q \sin(ql a) \qquad (27b)$$

$$M_q^{(s)} = \gamma_q \cos(ql a) \qquad (27c)$$

$$\omega_q = \left[\frac{2B_1}{M}(1-\cos qa)\right]^{1/2} = \left(\frac{B_1 a^2}{M}\right)^{1/2} q = cq \qquad (27d)$$

where M is the ion mass, B_1 is the force constant, $C_1 = 4\,|\Delta l = 2h^2/2ma^2$ is derivative of the bottom of the band with respect to the relative length change ($L=Na$).

Regarding the electron Green's functions, we can choose in this problem $l_a = l_b = 0$, since, on both sides of l=0, the propagation is free. Thus we only need Green's functions of the form $G(l,0;E)$, where $l = 1', 2', \ldots, N'$ labels the sites of the stub. It is a relatively straightforward exercise in tight-binding techniques (Guinea and Verges, 1987; Sols *et al.*, 1989a) to show that

$$G_0^{(+)}(l,0;E) = G_0^{(+)}(0,l;E) = [G_0^{(-)}(l,0;E)]^* = [G_0^{(-)}(0,l;E)]^* = \frac{i\,\psi(l)}{2\Delta \sin\theta} \qquad (28)$$

where

$$\psi_E(l) = \frac{\sin[(N-l+1)\theta]}{\sin[(N+1)\theta]} \left[1 - i\,\frac{\sin(N\theta)}{2\sin\theta \sin[(N+1)\theta]}\right]^{-1} \qquad (29)$$

is the wave function in the stub of the scattering state whose incident wave is $e^{il\theta}$ from the left or $e^{-il\theta}$ from the right, and which can be obtained by applying the relation $\psi^{(+)} = (1+G^{(+)}V)\Phi$ and projecting it on the stub sites, where Φ is a plane wave and V connects the stub and the main chain.

The final result is

$$T(E_f,E_i) = \frac{\lambda^2}{4\Delta^2 \sin\theta_i \sin\theta_f} \sum_{(l,l')} \sum_q M_q^*(l)\, M_q(l)\, [(N_q+1)\delta(E_f - E_i + h\omega_q)$$
$$+ N_q \delta(E_f - E_i - h\omega_q)] \psi_{E_i}^*(l')\, \psi_{E_f}^*(l')\, \psi_{E_f}(l)\, \psi_{E_i}(l) \,. \qquad (30)$$

At first sight, one may be surprised to see $\psi_{Ef}(l')\psi_{Ei}(l)$ instead of $\psi^*_{Ef}(l')\psi^*_{Ei}(l)$, as one would expect from squaring matrix elements. The reason for this apparent anomaly becomes clear if one tries to reproduce (30) by using the Fermi golden rule. Since in (22) we are doing perturbation theory on states that are not plane waves but scattering states of the Hamiltonian H_0, we have to calculate the matrix elements

$$\langle \psi_{k_f}^{(-)}, \beta |\, V_{e\text{-}ph} |\, \psi_{k_i}^{(+)}, \alpha \rangle$$

where (Taylor, 1972)

$$|\psi_k^{(\pm)}\rangle = (1 + G^{(\pm)}V)\,|\phi_k\rangle \,. \qquad (31)$$

This is the so-called distorted wave Born approximation (DWBA). If $\langle l|\phi_k\rangle = \exp(ikla)$ is a plane wave, and V connects stub and main chain, then $|\psi_k^{(+)}\rangle$ describes a scattering state formed by a plane wave coming from the left (if $k>0$) that is reflected or transmitted at the stub with certain probability and whose value within the stub is given in (29). The ket $|\psi_k^{(-)}\rangle$ corresponds to a coherent combination of incoming waves from the left and the right that result in an emerging wave moving to the right (if $k>0$). Since, within the stub, $\psi_k^{(+)}(l) = \psi_{-k}^{(+)}(l)$, and, in general, $\psi_k^{(+)}(l) = [\psi_{-k}^{(-)}(l)]^*$, we obtain $[\psi_{-k}^{(-)}(l)]^* = \psi_k^{(+)}(l)$. As a consequence, in a matrix element of the type $\langle \psi_{kf}^{(-)}| \ldots |\psi_{ki}^{(+)}\rangle$ that only has contributions from the stub region, we must expect to have terms of the type $\psi_{kf}(l)\,\psi_{ki}(l)$.

By using (22) with the modification described above and multiplying the result by $L^2/2\pi\hbar v_i v_f$, as in the previous section, we exactly reproduce (30). We remark that the expression (30) has been obtained first as a simple and unambiguous application of our perturbative method. This is in contrast to the use of the Fermi golden rule, which in the case of DWBA is more based on analogy with the plane-wave case and on a somewhat vague physical intuition, rather than on firm and clear theoretical arguments.

NUMERICAL APPROACH BY PATH INTEGRATION

The path integral approach as introduced by Feynman and Hibbs (1965) is an alternative to the techniques described before. The propagator can be calculated in this formalism from the knowledge of the classical Lagrangian alone and integration (and summation) of this Lagrangian over all possible path.

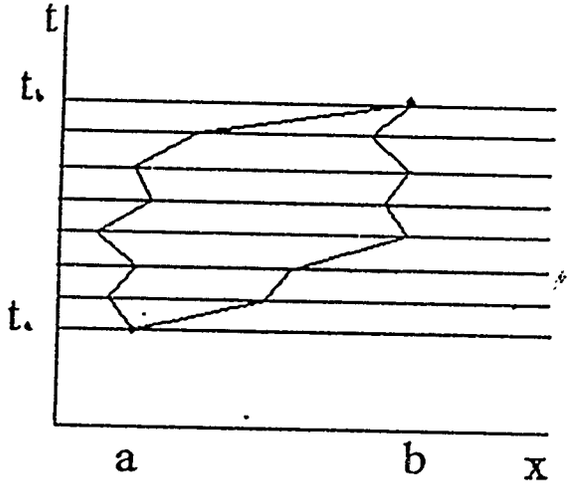


Fig. 2. Schematic for path integration.

Figure 2 shows two possible paths to propagate from a to b in the time period $[t_a,t_b]$. The assumption of discrete time steps is for our numerical approach. Feynman showed that each path $x(t)$ is connected to an amplitude

$$\phi(x(t)) \sim \exp(\tfrac{i}{\hbar} s(x(t))$$

where

$$s(x(t)) = \int_{t_b}^{t_a} L(x',x,t)dt$$

and L is the lagrangian function for the path $x(t)$. The sum of this amplitude over all path is then resulting in the propagator

$$G(b,a,t_a-t_b) \sim \sum_{\text{all path}} \phi(x(t))$$

This method of calculation has been proven to be superior for cases of strong electron-phonon coupling. The coupling results in a so-called "influence functional" which accompanies $\phi[x(t)]$ and which is well known in its functional form for many practical cases (Feynman and Vernon, 1963). If one performs the integration over all paths numerically, the complicated form of the influence functional does not play any significant role and the computation of G is possible without using perturbation theory. Indeed, such a complete approach is possible and has been demonstrated by Mason and Hess (1989). Figure 3 shows the numerical result in reduced units for the diagonal of the density matrix (the inset of the figure shows the propagator without electron phonon interaction).

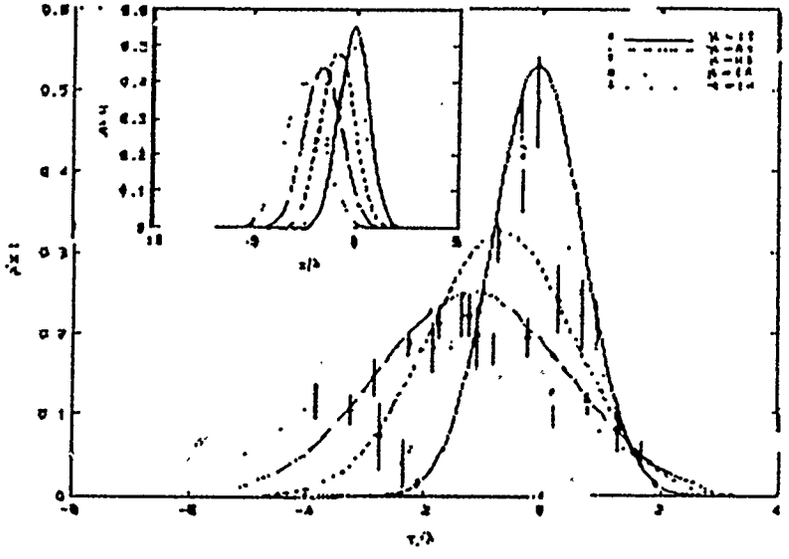


Fig. 3. Density matrix $\rho(x,t)$ as a function of x for five different times. The points and error bars are the full path-integral results. The inset shows the propagation of the density matrix without interaction with phonons. All quantities are normalized with respect to the length scales λ and reciprocal time scales γ. [After Mason and Hess (1989).]

Around one hundred million integrations are necessary to achieve this result for a one dimensional problem. Extension to a T-shape is a major challenge for even the largest existing computers. The work of Mason and Hess (1989) therefore demonstrated that in principle such quantum electron dynamics problems can be computed numerically using the path integration techniques. However, for complicated geometries and more than one dimension, current computational resources are insufficient. This work also demonstrated, however, that the use of the numerical path integration is, as least conceptually, simple. Intuitive concepts can be applied more easily (for example for boundary conditions) than in the case of the operator and Green's function formalism. Furthermore the work showed that a different numerical "space" is explored with different properties with respect to error propagation and also with respect to computer architecture, parallelization, and vectorization. We therefore proceed to explore this

avenue from different viewpoints. It is clear that some approximations are necessary to achieve numerical solutions for complex structures such as mesoscopic systems. One of the authors (V.P.) proposed to use the conjecture of Schulman (1981) that summation over classical paths may be sufficient in many geometries with hard-wall boundaries. One then has to determine only all classical ray trajectories and sum certain amplitudes to arrive at the propagator. From a numerical viewpoint this method is attractive since ray-tracing algorithms are well known from computer visualization. It turns out, as described below, that a numerical approach of this type leads to the exact solution of many interesting cases and even permits the inclusion of phonon scattering, electric and magnetic fields and so on. We refer to this method, introduced by Pevzner, as Quantum Ray Tracing.

Quantum Ray Tracing

The method we are proposing is based on the aforementioned conjecture which is analogous to the method of images used in electrostatics. This conjecture asserts that the propagator for a particle moving in a geometry formed by perfectly reflecting walls can be calculated by summing over the contributions from all the "classical" trajectories (that is, trajectories from all the "images" of the starting point to the final point).

(1) Method of Images

The method of images is well known in electrostatics (Morse and Feshbach, 1953). By utilizing the symmetry of the problem one can replace the boundary with a distribution of images such that the boundary condition for a given differential equation is automatically satisfied. Formally, one can proceed as follows.

Consider a field ψ that is restricted to some geometry formed by the boundary σ. Let us assume that this boundary is a Cauchy polygon. In addition, suppose the field ψ obeys the following differential equation

$$\mathcal{L}\psi(x)=0$$

with an inhomogeneous boundary condition (of Dirichlet or Neumann or mixed type). Instead of solving for the field directly, one typically proceeds by solving an inhomogeneous differential equation

$$\mathcal{L}G_1(x,x')=\delta(x-x')$$

with a homogeneous boundary condition which is isomorphic to $\mathcal{L}\psi(x)=0$.

First we construct an operator, T°, which generates all the images. Suppose the restricted region is an equilateral triangle. Then we can define a set of image generators $\{I,A,B,C\}$ as shown in Fig. 4. The set of these generators combined with defining relations completely specifies all elements of the group. Since every finite group is isomorphic to a permutation group, images (words) can be simply generated by appropriate permutations in a representation of the corresponding permutation group (Grossman and Magnus, 1964). Thus the mapping operator T° is

$$T^\circ = I^\circ + \sum_{j=1}^{\infty} \beta^j \, T_j^\circ \qquad (32)$$

where $\beta = 0$ or -1 depending on the type of boundary condition (-1 for Dirichlet and 0 for Neumann) and

$$T_j^\circ = P_j\{\alpha\} \prod_{i=1}^{j} A_{\alpha_i}^\circ,$$

where $A_{\alpha_i}^\circ$ = {set of all generators} and $P_j\{\alpha\}$ is a permutation operator that generates all the distinct non-trivial words (images).

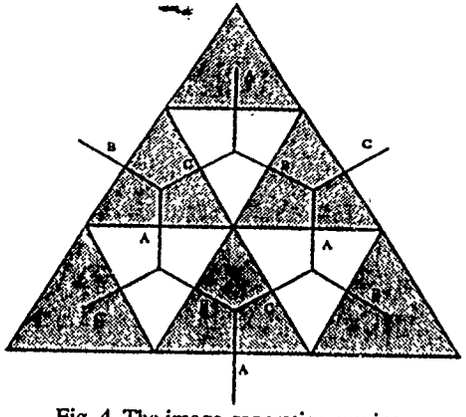


Fig. 4 The image generation routine.

After obtaining the solution to $\mathcal{L}G_\lambda(x,x') = \delta(x-x')$ for free space, $G_{\lambda 0}(x,x')$, we can get the solution in the constrained geometry by the use of T°. If x' is the position of the source and x is the point of observation, then the full solution is

$$\begin{aligned} G_\lambda(x,x') &= G_{\lambda 0}(x,x') + \sum_{j=1}^{\infty} \beta^j \, T_j^\circ \, G_{\lambda 0}(x,x') \\ &= G_{\lambda 0}(x,x') + \sum_{j=1}^{\infty} \beta^j \, G_{\lambda 0}(x, T_j^\circ x') \\ &= G_{\lambda 0}(x,x') + \sum_{j=1}^{\infty} \beta^j \, P_j\{\alpha\} \, G_{\lambda 0}(x, \Pi_{i=1}^{j} A_{\alpha_i}^\circ x') \\ &= G_{\lambda 0}(x,x') + \sum_{j=1}^{\infty} \beta^j \, \Sigma_\alpha P_j\{\alpha\} G_{\lambda 0}(x,x_j') \end{aligned} \qquad (33)$$

where

$$x'_j = \Pi_{i=1}^{j} A_{\alpha_i}^\circ x'$$

is the position of an image corresponding to a word of order j. For instance, for an equilateral triangle with the source in the center, images fall on the hexagonal lattice. [Each corner, due to the six-fold symmetry, is represented by the D_3 group with a set of defining relations, say $(AB)^3 = A^2 = B^2 = I$]. As the geometry becomes more complicated, so does the network of images (since the set of group generators increases), then the calculation of T° becomes a considerable permutational task. An alternative approach for generation of images is ray

tracing. This approach offers both a clear physical insight as well as an efficient numerical algorithm. The method of ray tracing is equivalent to the method of images; the symmetry of geometry is utilized in the same way in both cases. By folding the path from images to an observation point, one easily sees the one-to-one correspondence between rays and images. Suppose an operator, $R\circ$, performs such folding. Then all the trajectories are generated from the direct path, D_0, as in $D_j = R\circ T_j\circ D_0$ where $T_j\circ$ first generates paths from all images of class j and $R\circ$ folds these paths into the constrained region thus forming the ray trajectory.

(2) Quantum Ray Tracing in 2-d Pipe

Let us demonstrate the above discussion with a simple example. Consider a particle confined to a 2-d pipe with infinitely hard walls as shown in Fig. 4. We can express its propagator as

$$= \sum_k \langle x \mid k\rangle \langle k \mid x'\rangle \; e^{\frac{-iE_k t}{h}} \sum_n \langle y \mid n\rangle \langle n \mid y'\rangle \; e^{\frac{-iE_n t}{h}} \qquad (34)$$

where $\langle x|n\rangle^* = u_n^*(y)$ is the square well wave function and

$$E_n = \left(\frac{h^2}{8ma^2}\right) n^2 = E_0 n^2$$

is its energy, while $\langle k|y'\rangle$ is a plane wave and

$$E_k = \frac{h^2k^2}{2m}$$

is its energy.

Since the coordinates are separable the 2-d propagator is a product of two 1-d propagators,

$$K(x,t;x',0) = \left(\int_x^{x'} D_x(t)\, e^{\frac{iS[x(t)]}{h}}\right)\left(\int_y^{y'} D_y(t)\, e^{\frac{iS[y(t)]}{h}}\right) = K(x,t,x',0)\; K(y,t;y',0)\;. \qquad (35)$$

The propagator for the unconstrained degree of freedom is trivial to obtain,

$$K(x,t,x',0) = \left(\frac{M}{2\pi i h t}\right)^{\frac{1}{2}} \exp\left\{\frac{im(x-x')^2}{2ht}\right\}. \qquad (36)$$

However, the calculation of $K(y,t,y',0)$ directly from a path integral formulation is not possible unless one resorts to the method of images (or ray tracing) as discussed above. To illustrate this point, let us first show how the eigenfunction expansion can be put into a form that is clearly identical to the ray tracing solution,

$$\begin{aligned} K(y,t,y',0) &= \left(\frac{2}{a}\right)\sum_{n=0}^{\infty} \sin\left(\frac{\pi R y}{a}\right)\sin\left(\frac{\pi R y'}{a}\right) e^{\left(\frac{-iE_0 n^2}{h}\, t\right)} \\ &= \left(\frac{1}{a}\right)\sum_{n=0}^{\infty}\left(e^{\frac{+i\pi n}{a}(y-y')}\right) - \left(e^{\frac{+i\pi n}{a}(y+y')}\right) e^{\left(\frac{-iE_0 n^2}{h}\, t\right)} \\ &= \left(\frac{1}{a}\right)\left(\theta_3\left(\frac{\pi(y-y')}{2a}, \frac{-E_0 t}{\pi h}\right) + \theta_3\left(\frac{\pi(y+y')}{2a}, \frac{-E_0 t}{\pi h}\right)\right). \end{aligned} \qquad (37)$$

where $\theta_3(a,b)$ is Jacobi theta function (Abramowitz and Stegun, 1964). Remember that the Poisson summation formula (Sols *et al.*, 1989a) is

$$\sum_{n=-\infty}^{\infty} f(\alpha n) = \left(\sqrt{2\pi/\alpha^2}\right)\sum_{m=-\infty}^{\infty} F(2m\pi/\alpha) \qquad (38)$$

where $F(k) = \int (2\pi)^{-1/2} e^{ikx} f(x)$ and is evaluated by the usual calculus of residues. Using this formula, we can establish the following property of the function

$$\theta_3(a,b) = (-ib)^{\frac{-1}{2}}\, e^{\frac{a^2}{ib}}\, \theta_3\left(\frac{a}{b}; \frac{-1}{b}\right).$$

Using this property the propagator can be cast in the form (Schulman, 1981):

$$\begin{aligned} K(y,t,y'0) &= \left(\frac{m}{2\pi h t}\right)^{\frac{1}{2}} \sum_{n=-\infty}^{\infty}\left[e^{\frac{im}{2ht}(y-y'+2na)^2} - e^{\frac{im}{2ht}(y+y'+2na)^2}\right] \\ &= \left(\frac{m}{2\pi h t}\right)^{\frac{1}{2}} \sum_{n=-\infty}^{\infty}\left[e^{\frac{im}{2ht}(y-y'_n)^2} - e^{\frac{im}{2ht}(y+y'_n)^2}\right]. \end{aligned} \qquad (39)$$

The first of these terms corresponds to the propagator associated with the trajectory having an even number of reflections, while the second one is associated with an odd number of reflections. This is, of course, equivalent to the solution that one would obtain from the method of quantum ray tracing (images) directly. The quantum solution for a 2-d pipe in the language of quantum ray tracing would be

$$K(y,t,y',0) = \sum_{j=0}^{\infty} \Sigma_\alpha P_j\{\alpha\} K(y,t,\prod_{i=0}^{j} R\circ A_{\alpha i}\circ y',0) \qquad (40)$$

where we have used (33). This geometry corresponds to the "city-streets group" (Grossman and Magnus, 1964), $(C_\infty)^2$, and therefore has infinitely many elements. (Note that there are only two non-trivial generators here, one for the reflections about each boundary.)

This example is a clear manifestation of the Feynman picture of quantum mechanics (Economou, 1983) — the set of all the possible paths, each having an amplitude $e^{iS/h}$, has been reduced to the set of all the possible *classical* paths. This is a consequence of multi-minima present in the action S. To find these minima we vary $x = x_{cl} + \eta$,

$$S[x] = S_{cl}[x] + \left(\frac{\delta S}{\delta x}\right)_{cl}\eta + \left(\frac{\delta^2 S}{\delta x^2}\right)_{cl}\eta^2 + O(\eta^3) \qquad (41)$$

By solving the Euler-Lagrange equation, $\left(\frac{\delta S}{\delta x}\right)_{cl}=0$, we should find all the classical paths, $\{x^{\beta}_{cl}\}$ (provided that they exist) and the action is then

$$S[x] = S_{cl}[x] + \left(\frac{\delta^2 S}{\delta x^2}\right)_{cl}\eta^2 . \tag{42}$$

Then

$$K(x_f,t_f,x_i,0) = \int_i^f D\,x(t)\,\exp\,iS[x(t)]/\hbar$$

$$= \sum_{\beta}\int_0^0 D\eta_{\beta}(t)\,\exp\frac{i}{\hbar}\left(S^{\beta}_{cl} + \left(\frac{\delta^2 S^{\beta}}{\delta x^2}\right)_{cl}\eta^2{}_{\beta}\right) \tag{43}$$

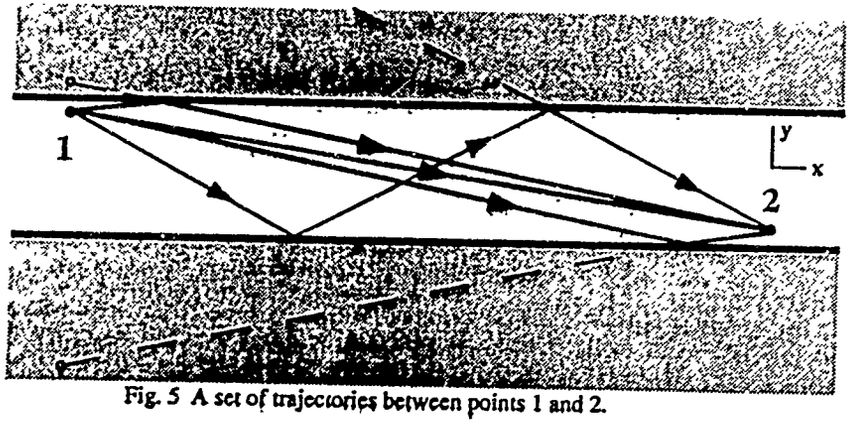


Fig. 5 A set of trajectories between points 1 and 2.

The Jacobian is unity since the transformation is linear, and the propagator is

$$K(x_f,t,x_i,0) = \sum_{\beta}\left(\det\left(\frac{\delta^2 S}{\delta x^2}\right)_{cl}\right)^{d/2}\exp\,iS^{\beta}_{cl}/\hbar . \tag{44}$$

Since the number of minima $\{x_{cl}{}^{\beta}\}$ in the 2-d pipe problem is infinite, so is the total number of terms in (39) and (40).

The derivation of (44) relies on the assumptions that the end points do not coincide with the classical caustic (otherwise the $\left(\det\left(\frac{\delta^2 S}{\delta x^2}\right)_{cl}\right)$ is infinite), and that there are no 'flawed' classical trajectories (Schulman and Ziolkovski, 1990) for which $\left(\frac{\delta S}{\delta x}\right)\neq 0$.

(3) Problems in Quantum Ray Tracing

This brings us to the issue of "shadow" regions in the context of ray tracing, and prompts us to review a problem of the nineteenth century – the diffraction of light by the wedge. The pioneering work of Sommerfield, Riemann, Carslaw, and Macdonald has recently been given renewed attention in the context of Feynman path integrals by DeWitt-Morette *et al.* (1986) and in the context of stochastic diffusions by Molchanov (1990). Although we encourage readers to follow the detailed discussion on this problem, we simply quote the results of these investigations. The geometry and coordinates are shown in Fig. 6.

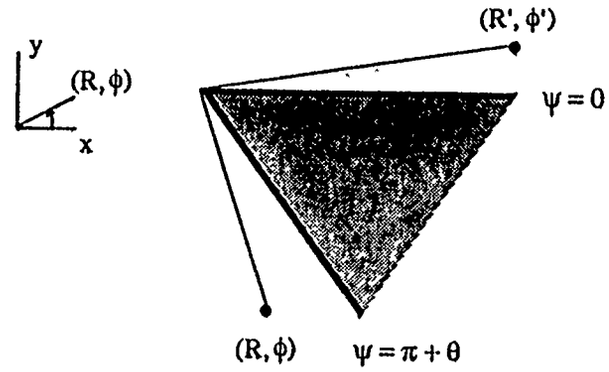


Fig. 6 Geometry for edge scattering.

The propagator for the wedge with Dirichlet boundary condition is (DeWitt-Morette *et al.*, 1986)

$$K^{-}_{w}(\phi,R,t,\phi',R') = \left(\frac{m}{2\pi i\hbar t}\right)^{d/2}\exp\left(\frac{im(R^2+R'^2)}{2\hbar t}\right)\sum_{j=1}^{2}F(z,\phi_j) \tag{45}$$

where

$$F(z,\phi_j) = \sum_{\alpha}^{\nu}\exp(-iz\cos\phi_j)\left(-i\sum_{\gamma=1}^{\mu-1}\cos(\phi_j(\mu-\gamma)/\mu)\,I(z,\phi_j,\gamma)\right) \tag{49}$$

$$I(z,f_j,\gamma) = \int_0^z\left(\left(e^{i\pi\gamma/2\mu}J_{-\gamma/\mu}(u) - e^{-i\pi\gamma/2\mu}J_{\gamma/\mu}(u)\right)e^{iu\cos f_j}\right)du \tag{50}$$

and $z=\frac{mRR'}{\hbar t}$, $\phi_1=\phi-\phi'+\frac{2\pi\alpha\mu}{\nu}$, $\phi_2=\phi+\phi'+\frac{2\pi\alpha\mu}{\nu}$, and integers μ and ν are defined by the wedge angle $\theta=\frac{\pi\mu}{\nu}$. This expression is cumbersome; in our numerical calculations we will use its asymptotic form whenever appropriate. The propagator in a general polygonal geometry can be approximated by the contributions from ordinary quantum ray tracing solutions as discussed earlier and from "flawed" classical trajectories which are calculated by using the asymptotic expansion of (45). The final result must involve the sum of K^{-} from all the images of x' to the observation point x.

The example of the pipe clearly demonstrates the use of ray tracing. However, due to the particular symmetry of this geometry, we could have solved for the propagator by summing over eigenfunctions. In more complicated geometries the use of eigenfunctions becomes prohibitive, while the ray tracing, at least from an algorithmic point of view, presents no difficulties for arbitrarily complex geometries. For the problem of general interest one would have to introduce a cutoff on the total number of reflections for any given ray much like the upper eigenvalue cutoff in the eigenfunction sum. The cutoff is of no physical consequence provided that the characteristic length scale associated with the cutoff is smaller then the finest geometric feature of geometry in the problem.

(4) Phonons and Quantum Ray Tracing

The full utility of this method becomes evident when one attempts to include the effects of dissipation on quantum transport in mesoscopic systems. We illustrate this point by making a formal connection between ray tracing and the conventional perturbation theory. The perturbative expansion obtained offers both a clear physical insight and permits generalization of the treatment of phonons in the restricted geometries. The typical Hamiltonian of electron interactions with the environment is of the form

$$H = \frac{p^2}{2m} + V_{ext}(x) + \sum_q \hbar\omega_q a_q a_q^+ + \sum_q M_q(a_q + a_{-q}^+) \ . \qquad (47)$$

where ω_q and $M_q(x)$ are the frequency and the coupling matrix element of the nth phonon mode. $V_{ext}(x)$ is the external potential, and does not contain the boundary condition which we specify explicitly. The typical coupling is of the form

$$M_q = C_q \exp -iqx \qquad (48)$$

where C_q is a c-number. In the case of the Frohlich polaron $C_q = Cq^{1/2}$, while it is constant for optical deformation potential, and for acoustic deformation potential $C_q = C/q$.

The evolution of the reduced density matrix of an electron is (Feynman and Vernon, 1963)

$$\rho(x_f x'_f,t,\beta) = \int dx \int dx' \ J(x_f x'_f,t,x,x'0,\beta)\rho(x,x',0,\beta) \qquad (49)$$

and

$$J = \iint_{ii'}^{ff'} Dx(t)\, Dx'(s) \exp\left\{\frac{i}{\hbar}\left(S_e[x(t)] - S_e[x'(s)] + i\Phi^{\beta}[x(t);x'(s)]\right)\right\}, \qquad (50)$$

where $S_e[x(t)]$ is the electronic action associated with $L_e = \frac{m\dot{x}^2}{2} - V_{ext}(x)$ and $\Phi\beta[x,x']$ is the exponent of the influence functional due to the phonon bath and the relation of $J(f,i)$ to two particle Green's function has been discussed in the previous section. If the electron was decoupled from the bath, J would simply be $K^*(x_f',x_i',t,0)K(x_f,x_i,t,0)$. Since all the modes of the bath are independent from one another (we omit superscript β from now on), $\Phi[x,x'] = \sum_q \Phi_q[x,x']$, or

$$\Phi_q[x,x'] = \int_0^T dt \int_0^T ds \ \{ A_q(t-s)\, e^{-iq(x'(t)-x'(s))} + A_q^*(t-s)\, e^{iq(x(t)-x(s))} + A_q(t-s)\, e^{iq(x(t)-x'(s))} + A_q^*(t-s)\, e^{-iq(x'(t)-x(s))} \} \qquad (51)$$

where the correlation function is defined as

$$A_q(t) = |C_q|^2 \{ \coth\left(\frac{\beta\omega_q}{2}\right)\cos\omega_q t + i \sin\omega_q t \} \ . \qquad (52)$$

In the case of the optical deformation potential interaction or the Frohlich polaron, we can approximate ω_q independent of q. The terms of the exponent Φ in (51) are then of the form $\frac{A(t-s)}{|x(t)-x'(s)|}$ and $A(t-s)\delta(x(t) - x'(s))$, respectively with the interaction type, where $A(t) = |C|^2$ $\{ \coth\left(\frac{\beta\omega}{2}\right) \cos\omega t + i \sin\omega t \}$.

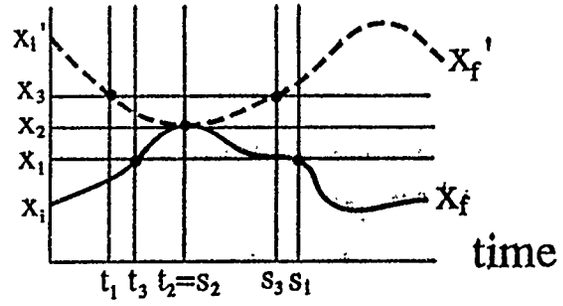


Fig. 7 An illustration of correlated paths.

The optical deformation potential establishes correlation between all shared points of various path. As shown in Fig. 7, path x' is correlated with itself at x_3 since $x'(t_3)=x'(s_3)$, while x and x' are correlated at x_2 since $x(t_2)=x'(s_2)$. The effects of the bath are felt strongest at the caustic surface (where many different paths intersect) and in the regions where paths spend a long time as shown in Fig. 8. The first of these affects the dissipative features of J while the second contributes to both dynamics and renormalization.

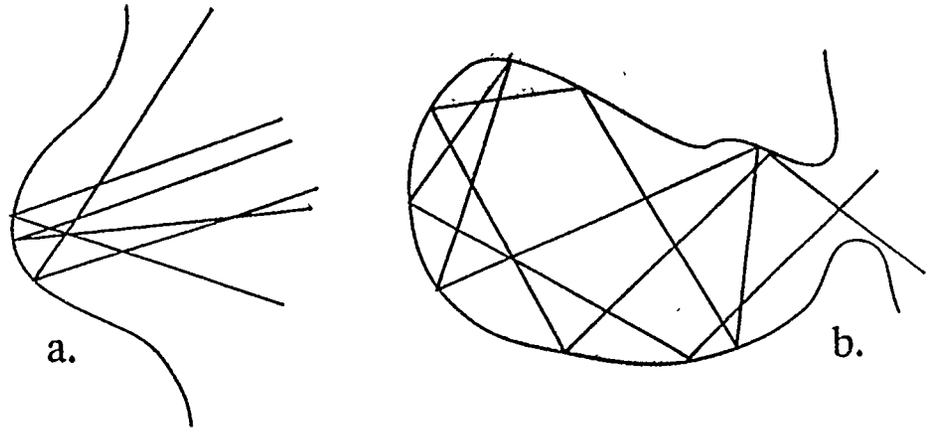
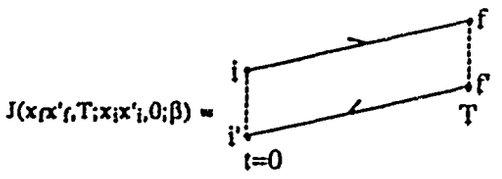
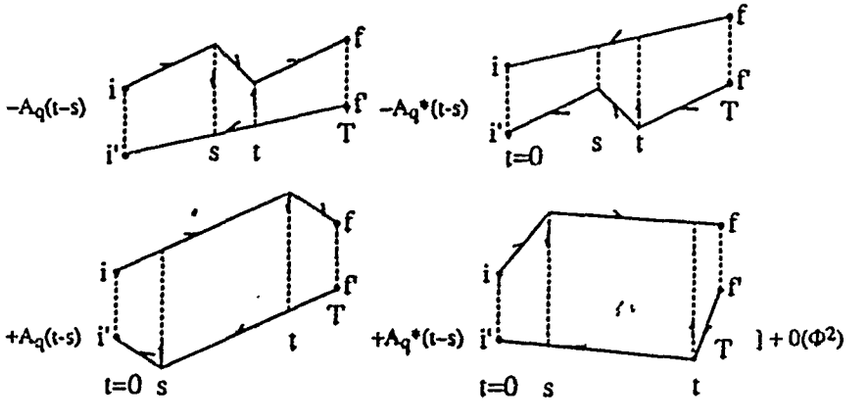


Fig. 8. Schematic representation of (a) caustic and (b) trapped trajectory.

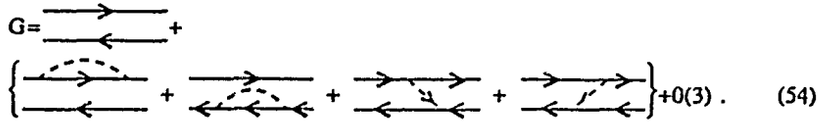
Let us return to the general case, the perturbation expansion in $\frac{\Phi}{\hbar}$ is



$$+\sum_q \iint \frac{dt\,ds}{h}\{$$



This expansion is equivalent to the conventional perturbation expansion of two-particle Green' functions which is often represented schematically as



(54)

The formal relation of $J(x_f x'_f, T, x_i x'_i, 0)$ to $G(x_f, x'_f, x_i, x'_i, t_f - t'_i, t'_f - t'_i, t_i - t'_i, 0)$ is discussed i detail in the subsequent section. This result has also been obtained by Rammer (1990) with th application of the methods of Keldysh (1964), Schwinger (1961), and Feynman and Hibb (1965).

The typical nth order "cross" term is of the form

$$\left(\frac{-1}{h}\right)^n \sum_{q_1}\sum_{q_n} \iint dt_1 ds_1 \cdots \iint dt_n ds_n$$

$$A_{q_1}^{\alpha_n}(t_n - s_n)\; A_{q_1}^{\alpha_1}(t_1 - s_1)\; \exp i \sum_{j=1}^{n} d_j q_j (x(t_j) - x'(t_j)) \qquad (55)$$

where by A^{α_n} we mean $|A|\exp i\phi_a \alpha_n$ with $\alpha_n = \{-1,1\}$ and $t_n > s_n > t_{n-1} > s_{n-1}$ for all n. The terms in (53) then separate into two single path integrals with the Lagrangian of the form

$$\mathcal{L} = \mathcal{L}_o + \int_0^T f(t)x(t)dt \qquad (56)$$

and the force is

$$f(t) = \sum_{n=1}^{N} h\alpha_n q_n \delta(t - t_n) = \sum_{n=1}^{N} f_n \delta(t - t_n) \qquad t_i < t_n < t_f . \qquad (57)$$

The propagator for this problem can easily be obtained analytically, it is of the form $(\frac{m}{2\pi i\hbar t})^{d/2} \exp iS_{cl}/\hbar$ where the classical action is

$$S_{cl} = \frac{m(x_f - x_i)^2}{2t} - \left\{ \sum_{n=1}^{N} f_n \left(\frac{(t_n - t_i)(x_f - x_n)}{(t_f - t_i)} + x_i \right) + \sum_{n=1}^{N} f_n^2 \frac{2(t_n - t_i)(t_f - t_n)}{2m(t_f - t_i)^2} \right\} q(t_f, t_i) . \qquad (58)$$

Furthermore, if we add potential $V(x) = 0$ for $x>0$ and zero elsewhere, using the group property of the the propagator

$$K(x_f, t_f, x_i, t_i) = \int dx\; K(x_f, t_i, x, t)\; K(x, t, x_i, t_i) , \qquad (59)$$

we obtain a kernal equivalent to that obtained by the method of ray tracing (images) (see Fig. 9):

$$K(x_f, x_i, T, 0) = \left(\frac{m}{2\pi\hbar t}\right)^{\frac{1}{2}} \left\{ \exp\frac{i}{\hbar}\left(\frac{m(x_f - x_i)^2}{2T} - \sum_n \mathrm{sign}(t_n)\, f_n \left(\frac{t_n}{T}(x_f - x_i) + x_i \right) - \sum \frac{f_n^2 (T - t_n)\, t_n}{2m\, T^2} \right) - \exp\frac{i}{\hbar}\left(\frac{m(x_f + x_i)^2}{2T} - \sum_n \mathrm{sign}(t_n)\, f_n \left(\frac{t_n}{T}(x_f + x_i) + x_i \right) - \sum \frac{f_n^2 (T - t_n)\, t_n}{2m\, T^2} \right) \right\} \qquad (60)$$

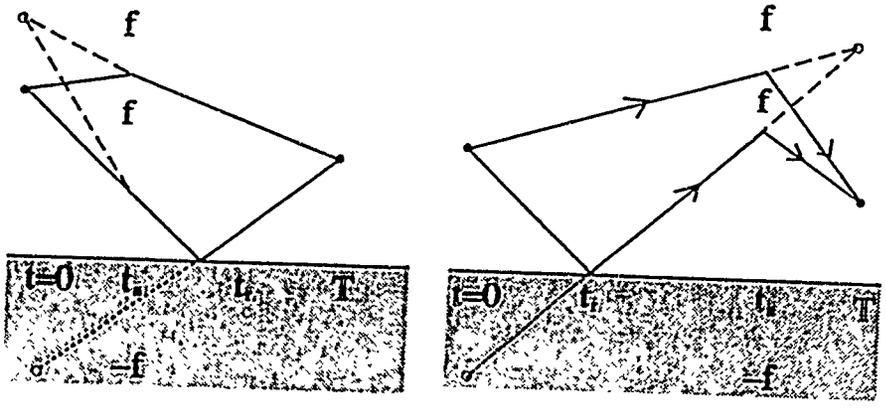
where $\mathrm{sign}(t_h) = (2\theta(t_n - t_r) - 1)$ and $t_r = Tx_i(x_f + x_i)^{-1}$ is the time at which a classical particle would reflect. In the case of the square well, paths will have more than one reflection, thus $\mathrm{sign}(t_n) = (-1)^{N_{max}+1}$ and $N_{max} = \sum_{m=1}^{N_{ref}} \int_{t_i}^{t_n} dt\, \delta(t - t_m^r)$. A perturbative approach is also applicable when electric or magnetic fields are present. To obtain a propagator for this case, we have performed calculations similar to the ones above. However, they may be omitted at the present. It is worth mentioning that a numerical path integral solution for even the simplest of propagators is a considerable task. Yet by using explicit (simple) analytical expressions, and summing numerically advantageous ways of solution can be realized.

(5) Field Effects

Weak electric and magnetic fields can easily be included in propagators in the path integral formalism (see Appendix B). In fact the propagator of a particle is completely described by its classical equations of motion. By utilizing the discussion above, we conjecture that the propagator for the particle in the presence of classical fields is equal to the sum of all the "classical" trajectories inside the restricted region . When fields are linearly coupled to the particle, the propagator is

$$G(1,2,t) = \Sigma_{\beta} \left\{ \frac{-iM}{2\pi t} \right\}^{d/2} \cdot \exp\{ iS_{cl}(\beta) \} \cdot (-1)^{N_{\beta}} , \qquad (61)$$

where ß is the classical path index, d is the dimensionality, $S_{cl}(\beta)$ and N_{β} are the action and the total number of reflections along the ß-th path, respectively.



a. b.

Fig. 9. Instanteneous momentum transfer at time t_n (a) before and (b) after classical reflection time t_r.

Applications and Conclusions

The nature of the algorithm, described in a formal manner above, allows us to achieve a high degree of parallelism in computation, hence numerical investigations are particularly well suited for machines with parallel architecture. We are currently investigating arbitrary 2-d geometries by performing the sum over a large set of classical trajectories numerically. We have restricted the number of trajectories to 4×10^3 with an upper cutoff in the total number of the reflections from the boundaries. Equivalently, we are also calculating G(1,2,E) and from it the transmission and reflections coefficients. In 2-d, for $E > L^{-2}$, the propagator is

$$G(1,2,E) = \Sigma_{\beta} \left\{ EL_{\beta}^2 \right\}^{-1/4} \cdot \exp\left\{ -iIL_{\beta}|E^{1/2} \right\} \cdot (-1)^{N_{\beta}} \qquad (62)$$

One particular example of such geometry is that of the Quantum Modulated Transistor (Sols *et al.*, 1989a, 1989b). As mentioned above, Sols *et al.* (1989a,1989b) have used tight-binding Green's functions to investigate the transport in the absence of impurities. We plan to use geometric models of impurities and to investigate their effect on quantum interference by ray tracing as shown schematically in Fig. 10.

In summary, we have shown the equivalence of the method of ray tracing with the method of images. Hence we have demonstrated that the method of ray tracing is formally correct for an arbitrary 2-d geometry. We have also proposed a conjecture, which we believe will permit us to investigate the effects of weak electric and magnetic fields.

We are comparing the results of our calculations with those obtained by other methods (i.e., tight-binding Green's function, Schrödinger's equation, and the analytic calculations available for a small class of tractable problems) for a wide range of geometries. Also, we intend to generalize our numerical algorithm to the cases where magnetic or/and electric fields are present.

The method which we are developing has three main advantages as compared to standard methods. Firstly, based on our conjecture, the path integral nature of our calculation permits an exact treatment of transport in the presence of (at least weak) electric and magnetic fields. Secondly, for the same reason, there is a straightforward prescription for the perturbative treatment of phonons. Finally, and possibly of greatest importance, the numerical algorithm of ray tracing is well suited for an efficient use on a computer and is unaffected by the complexity of the geometry.

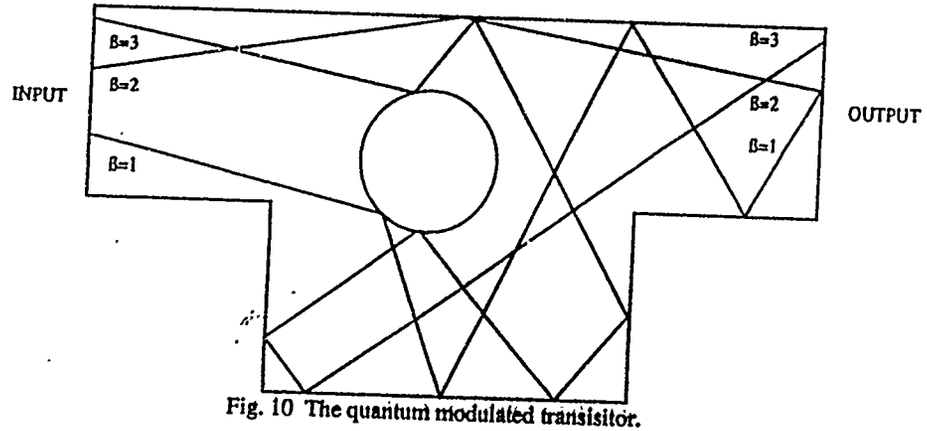


Fig. 10 The quantum modulated transisitor.

ACKNOWLEDGEMENT

This work was supported by the Office of Naval Research and the Army Research Office. We also wish to thank John Larson for his contributions to the quantum ray tracing project.

APPENDIX A

Diagrammatic Analysis of the Inelastic Transmission Probability

In this appendix we present a perturbative analysis of the effect of the electron-phonon interaction on the transmission proability in a mesoscopic structure like the one shown in Fig. 1. For the electron-phonon interaction, we take the Hamiltonian given in (11). We derive general Feynman rules for the construction of diagrams in energy space and illustrate our method by applying it to the case of a free particle in one dimension. In order to develop a diagrammatic perturbation theory in the electron-phonon interaction, we would like to write the field operators of (9) in a time-ordered form, so that Wick's theorem can be directly applied. To that end, we note that the two-particle Green's function in (9) can be written

$$G_{mnnm}(x_a,x_b,x_b,x_a;\, t_a,t_b,s_b,s_a)$$
$$\equiv \theta(t_b - t_a)\theta(s_b - s_a)\langle\psi_m(x_a,t_a)\psi_n^+(x_b,t_b)\psi_n(x_b,s_b)\psi_m^+(x_a,s_a)\rangle$$
$$= \langle T_\gamma \psi_m(x_a,t_a)\psi_n^+(x_b,t_b)\psi_n(x_b,s_b)\psi_m^+(x_a,s_a)\rangle \qquad (A1)$$

where T_γ indicates time ordering in the time contour γ shown in Fig. A1, sometimes called the Keldysh contour. To obtain (A1), we have used the fact that, by definition, the times t_a and t_b always are later in γ than s_a and s_b and that, in the absence of other electrons, (A1) vanishes if $s_b<s_a$ or $t_b<t_a$ (equivalent to $t_b>t_a$ in the contour γ), since in those cases a destruction operator is on the right or a creation operator is on the left.

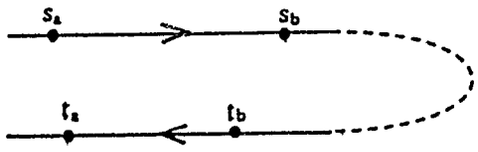


Fig. A1 Time ordering on the Keldysh path.

To obtain a more symmetric notation we will introduce some changes. First, we adopt a position representation for both the x and y coordinates:

$$\psi_m^+(x_a) = \int dy_a \chi_m(y_a)\psi^+(x_a) \qquad (A2)$$

where $x_a \equiv (x_a,y_a)$. Equation (A1) becomes

$$\iiiint dy_a\, dy'_a\, dy_b\, dy'_b\, \chi_m(y_a)\chi^*_n(y_b)\chi_n(y'_b)\chi^*_m(y'_a)$$
$$G(x_a,x_b,x_b,x_a;t_a,t_b,s_b,s_a) \qquad (A3)$$

where

$$G(x_a,x_b,x_b,x_a;t_a,t_b,s_b,s_a)$$
$$\equiv \theta(t_b - t_a)\theta(s_b - s_a)\langle\psi_m(x_a,t_a)\psi^+(x_b,t_b)\psi(x_b,s_b)\psi^+(x_a,s_a)\rangle \qquad (A4)$$

At this point we introduce the interaction picture, in which

$$\hat{A}(t) \equiv e^{i(H_0+H_B)t/\hbar}\, A\, e^{-i(H_0+H_B)t/\hbar}. \qquad (A5)$$

The operators in the Heisenberg and interaction picture are related by

$$A(t) = \hat{U}(0,t)\,\hat{A}(t)\,\hat{U}(t,0), \qquad (A6)$$

where

$$\hat{U}(t,t_0) \equiv T\exp\Big\{\frac{-i}{\hbar}\int_{t_0}^{t} dt'\ \hat{V}(t')\Big\} \qquad (A7)$$

is the evolution operator in the interaction picture. Equation (16b) can then be written

$$\langle T_\gamma \hat{U}(0,t_a)\hat{\Psi}(x_a,t_a)\hat{U}(t_a,t_b)\hat{\Psi}^+(x_b,t_b)\hat{U}(t_b,s_b)\hat{\Psi}(x_b,s_b)\hat{U}(s_b,s_a)\hat{\Psi}^+(x_a,s_a)\hat{U}(s_a,0)\rangle \qquad (A8)$$

Note that here $\hat{U}(0,t_a)$ and $\hat{U}(s_a,0)$ can be replaced by the identity since they are acting on states without electrons, but for the same reason they can be replaced by $\hat{U}(-\infty,t_a)$ and $\hat{U}(s_a,-\infty)$. On the other hand, $\hat{U}(t_b,s_b)$ is equivalent to $\hat{U}(t_b,\infty)\hat{U}(\infty,s_b)$ and (A4) can thus be written formally as

$$\langle T_\gamma \psi(x_a,t_a)U(-\infty,\infty)\psi^+(x_b,t_b)\psi(x_b,s_b)U(\infty,-\infty)\psi^+(x_a,s_a)\rangle . \qquad (A9)$$

Due to time-translational invariance the Green's function in (A1) remains invariant under a global time shift:

$$G(x_a,x_b,x_b,x_a;\tau\text{-}s,\tau,t,0) = G(x_a,x_b,x_b,x_a;t_0+\tau\text{-}s,t_0+\tau,t_0+t,t_0). \qquad (A10)$$

If, instead of $T_{nm,ba}(E_f,E_i)$, we decide to compute

$$P_{nm,ba}(E_f;E_i,E_i') \equiv T_{nm,ba}(E_f,E_i)\,\delta(E_i-E_i'), \qquad (A11)$$

the resulting expression is more symmetric in the four times variables since, due to (A10), the delta function can be replaced by

$$\delta(E_i-E_i') = \frac{1}{2\pi\hbar}\int_{-\infty}^{\infty} dt_0\, e^{i(E_i-E_i')t_0/\hbar}. \qquad (A12)$$

By introducing some obvious changes of variables, we can finally write

$$P_{nm,ba}(E_f;E_i,E_i') = \frac{\nu_f(\nu_i\,\nu_i')^{1/2}}{(2\pi\hbar)^2}\int\int\int\int_{-\infty}^{\infty} ds_a\, ds_b\, dt_b\, dt_a$$
$$e^{i[E_f't_a - E_i's_a - E_f(t_b-s_b)]}\int\int\int\int dy_a\, dy_b\, dy_b'\, dy_a'$$
$$\chi_m(y_a)\chi_n^*(y_b)\chi_n(y_b')\chi_m^*(y_a')\, G(x_a,x_b,x_b,x_a;t_a,t_b,s_b,s_a) \qquad (A13)$$

We prefer to compute (A13) instead of (A9) because when the $\hat{U}$'s in (A10) are expanded in powers of the interaction and Wick's theorem is applied to each term, the resulting contractions will depend on the time differences and, after Fourier transforming, the time integrals in (A13) will be evaluated exactly and we will then be able to formulate the Feynman rules for the calculation of (A13) directly in the energy space. We will devote the rest of this appendix to carrying out this program.

By expanding the evolution operators in (A10), we obtain

$$G(x_a,x_b,x_b,x_a;t_a,t_b,s_b,s_a) = \sum_{n=0}^{\infty}\sum_{m=0}^{\infty}\Big(\frac{i}{\hbar}\Big)^{n-m}\frac{\lambda^{n+m}}{n!\,m!}$$

$$\int_{-\infty}^{\infty} dt_1 \dots \int_{-\infty}^{\infty} dt_n \int_{-\infty}^{\infty} ds_n \dots \int_{-\infty}^{\infty} ds_1 \int dx_1' \dots dx_n' \, dx_m \dots dx_1 \sum_{p_1} \dots \sum_{p_n} \sum_{q_m} \dots \sum_{q_1}$$
$$M_{p_1}(x_1') \dots M_{p_n}(x_n') M_{q_m}(x_m) \dots M_{q_1}(x_1) \qquad \text{(A14a)}$$

$$\langle T_\gamma \hat{\Psi}(x_a,t_a)\hat{n}(x_1',t_1) \dots \hat{n}(x_n',t_n)\hat{\Psi}^+(x_b,t_b)\hat{\Psi}(x_b,s_b)\hat{n}(x_m,s_m) \dots \hat{n}(x_1,s_1)\hat{\Psi}^+(x_a,s_a)\rangle$$
$$\langle T_\gamma \hat{A}_{p_1}(t_1) \dots \hat{A}_{p_n}(t_n) \hat{A}_{q_m}(s_m) \dots \hat{A}_{q_1}(s_1)\rangle \qquad \text{(A14b)}$$

where $\hat{n} = \hat{\Psi}^+ \hat{\Psi}$ and $\hat{A}_q(t) \equiv \hat{a}_q(t) + \hat{a}^+_{-q}(t)$. (The remaining indices n,m indicating the order of the interaction should not be confused with the transverse modes.)

We now apply the statistical Wick's theorem to the expectation value of the time-ordered product of field operators. One can easily convince oneself that, in the absence of a many-body background, the only nonvanishing contractions are those that pair electron operators in the same time branch, since for each contraction of the type $\langle T_\gamma \psi(t)\psi^+(s)\rangle = \langle \psi(t)\psi^+(s)\rangle$, there must be in the same term another contraction of the type $\langle T_\gamma \psi^+(t)\psi(s)\rangle = 0$. This is not the case for the phonon operators, for which we can have contractions within the same time branch, giving rise to terms of the type $\langle TA_q(s)A^q_+(s')\rangle$ or $\langle \tilde{T}A_q(t)A^q_+(t')\rangle$, or between the two branches, $\langle A_q(t)A^q_+(s)\rangle$ ($\tilde{T}$ stands for anti-time ordering).

The expectation value of the time-ordered product of electron operators can be effectively replaced by

$$n!\, m!\, i^{m-n} G_0^{(-)}(x_a,x_1';t_a-t_1)\dots G_0^{(-)}(x_n',x_b;t_n-t_b)$$
$$G_0^{(+)}(x_b,x_m;s_b-s_m)\dots G_0^{(+)}(x_1,x_a;s_1-s_a) \qquad \text{(A15)}$$

where

$$G_0^{(+)}(x,x';t) = -i\,\langle T\,\psi(x,t)\,\psi^+(x',0)\rangle = -i\,\theta(t)\,\langle \psi(x,t)\,\psi^+(x',0)\rangle \qquad \text{(A16)}$$

is the retarded Green's function and $G_0^{(-)}$ is the advanced Green's function (obtained by replacing $-i\theta(t)$ by $i\theta(-t)$ in (A16)). It is not entirely surprising that we cast our results in terms of quantum-mechanical one-particle Green's functions, since the field-theoretical notation for the electron was introduced by pure convenience.

The contraction of the phonon operators yields three types of phonon Green's functions, depending on whether the contraction takes place (a) within the positive time branch, (b) within the negative branch, or (c) between the two branches. The corresponding expressions are

$$\langle \tilde{T}\,\hat{A}_q(s)\,\hat{A}_q^+(s')\rangle = i\,D_q(s-s') = (N_q+1)\,e^{-i\omega_q|s-s'|} + N_q\,e^{i\omega_q|s-s'|} \qquad \text{(A17a)}$$

$$\langle \tilde{T}\,\hat{A}_q(t)\,\hat{A}_q^+(t')\rangle = i\,\tilde{D}_q(t-t') = -i[D_q(t'-t)]^* \qquad \text{(A17b)}$$

$$\langle \hat{A}_q(t)\,\hat{A}_q^+(s)\rangle = i\,D_q^>(t-s) = (N_q+1)\,e^{-i\omega_q(t-s)} + N_q\,e^{i\omega_q(t-s)} \qquad \text{(A17c)}$$

where N_q is the Bose-Einstein occupation factor (we have used $N_q = N_{-q}$).

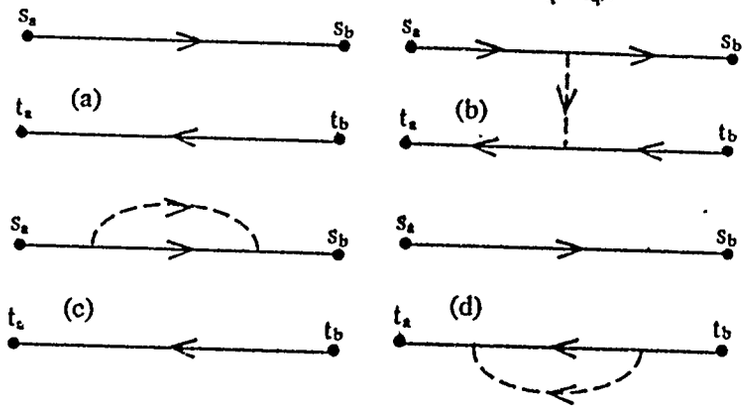


Fig. A2 Expansion diagrams of the lowest order terms.

As a result of Wick's theorem, the perturbative expansion (A14) admits a natural diagrammatic representation. The diagrams for the lowest order terms are shown in Fig. A2. The sign of the arrows is important because it indicates the convention for the sign of the time argument. When inserting an appropriate combination of (A14)-(A17) into (A13), we are left with a multi-dimensional time integral of a function that depends only on the differences between the various times. It seems natural to expand these propagators in terms of their Fourier transform and evaluate the time integrals exactly. We will need the following Fourier transforms

$$iD_q(\omega) \equiv \int \frac{d\omega}{2\pi} e^{-i\omega t}\, iD_q(t) = 2i\omega_q\left[\frac{N_q+1}{\omega^2-\omega_q{}^2+i\eta} - \frac{N_q}{\omega^2-\omega_q{}^2+i\eta}\right] \qquad \text{(A18a)}$$

$$iD_q(\omega) = 2i\omega_q\left[\frac{N_q+1}{\omega^2-\omega_q{}^2+i\eta} - \frac{N_q}{\omega^2-\omega_q{}^2+i\eta}\right] \qquad \text{(A18b)}$$

$$iD_q^>(\omega) = 2\pi[(N_q+1)\delta(\omega-\omega_q) + N_q\delta(\omega+\omega_q) \qquad \text{(A18c)}$$

where $\omega_q > 0$ and $\eta \to 0^+$.

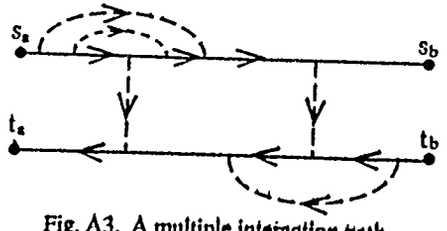


Fig. A3. A multiple interaction path.

At each vertex, there is a time integration whose evaluation yields a delta function that guarantees the conservation of energy. For example, the vertex of Fig. 4 corresponds to

$$\int dt\, e^{-iEt/\hbar}\, e^{iE't/\hbar}\, e^{i\omega t} = 2\pi\hbar\, \delta(\hbar\omega - E + E') \tag{A19}$$

We have adopted an arrow convention in which a given vertex accommodates both the absorption and the emission of phonons, as can be seen by inspection of (A18). By comparing (A18c) for the phonon propagator that connects the two time branches with (A18a,A18b) for those which remain within a branch, it becomes clear that the former corresponds to the real excitation of phonons (and require conservation of energy, as guaranteed by the delta functions) while the latter correspond to virtual phonons that renormalize the particle motion. This is a confirmation of the interpretation that is discussed in the text. There is here a strong analogy with the theory of Raman scattering, where the photon plays the role of the electron in our scattering problem (Toyozawa, 1977). The Feynman rules for the calculation of $P(E_f,E_i,E_i')$ can be formulated easily as the rules are given in Mahan (1981).

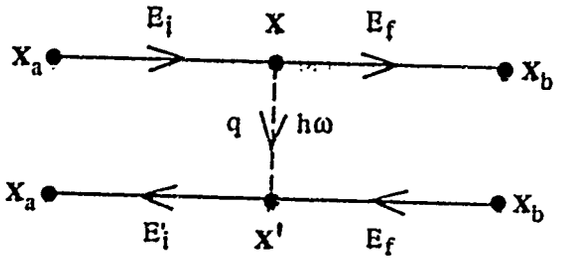


Fig. A3 A typical interaction diagram.

As an illustration, we apply these diagram rules to the calculation of the diagram in Fig. A3, which describes the real creation or absorption of a phonon. We will assume motion in one dimension. We obtain

$$T(E_f,E_i) = h v_f v_i \lambda^2 \iint dx\, dx' \sum_q M^*_q(x)\, M_q(x')$$
$$[(N_q+1)\delta(E_f - E_i + \hbar\omega_q) + N_q\delta(E_f - E_i - \hbar\omega_q)]$$
$$G_0^{(-)}(x_a,x';E_i)\; G_0^{(-)}(x',x_b,E_f)\; G_0^{(+)}(x_b,x;E_f)\; G_0^{(+)}(x,x_a;E_i)\,. \tag{A20}$$

We remark that, although (A20) includes the electron-phonon interaction only to the lowest order, it may well contain the elastic scattering to all orders through the electron Green's functions $G^{(+)}$ and $G^{(-)}$, which can often be calculated exactly by a variety of methods. In the case where there is no elastic scattering (Economou, 1983),

$$G_0^{(+)}(x,x';E) = \frac{e^{ik|x-x'|}}{ihv} = [G_0^{(-)}(x',x;E)]^* \tag{A21}$$

and the final result is

$$T(E_f,E_i) = \frac{\lambda^2}{h^2 v_i v_f} \iint dx\, dx' \sum_q M_q(x) M_q(x') e^{i(k_i - k_f)(x-x')}$$
$$[(N_q+1)\delta(E_f - E_i + \hbar\omega_q) + N_q\delta(E_f - E_i - \hbar\omega_q)] \tag{A22}$$

which agrees exactly with the result that one would have been obtained by using directly the Fermi golden rule in the Born approximation. The alternative derivation based on the Fermi golden rule is given in the text.

APPENDIX B

The propagator for $V(x) = f(t)x$ is

$$G(x_1,x_2,t_1,t_2) = \left\{\frac{mw}{2\pi i \sin wt}\right\}^{d/2} \times$$
$$\times \exp\left\{ \frac{im}{2t}\left[(x_1-x_2)^2 - \frac{2x_1}{m}\int_1^2 f(t')(t'-t_1)dt' - \frac{2x_2}{m}\int_1^2 f(t')(t_2-t')dt' \right] \right\}$$
$$\exp\left\{ \frac{im}{2t}\left[\frac{2}{m^2}\int_1^2\int_1^2 f(t')f(t'')(t_2-t')(t''-t_1)\theta(t'-t'')\,dt'dt'' \right] \right\} \tag{B1}$$

In the case of constant electric and magnetic field, the Lagrangian is $\mathcal{L} = \frac{mx^2}{2} - V(x)$ where the potential for electric field is $V(x) = xF$ and the potential for magnetic field $\mathbf{B} = B(z)x$ is $V(x) = \frac{eB_z}{2c}(xy - yx)$. The action for these elementary problems is

$$S_e(1,2) = \left\{\frac{mw}{2\pi i \sin wt}\right\}^{d/2} \exp\left\{ \frac{im}{2}\left[\frac{(x_1-x_2)^2}{t} - \frac{ft(x_1-x_2)}{m} - \frac{f^2t^3}{24m^2} \right] \right\} \tag{B2}$$

$$S_b(1,2) = \left\{\frac{mw}{2\pi i \sin wt}\right\}^{d/2} \tag{B3}$$

$$\exp\left\{ \frac{im}{2}\left[\frac{(z_1-z_2)^2}{t} - \frac{w\cot(wt)}{2}\left\{ (y_1-y_2)^2 + (x_1-x_2)^2 \right\} + w(x_1y_2 - x_2y_1) \right] \right\}$$

where $\omega = eB_z/cm$.

REFERENCES

Abramowitz, M., and Stegun, I. A., 1964, "Handbook of Mathematical Functions," National Bureau of Standards, U. S. Dept. of Commerce.

Büttiker, M. , 1987, Phys. Rev. B, 35:4123.

Büttiker, M. , 1988a, Phys. Rev. B, 38:12724.

Büttiker, M. , 1988b, IBM J. Res. Dev., 32:306.

Caldeira,A. O., and Leggett, A. J., 1983, Ann Phys. (New York), 149:374.

Datta, S., 1989, Superlattices and Microstructures, 6:86.

DeWitt-Morette, C., Low, S. G., Schulman, L. S., and Shiekh, A. Y., 1986, Foundations of Physics, 16:4, 311.

Economou,E. N., 1983, in "Green's Functions in Quantum Physics", Vol. 7 of "Springer Series in Solid State Sciences," Ed. by Cardona, M., Fulde, P., and Queisser, H. J. Springer, Berlin.
Feynman, R. P., and Hibbs, A. R., 1965, "Quantum Mechanics and Path Integrals," McGraw-Hill, New York.
Feynman, R. P., and Vernon, F. L., 1963, Ann. Phys. (New York), 24:118.
Fisher, D. S., and Lee, P. A., 1981, Phys. Rev. B , 23:6851.
Grossman, I., and Magnus, W., 1964, "Groups and Their Graphs," L. W. Singer, New York.
Guinea, F., and Verges, J.A., 1987, Phys. Rev. B, 35:979.
Keldysh, L. V., 1964, Zh. Eksp. Teor., 47:1515.
Kittel, C., 1983, "Quantum Theory of Solids," John Wiley & Sons, New York.
Landauer, R., 1957, IBM J. Res. Dev., 1:223.
Landauer, R., 1970, Philos. Mag., 21:863.
Landauer, R., 1988, IBM J. Res. Dev., 32:306.
Leggett, A. J., Chakravarty,S., Dorsey, A. T., Fisher, M. P. A., Garg, A., and Zwerger, W., 1987, Rev. Mod. Phys., 59:1.
Loenen, E.T., Dijkhanp, D., Hoeven, A.T., Lenssinck, T.M., and Dieleman, T., 1989, Appl. Phys. Lett., 55:1312.
Lyding, J. W., and Tucker, J. R., Higman, T., Sols, F., Pevzner, V., and Hess, K., 1990, unpublished.
Mahan, G., 1981, "Many-Particle Physics," Plenum, New York.
Mason, B. A., and Hess, K., 1989, Phys. Rev. B, 39:5051.
Molchanov, S. A., 1990, Russ. Math. Surv., 30:1 (translation from Usp. Mat. Nauk, 30:3).
Morse, P. M., and Feschbach, H., 1953, "Methods of Theoretical Physics," McGraw-Hill, New York.
Rammer, J., 1990, Phys. Rev. B, submitted for publication.
Schulman, L. S., 1981, "Techniques and Applications of Path Integration," John Wiley and Sons, New York.
Schulman, L. S., and Ziolkovski, R. W., 1990, in "Proceeding of Third Path Integrals from mev to Mev Conference".
Schwinger, J., 1961, J. Math. Phys., 2:407.
Sols, F., 1990, unpublished.
Sols, F., Macucci, M., Ravaioli, U., and Hess, K., 1989a, J. Appl. Phys., 66:3892.
Sols, F., Macucci, M., Ravaioli, U., and Hess, K., 1989b, Appl. Phys. Lett., 54:350.
Stone, A. D., and Szafer, A., 1988, IBM J. Res. Dev., 32:384.
Taylor, J. R., 1972, "Scattering Theory," John Wiley and Sons, New York.
Thornber, K. K., and Feynman, R. P., 1970, Phys. Rev. B1, 10:4099.
Toyozawa, Y., 1976, J. Phys. Soc. Jpn., 41:400.
Toyozawa, Y., 1977, J. Phys. Soc. Jpn.,42:1495.
Wingreen, N. S., Jacobsen, K. W., and Wilkins, J. W., 1988, Phys. Rev. Lett., 61:1396.

5

CONTACTS AND THE QUANTIZED HALL EFFECT

Markus Büttiker

IBM T.J. Watson Research Center
P.O. Box 218
Yorktown Heights, N.Y. 10598, USA

INTRODUCTION

Ten years after the discovery of the quantum Hall effect by von Klitzing *et al.* (1980,1986), research in this field is still very active and, what is more astonishing, the research is concerned with very fundamental and basic aspects of this phenomena. In the past three years, new theories and experiments have considerably deepened our understanding of this effect (Büttiker, 1990; van Houten *et al.*, 1990).

From the theoretical side, the essential impetus came from a realization that we cannot characterize the resistances of a quantum Hall sample in terms of ρ_{xx} and ρ_{xy}, but must resort to a more general description in terms of global conductances G_{ij}. The indices of the global transport coefficients refer to the contacts of the sample. The net carrier flux incident at a contact i is related to the voltages V_i at these contacts via

$$I_i = \sum_{j \neq i} G_{ij}\left[V_i - V_j\right] , \qquad (1)$$

where V_j are the voltages applied at the contacts. Equation (1) is valid in the linear response regime. According to Onsager and Casimir (1945), in the presence of a magnetic field the transport coefficients of (1) must obey $G_{ij}(B)=G_{ji}(-B)$. A description of a conductance problem in terms of global transport coefficients is necessary as soon as we are interested in phenomena which appear on length scales short compared to an inelastic length. One of the surprises to be discussed below is the discovery of macroscopic equilibration lengths in high mobility GaAs samples.

Büttiker (1986) has given a simple quantum mechanical derivation of (1) in which he expands on a pointof view long advocated by Landauer (1957,1987). It is assumed that the contacts can be described as electron reservoirs which at zero temperature can be characterized by their chemical potentials μ_i. The sample is viewed as an elastic scatterer which permits transmission and reflection of carriers emitted by the reservoirs (see Fig. 1). Inelastic scattering occurs only in the reservoirs which are a source of irreversibility. The reservoir is characterized by M_i quantum channels which allow for incident electron waves and for reflected electron waves. The scattering properties of the sample are characterized by a scattering matrix S which relates the incident current amplitudes to the out-going current amplitudes. Transport can then be characterized by the total transmission probabilities T_{ij} which describe transmission from contact j to contact i and by total reflection probabilities R_{ii} which describe the total probability of carriers incident in probe i to be reflected back into probe i. Büttiker (1986) finds, for the incident carrier flux,

$$I_i = \frac{e}{h}\left[(M_i - R_{ii})\mu_i - \sum_{j \neq i} T_{ij}\mu_j \right] . \qquad (2)$$

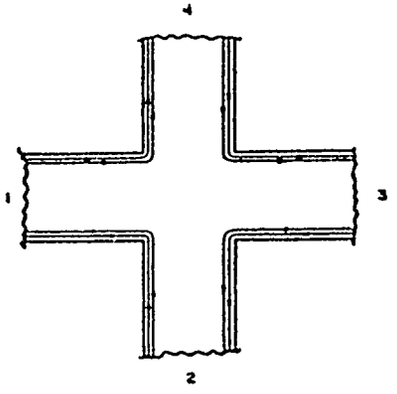


Fig. 1 Four probe conductor connected to electron reservoirs at chemical potentials μ_i. The faint lines indicate the edge states at the Fermi energy.

Due to current conservation all the rows and columns of the matrix of transport coefficients on the right side of (2) add to zero. If this is taken into account, (1) is recovered with $G_{ij}=(e^2/h)T_{ij}$ and $eV_i=\mu_i$. The symmetry of the transport coefficients $T_{ij}(B)=T_{ji}(-B)$ is a direct consequence of the microreversibility of the scattering matrix S. The transport coefficients G_{ij} *are not what is directly measured in an experiment.* But, (1) and (2) can be used to obtain the measured resistances. In a configuration in which contact *m* is the carrier source, contact *n* is the carrier sink, and *k* and *l* are voltage probes, we find that the measured resistance is (Büttiker, 1986)

$$R_{mn,kl} = \frac{h}{e^2}\,\frac{T_{km}T_{ln} - T_{kn}T_{lm}}{D} . \qquad (3)$$

In (3), D is also a function of the transmission probabilities but is independent of the indices *mn,kl*. A resistance, for which the voltage contacts are on opposite sides of an imaginary line connecting the current source contact and the current sink contact, is called a generalized Hall resistance; a resistance, for which the voltage contacts are on the same side of this imaginary line, is called a generalized longitudinal resistance. The important features of this result are the following: (3) results from an equivalent treatment of both current contacts and voltage contacts. As a consequence, both the longitudinal resistances and the Hall resistances are treated on an equivalent footing. The resistance is expressed in terms of the chemical potentials of the reservoirs and the net current flowing at these contacts. The electric field distribution inside the conductor might be very complicated and, in any case, depends on the screening properties of the conductor. But, to arrive at (3), it is not necessary to calculate the electric field. The voltages which enter are chemical potential differences of the reservoirs. If the role of the current contacts and voltage contacts is exchanged, and the magnetic field is reversed, microreversibility leads to the reciprocity of resistances $R_{mn,kl}(B)=R_{kl,mn}(-B)$. It is this symmetry (Büttiker, 1986, 1988a) which is observed experimentally. Here, we mention only the experiments on Aharonov-Bohm loops by Benoit *et al.* (1986) and the electron focusing experiments by van Houten *et al.* (1989). An additional feature of (2) and (3), which is crucial

for the discussion which follows, is that, in linear response, the transport coefficients (the transmission probabilities) are evaluated at the Fermi energy. It is the transmission of carriers from one contact to another at the Fermi energy which matters. Clearly, the typical explanation of the quantum Hall effect (Prange and Girvin, 1987), which considers only extended states at the center of bulk Landau levels and localized states away from the center of the Landau levels, leads to a paradox. In this picture there are no states available for transport at the Fermi energy, when the Fermi level lies between bulk Landau levels, i.e. exactly when the quantized Hall effect occurs. The fact that the quantum Hall effect is observed tells us that there must be extended states available at the Fermi energy even in the case when the Fermi energy is between bulk Landau levels.

Equations (2) and (3) have been successfully applied to a variety of problems. Here, we point especially to the low field magneto-transport anomalies found in ballistic conductors by Roukes *et al.* (1987), Takagaki *et al.* (1988), Ford *et al.* (1989), and Chang *et al.* (1989). A number of theoretical efforts by Kirczenow (1989a,1989b), Avishai and Band (1989) and Ravenhall *et al.* (1989) have recently led to more realistic descriptions by Baranger and Stone (1989a), and Beenakker and van Houten (1989), which successfully reproduce essential physical features. The success of the approach discussed above has also led to efforts to derive these equations from formal linear response theory and a recent paper by Baranger and Stone (1989b) achieves that.

Equations (2) and (3) were applied to the quantum Hall effect by Büttiker (1988a) and independently by Peeters (1988) and Beenakker and van Houten (1988). This work followed an attempt by Streda *et al.* (1987) and similar work by Jain and Kivelson (1988) who appealed to the build up of local charges to determine Hall voltages. In contrast, the discussion of Büttiker (1988a) emphasizes the need to consider all contacts.

MOTION IN HIGH MAGNETIC FIELDS

A simple picture of the electronic states at the Fermi energy can be obtained for potentials which fluctuate slowly on the scale of the magnetic length $L_B=(\hbar/eB)^{1/2}$. In high mobility samples, where donors are separated by a spacer layer (Esfarjani *et al.*, 1990) from the two dimensional electron gas, the fluctuations in the potential are determined by the width of the spacer layer. For such samples, a quasi-classical discussion of electron motion can be expected to provide an excellent starting point. We consider, therefore, the motion of the guiding center coordinate. In the limit of large fields, the trajectories of the guiding center are found as solutions of the equation

$$E_F = \hbar\omega_c(n+\tfrac{1}{2}) + V(x,y) \quad . \qquad (4)$$

where ω_c is the cyclotron frequency and V(x,y) is the potential. Equation (4) has (essentially) two types of solutions: trajectories which describe closed loops, and trajectories which emanate at one contact and terminate at another contact. Closed loop trajectories (localized states) occur predominantly in the bulk of the sample. Extended trajectories (edge states) occur along the boundaries of the sample, since near a sample edge the confining potential is much stronger than the fluctuating part of the potential. Each bulk Landau level below the Fermi energy gives rise to an edge state at the Fermi energy. Figure 2, taken from Büttiker (1990), illustrates the states at the Fermi energy. The Fermi energy is between the second and third bulk Landau levels such that there are only a few localized states at the Fermi energy. The arrows indicate the direction of motion of carriers along these trajectories. The box labeled A designates an area where the quasi-classical description breaks down due, for instance, to a residual impurity. The broken line B marks a spot where the innermost edge states approach so closely that tunneling or Mott hopping processes, via localized states, lead to scattering from one edge state on one side of the sample to another edge state on the other side of the sample. Process B is called a backscattering process: it reverses the direction of motion of a carrier. Let us return to the discussion of residual impurities marked A in Fig. 2, assuming for the moment that there are no backscattering processes. If there are no backscattering processes, each of the edge states entering box A is filled up to the same energy and, in a small energy interval at the Fermi energy, carries a unit carrier flux. Regardless of the scattering in box A, the edge states leaving this box are still filled up to the same energy and carry a unit flux, as long as the scattering does not lead to backscattering across the bulk of the sample. Hence, in the absence of backscattering, the edge states are immune to scattering (residual impurities, surface roughness,...). Clearly, this is not the case if backscattering occurs. Due to process B, the innermost edge state is partially depleted. Hence, at A the two edge states entering the box have differing fluxes (or might be filled up to different energies). The scattering at A, in the presence of backscattering, leads to a redistribution of carriers over the two edge states. The main conclusion from this discussion is the following: *in the absence of backscattering, the edge states provide perfect quantum channels for the transmission of carriers from one contact to another.* Since each edge state allows for the transmission of unit flux, the total transmission probability along a sample edge is equal to the number N of edge states.

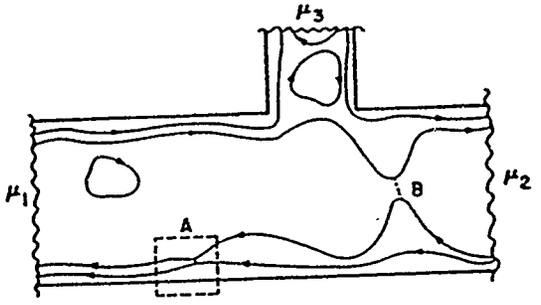


Fig. 2 States at the Fermi energy of a three probe conductor. At A the edge states interact due to a strong potential fluctuation. B denotes a backscattering channel.

THE QUANTIZED HALL EFFECT

The previous discussion illuminates the simple manner in which the quantized Hall effect arises (Büttiker, 1988a). In the absence of backscattering, each edge of the sample sustains N quantum channels which allow for perfect transmission of carriers from one contact to another in a circulatory fashion. The transmission probabilities, for the conductor of Fig. 1, are $T_{41}=T_{34}=T_{23}=T_{12}=N$. All other transmission probabilities are equal to zero. Using (2) and (3) it is easy to show that $D=N^3$ and that the Hall resistances are quantized to the values $R_{13,42}=-R_{42,13}=(h/e^2N)$, and the longitudinal resistances are zero, $R_{14,23}=R_{43,12}=0$. If the magnetic field is reversed, then $T_{21}=T_{32}=T_{43}=T_{14}=N$ are quantized, and all other transmission probabilities are zero. D is invariant under field reversal and consequently the Hall resistances change signs. If the field is increased, the innermost edge states move away from the sample boundary into the bulk of the sample and the density of localized states increases. Backscattering begins to occur and this causes the transmission probabilities mentioned above to become smaller than N. All transmission probabilities are now non-zero. Consequently, the Hall resistance is not quantized and the longitudinal resistances are different from zero. To arrive at these results, we have assumed that no backscattering processes occur and we have assumed that carriers reaching a contact leave the sample with probability 1. Below, we examine differing conductors for which these conditions are not obeyed.

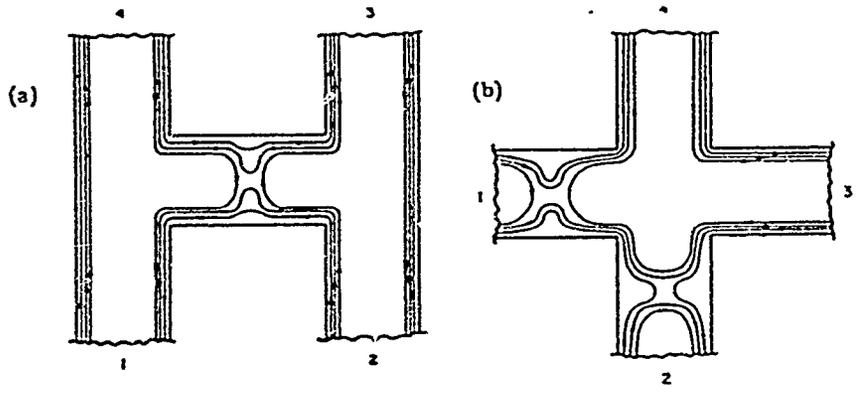


Fig. 3 (a) Conductor with a gate leading to quantized backscattering. (b) Conductor with two adjacent contacts with internal reflection and two ideal contacts.

QUANTIZED BACKSCATTERING

In the conductor of Fig. 3a, a barrier is induced with the help of a gate across the conductor. Suppose that, for zero gate voltage, the magnetic filed is adjusted such that we have N edge states at the Fermi energy. In this case, the discussion given for the conductor of Fig. 1 applies. If the gate voltage is increased, the density of carriers beneath the gate is depleted. One edge state after another, with guiding center energy $E_G=E_F-\hbar\omega_c(n+1)/2$ is reflected at the barrier. Suppose the gate voltage is in a range where K edge states are reflected and only $N_G=N-K$ edge states penetrate the barrier. Application of (3) yields Hall resistances (Büttiker, 1988a,1989)

$$R_{13,42} = \frac{h}{e^2 N_G} \quad . \tag{5}$$

$$R_{42,13} = R_{13,42} = -\frac{h}{e^2}\left(\frac{2}{N} - \frac{1}{N_G}\right) \quad . \tag{6}$$

and longitudinal resistances

$$R_{12,43} = R_{43,12} = \frac{h}{e^2}\left(\frac{1}{N_G} - \frac{1}{N}\right) \quad . \tag{7}$$

Interestingly, according to (5)-(7), the Hall resistances and the longitudinal resistances can be simultaneously quantized. Note also, that the Hall resistances are not anti-symmetric under field reversal. Equation (7) was found in Büttiker (1987) and independently by van Houten *et al.* (1988). Well resolved plateaus, as predicted by (5)-(7), have been seen in experiments by Washburn *et al.* (1988) in a small sample with a lithographic width of only 2 μm and a gate width of 0.1 μm and by Haug *et al.* (1988,1989) in a sample with considerably larger dimensions. The plateaus predicted by (5)-(7) have also been measured by Snell *et al.* (1989) and Main *et al.* (1989) in a sample consisting of split gate constrictions. Equation (7) can easily be generalized to the case where there is only partial transmission or reflection of one of the edge states.

CONTACTS WITH INTERNAL REFLECTION

The contacts which we have invoked so far are ideal. For quantizing fields, carriers approaching the contact leave the sample with probability 1. Contacts which are not ideal can arise for a variety of reasons. A contact which is not ideal reflects carriers approaching the contact back into the conductor. (Since the reservoir has a density of states which is large compared to the number of edge states even ideal contacts exhibit external reflection). Internal reflection of a contact (Büttiker,1988b) can be introduced by insufficient alloying of metallic contacts, building a gate over a contact, by using split gates or metallic "interior contacts", which are located a distance which is large compared to a magnetic length away from a sample, edge. In Fig. 2, contact 1 is ideal. Contact 2, in the presence of backscattering, acts like a contact with internal reflection. Contact 3 only couples to the outer edge state and thus exhibits internal reflection. Figure 3b is a schematic view of a sample with two adjoining contacts with internal reflection, as used in an experiment by van Wees *et al.* (1989a). If a current source contact exhibits internal reflection the edge states past the contact are not equally populated by the current source. If a voltage contact is used to measure voltage, and also exhibits internal reflection and is close enough to the carrier source such that inelastic scattering is not effective. the Hall resistance exhibits deviations from the normal quantized Hall effect (Büttiker, 1988b). For the situation of Fig. 3b, (3) yields a Hall resistance (van Wees *et al.*, 1989a)

$$R_{24,13} = \frac{h}{e^2}\frac{T_{12}}{T_1 T_2} \quad . \tag{8}$$

Here T_{12} is the overall transmission probability for carriers incident in contact 2 to reach contact 1, and $T_1=T_{41}$ and $T_2=T_{23}$ are the transmission probabilities which characterize the two-terminal resistance of the contacts 1 and 2 (measured by grounding the voltages at the other contacts). Consider the case shown in Fig. 2a, where the contact conductances are quantized. N_1 edge states are perfectly transmitted at contact 1 and N_2 states are perfectly transmitted at contact 2. Clearly, as seen form Fig. 2a, the overall transmission probability from contact 2 to contact 1 is determined by the contact with the lower transmission probability. Therefore, $T_{12}=\min(N_1,N_2)$ and (8) predicts a Hall resistance which is quantized and proportional to $(\max(N_1,N_2))^{-1}$. Quantized Hall resistance plateaus at such "anomalous" values have been observed by van Wees *et al.* (1989a). Note that these plateaus are entirely determined by the properties of the contacts.

MESOSCOPIC EFFECTS ON MACROSCOPIC LENGTH SCALES

In the experiment of van Wees *et al.* (1989a) described above, the contacts with internal reflection are only a distance of the order of 1.5 μm apart. Komiyama *et al.* (1989, 1990a, 1990b) performed experiments on gated structures similar to those of Washburn *et al.* (1988) and Haug *et al.* (1988,1989). However, in contrast to these latter experiments, they found deviations of the measured resistances form those predicted by (5)-(7). In their experiment, the contacts are more than 50 μm from the barrier. Could it be that these deviations arose because the contacts were not, as in the Washburn *et al.* and Haug *et al.* experiments, ideal? Measurement of the contact resistances (three terminal resistances) revealed that these were indeed not small or quantized, lending support to the idea that the deviations from quantization are a consequence of contacts with internal reflection. Subsequently, Komiyama and Hirai (1989) extended the Büttiker (1988a) theory of contacts and showed that such an extension was indeed capable of providing a quantitative explanation of their data.

Van Wees *et al.* (1989b) used a sample with point contacts separated more than 200 μm from one another and showed that the Shubnikov de Haas oscillations disappear when the contacts couple only to the outermost edge state. Both the Komiyama *et al.* (1989, 1990a, 1990b) experiments and the van Wees *et al.* (1989b) experiment suggested, therefore, that non-equilibrium populations created at current contacts in high mobility GaAs samples are maintained over macroscopic distances.

An intriguing experiment to further clarify the long range of non-equilibrium situations was made by Alphenaar *et al.* (1990). In this experiment, two point contacts are used which are separated by a distance of 80 μm. The carrier source contact is adjusted such that only the outermost edge state transmits the point contact. Therefore, $T_2=1$ in (8). Suppose now that, at the entrance to the voltage contact 1, a fraction α_1 of the injected current is in the outermost edge state and a fraction α_i has been scattered (elastically or inelastically) into the i-th edge state. (Current conservation requires that if we sum the individual α_i over the total number of edge states, this sum must be 1.) If the voltage contact is successively opened to permit perfect transmission of 1, 2, ..., N_1 edge states, (8) leads to a Hall resistance

$$R_{24,13} = \frac{h}{e^2}\frac{1}{N_1}\sum_{i=1}^{N_1}\alpha_i \quad . \tag{9}$$

At $T=0.45$ K, Alphenaar *et al.* (1990) found in their experiment $\alpha_1=0.48$, $\alpha_2=0.44$, and $\alpha_3=0.08$, for a field at which there are three edge states. This result indicates that the two outermost edge states equilibrate but the innermost edge state is basically decoupled from the other edge states. Alphenaar *et al.* substantiate these results by additional experiments which show that the N-1 outermost edge states equilibrate but that these states are decoupled from the innermost edge state.

The contacts used in the van Wees *et al.* (1989b), Alphenaar *et al.* (1990) and possibly also in the Komiyama *et al.* (1989, 1990a, 1990b) experiments preferentially populate the outermost edge states. But, it is also possible to populate or detect preferentially the innermost edge state. This can be achieved with contacts which are located in the interior of the sample but near the edges of the two-dimensional electron gas. Experiments on a sample with contacts of this type have recently been performed by Faist *et al.* (1990). The samples used by Faist *et al.* exhibit both interior contacts, which couple preferentially to the innermost edge state, and exterior contacts, which couple to all edge states. The contacts are more than 250 μm apart and the sample width is larger than 80 μm. For quantizing fields, longitudinal resistances measured over such distances with interior contacts can be negative even when the current source and the current sink are ideal contacts. Under similar conditions, deviations of the Hall resistance occurs. We must conclude from these results that, even for quantizing fields, there are very small backscattering processes which bring the innermost edge state out of equilibrium with the N-1 outer edge states. If current is fed through interior contacts, and the voltage is also measured with interior contacts, the Shubnikov-de Haas oscillations are enhanced compared to what is measured with exterior contacts. Negative resistance fluctuations at minima of the longitudinal resistance of submicron conductors have been observed before, and simple models have been proposed to explain them (Chang *et al.*, 1988). In contrast the experiment of Faist *et al.* (1990) measures negative longitudinal resistances over macroscopic distances.

A clear understanding of the reasons for the decoupling of the innermost edge state from the N-1 outer edge states has not been yet achieved. Martin and Feng (1990) have pointed out that, in high magnetic fields, scattering rates are suppressed. In the Born approximation, they obtain

$$\frac{1}{\tau_{el,n,n+1}} = \frac{1}{\tau_0}\exp\left\{-\frac{(y_n - y_{n+1})^2}{2L_B^2}\right\} .$$

where y_n is the guiding center position of the n-th edge state at the Fermi energy. Similarly, they find that the electron-phonon scattering rate is suppressed at high fields. The elastic scattering rate, however, is calculated for sharp impurity potentials (delta functions). If, instead, a smooth impurity potential is assumed proportional to

$$\exp\left\{-\frac{(r-r_0)^2}{2S^2}\right\} ,$$

where S can be taken to be equal to the width of the spacer layer, the suppression of elastic scattering is much stronger and proportional to

$$\exp\left\{-\frac{(y_n - y_{n+1})^2}{2S^2}\right\} ,$$

as pointed out by Ono and Ohtsuki (1989). A brief calculation shows that, in the Born approximation, the elastic impurity averaged scattering rate is

$$\frac{1}{\tau_{el,n,n+1}} = \frac{1}{\tau_0}\exp\left\{-\frac{(y_n - y_{n+1})^2}{2L_B^2}\left(1+\frac{S^2}{L_B^2}\right)\right\} .$$

The separation of the N-1 outermost edge states is chiefly determined by the confining potential and thus might not be much larger than a magnetic length. The separation between the N-1 outermost edge states and the innermost edge state in contrast is chiefly determined by the weakly, and slowly, fluctuating impurity potential. The result given above might then be applicable leading at high fields to an effective suppression of elastic scattering between the innermost edge state and the N-1 outermost edge states. An additional effect, which might help to explain why very little current is detected in the innermost edge state, is the following. The backscattering processes, which exist under quantizing conditions, are very small, but they might be comparable to the scattering rate between the N-1 outermost edge states and the innermost edge state. Backscattering then helps to deplete the innermost edge state. Similarly, it can be shown that scattering of electrons by acoustic phonons is not effective in equilibrating neighboring edge states in a region of weak field. It is known that if the drift velocity $v_d=cE/B$ is smaller than the velocity of sound v_s, phonon emission does not occur (Heinonen *et al.*, 1984). Emission of phonons with a small electronic energy change requires that the drift velocity exceeds the velocity of sound. Both the weak elastic scattering rate and the absence of phonon emission in weak fields are thus compatible with the absence of interactions between the innermost edge state and the N-1 outermost edge states. On the other hand, the Born approximation is not very suitable to treat scattering in a highly non-unifrom system.

The long range over which non-equilibrium populations are maintained is also revealed in a paper by Haug and von Klitzing (1989). They point out that the Shubnikov-de Haas oscillations do not scale with sample width and length between contacts as one would expect from a classical result, i. e. the longitudinal resistances are not proportional to $\rho_{xx}L/W$. Clearly, this is to be expected if, even away from the plateaus, N-1 edge states persist and are decoupled from the innermost edge state. Away from a Hall plateau, all contacts exhibit internal reflection. A simple phenomenological model proposed by Szafer *et al.* (1990) takes the decoupling into account and attributes, to each segment of width W and length L of the conductor, a transmission probability

$$N + 1 + \frac{1}{1+\frac{\rho L}{W}} .$$

Here ρ is a bulk resistivity characterizing backscattering between the innermost edge states, i.e. in the highest occupied Landau level. No backscattering is taken into account in the junctions. As shown by McEuen *et al.* (1990), this model gives good agreement with four-terminal resistances measured on macroscopic conductors with aspect ratios W<L.

Clearly, recent theoretical progress, and the many experiments discussed here, have led to a number of novel and fascinating questions. Of technical importance are possible limitations on the high precision measurement of the quantized Hall resistance. A paper by Hirai and Komiyama (1990) estimates deviations due to weak internal reflection at contacts. Our discussion in this paper was restricted to the integer quantum Hall effect. But a set of interesting experiments in the fractional quantum Hall regime by Chang and Cunningham (1989), Kouwenhoven *et al.* (1990), and Simmons *et al.* (1989) point to effects similar to those we discussed for the integer quantum Hall effect. Recent work by Beenakker (1990), MacDonald (1990), Kivelson and Prokovsky (1989) and Chang (1990) has begun to address the nature of edge states in the fractional quantum Hall regime. It is, therefore, likely that in the coming years we will witness a continued interest in the quantized Hall effect.

REFERENCES

Alphenaar, B. W., McEuen, P. L., Wheeler, R. G., and Sacks, R. N., 1990, Phys. Rev. Letters, 64:677

Avishai, Y., and Band, Y., 1989, Phys. Rev. Letters, 62:2527.

Baranger, H. U., and Stone, A. D., 1989a, Phys. Rev. Letters, 63:414.

Baranger, H. U., and Stone, A. D., 1989b, Phys. Rev. B, 40:8169.

Beenakker, C. W. J., 1990, Phys. Rev. Letters, 64:216.

Beenakker, C. W. J., and van Houten, H., 1988, Phys. Rev. Letters, 60:2406.

Beenakker, C. W. J., and van Houten, H., 1989, Phys. Rev. B, 39:10445.

Benoit, A. D., Washburn, S., Umbach, C. P., Laibowitz, R. B., and Webb, R. A., 1986, Phys. Rev. Letters, 57:1765.

Büttiker, M., 1986, Phys. Rev. Letters, 57:176.

Büttiker, M., 1987, IBM J. Res. Devel., 32:317.

Büttiker, M., 1988a, Phys. Rev. B, 38:9375.

Büttiker, M., 1988b, Phys. Rev. B, 38:12724.

Büttiker, M., 1989, Phys. Rev. Letters, 62:229.

Büttiker, M., 1990, in "Nanostructured Systems", M. A. Reed, Ed., Academic Press, New York.

Casimir, H. B. G., 1945, Rev. Mod. Phys., 17:343.

Chang, A. M., 1990, unpublished.

Chang, A. M., and Cunningham, J. E., 1989, Sol. State Commun., 72:651.

Chang, A. M., Timp, G., Cunningham, J. E., Mankiewich, P. M., Behringer, R. E., and Howard, R. E., 1988, Sol. State Commun., 76:769.

Chang, A. M., Chang, T. Y., and Baranger, H. U., 1989, Phys. Rev. Letters, 63:2268.

Esfarjani, K., Glyde, H. R., and Sa-yakanit, V., 1990, Phys. Rev. B, 41:1990.

Faist, J., Meier, H. P., and Gueret, P., 1990, Interior Contacts for Probing the Equilibrium between Magnetic Edge Channels in the Quantum Hall Effect, unpublished.

Ford, C. J. B., Washburn, S., Büttiker, M., Knoedler, C. M., and Hong, J. M., 1989, Phys. Rev. Letters, 62:2724.

Harai, H., and Komiyama, S., 1990, A Contact Limited Precision of the Quantized Hall Resistance, unpublished.

Haug, R. J., and von Klitzing, K., 1989, Europhys. Letters, 10:489.

Haug, R. J., MacDonald, A. H., Streda, P., and von Klitzing, K., 1988, Phys. Rev. Letters, 61:2797.

Haug, R. J., Kucera, J., Sueda, P., and von Klitzing, K., 1989, Phys. Rev. B, 39:10892.

Heinonen, O., Taylor, P. L., and Girvin, S. M., 1984, Phys. Rev. B, 30:3016.

Jain, J. K., and Kivelson, S. A., 1988, Phys. Rev. Letters, 60:1542.

Kirczenow, G., 1989a, Phys. Rev. Letters, 62:2993.

Kirczenow, G., 1989b, Sol. State Commun., 71:469.

Kivelson, S. A., and Prokovsky, V. L., 1989, Phys. Rev. B, 40:1373

Kouvenhoven, L. P., van Wees, B. J., van der Vaart, N. C., Harmans, C. J. P. M., Timmering, C. E., and Foxon, C. T., 1990, Phys. Rev. Letters, 64:685.

Komiyama, S., and Hirai, H., 1989, Phys. Rev. B, 40:7767.

Komiyama, S., Hirai, H. Sasa, S., and Fujii, T., 1989, Phys. Rev. B, 40:12566.

Komiyama, S., Hirai, H., Sasa, S., and Fujii, T., 1990a, Sol. State Commun., 73:91.

Komiyama, S., Hirai, H., Sasa, S., and Fujii, T., 1990b, J. Phys. Soc. Jpn., 58:4086.

Landauer, R., 1986, Phys. Rev. Letters, 57:1761.

Landauer, R., 1987, Z. Phys., B68:217.

MacDonald, A. H., 1990, Phys. Rev. Letters, 64:220.

Main, P. C., Beton, P. H., Snell, B. R., Neves, A. J. M., Owers-Bradley, J. R., Eaves, L., Beaumont, S. P., and Wilkinson, C. D. W., 1989, Phys. Rev. B, 40:10003.

Martin, T., and Feng, S., 1990, Phys. Rev. Letters, 64:1971.

McEuen, P. L., Szafer, A., Richter, C. A., Alphanaar, B. W., Jain, J. K., Stone, A. D., Wheeler, R. G., and Sacks, R. N., 1990, unpublished.

Ohtsuki, T., and Ono, Y., 1989, J. Phys. Soc. Jpn., 58:3863.

Peeters, F. M., 1988, Phys. Rev. Letters, 61:589.

Prange, R. E., and Girvin, S. M., 1987, "The Quantum Hall Effect", Springer-Verlag, New York.

Ravenhall, D. G., Wyld, H. W., and Schult, R. L., 1989, Phys. Rev. Letters, 62:1780.

Roukes, M. L., Scherer, A., Allen, S. J., Jr., Craighead, H. G., Ruthen, R. M., Beebe, E. D., and Harbison, J. P., 1987, Phys. Rev. Letters, 59:3011.

Simmons, J. A., Wei, H. P., Engel, L. W., Tsui, D. C., and Shayegan, M., 1989, Phys. Rev. Letters, 63:1731.

Snell, B. R., Beton, P. H., Main, P. C., Neves, A., Owers-Bradley, J. R., Eaves, L., Hennini, M., Hughes, O. H., Beaumont, S. P., and Wilkinson, C. D. W., 1989, J. Phys. C, 1:7499.

Streda, P., Kucera, J., and MacDonald, A. H., 1987, Phys. Rev. Letters, 59:1973.

Szafer, A., McEuen, P. L., Jain, J. K., and Stone, A. D., 1990, unpublished.

Takagaki, Y., Gamo, K., Namba, S., Ishida, S., Takaoka, S., Murase, K., Ishibashi, K, and Aoyagi, Y., 1988, Sol. State Commun., 68:1051.

van Houten, H., Beenakker, C. W. J., van Loosdrecht, P. H. M., Thornton, T. J., Ahmed, H., Pepper, M., Foxon, C. T., and Harris, J. J., 1988, Phys. Rev. B, 37:8534.

van Houten, H., Beenakker, C. W. J., Williamson, J. G., Broekaart, M. E. I., van Loosdrecht, P. H. M., wan Wees, B. J., Moij, J. E., Foxon, C. T., and Harris, J. J., 1989, Phys. Rev. B, 39:8556.

van Houten, H., Beenakker, C. W. J., and van Wees, B. J., 1990, in "Nanostructured Systems", M. A. Reed, Ed., Academic Press, New York.

van Wees, B. J., Willems, E. M. M., Harmans, C. J. P. M., Beenakker, C. W. J., van Houten, H., Williamson, J. G., Foxon, C. T., and Harris, J. J., 1989a, Phys. Rev. Letters, 62:1181.

van Wees, B. J., Willems, E. M. M., Kouwenhoven, L. P., Harmans, C. J. P. M., Williamson, H. G., Foxon, C. T., and Harris, J. J., 1989b, Phys. Rev. B, 39:8066.

von Klitzing, K., 1986, Rev. Mod. Phys., 58:519.

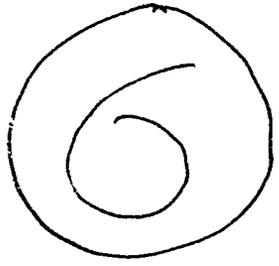
von Klitzing, K., Dorda, G., and Pepper, M., 1980, Phys. Rev. Letters, 45:494.

Washburn, S., Fowler, A. B., Schmid, H., and Kern, D., 1988, Phys. Rev. Letters, 61:2801.

INTERFERENCE DEVICES

Alan B. Fowler

IBM T. J. Watson Research Center
P. O. Box 218, Yorktown Heights, N.Y. 10598



INTRODUCTION

In two-dimensional gases, when the mean free paths of the electrons are long compared to sample dimensions, it is clear that the electrons will propagate ballistically until reflected by some sort of a boundary - either the sample edge or some other introduced sharp fluctuation in the potential. In the GaAs/GaAlAs heterojunction systems, it is relatively easy to attain mobilities high enough so that mean free paths are micrometers or more. It is also relatively easy at many laboratories to build structures lithographically that are much smaller. It has been demonstrated by many that structures can be made in which the dimensions can be reduced to a one-dimensional quantum wire - wherein the electronic states are quantized perpendicular to the wire (in both directions) and k, the wave vector, is a good quantum number only along the wire (Warren *et al.*, 1986; Kotthaus *et al.*, 1988; Berggren *et al.*, 1986; Smith *et al.*, 1987).

One might suppose that, given the coherence of the electrons over distances many times larger than the structures, it should be possible to construct active devices based upon interferometric principles - devices that take advantage of the coherent wave nature of the electrons passing through a structure in which interference of some sort exists. The final element to make an active device must be to provide a means for altering the interference patterns - normally by changing the wave vector (although one could change a length). The last is easily done in a field effect device where changing a gate voltage changes the number of induced electrons and therefore the Fermi energy E_F and the Fermi wave vector k_F. Since only the electrons within kBT of the Fermi surface contribute to the conduction, it is k_F that is critical to the interference condition. Note that raising the temperature or otherwise heating the electrons can destroy the interference effects.

WAVEGUIDES AND CONDUCTANCE

The relevant properties of an electron waveguide are surprisingly difficult to calculate. If the confining structure is well defined, it is possible, with considerable computing effort, (Laux and Stern, 1986; Laux and Warren, 1986) to simulate the structure in the sense that the quantum levels (subbands), their filling k_F for each subband, etc., can be determined at least for a wire of constant cross section, so that only a two-dimensional, self-consistent solution of Poisson's equation and Schrodinger's equation is required. It is also possible to attain solutions in three dimensions for simple cases (boxes, periodic boundary conditions, etc.). Further, there are approximate solutions that at least give insight (Berggren *et al.*, 1986; Shik, 1985).

The conductance through an arbitrary structure is most simply given by Landauer's equation

$$G = \frac{e^2}{h} \sum_n T_n \ , \qquad (1)$$

where the sum is taken over the quantum mechanical transmissivities of each of the channels. Each spin contributes a channel. Each occupied one-dimensional subband contributes a channel, because at 0 K only the electrons at the Fermi surface contribute to the conductance. In principle, then, the calculation of interference is easy. One simply needs to know $T_n(k_{Fn})$ and $k_{Fn}(V_g)$. It is obvious that, in practice, for complex structures, this is not easy and only a few idealized structures have been modeled in fact (Büttiker, 1985; Sols *et al.*, 1989).

Conceptually, the simplest interference structure would be a one-dimensional wire (Fowler and Hartstein, 1985) with a gate of length L across it. If the gate were adjusted so that $k_F=k_F'$ under the gate were different from that in the rest of the wire k_F, and if the boundaries were abrupt, then the transmission is just

$$T = \left[1 + (E_F - E_F')^2 \frac{\sinh(\beta L)}{4E_F E_F'} \right] , \qquad (2)$$

where

$$\beta = \sqrt{\frac{2mE_F'}{h^2}} \ , \qquad (3)$$

which is a classic result discussed in almost all elementary quantum mechanics texts. The transmission is step-like at $E_F'=0$, with a soft turn-on and some structure above that point.

A real structure does not have an abrupt step because the gates are at some distance from the wire. In the case of a device based upon a GaAs/GaAlAs modulation-doped structure, the gates are some 20 nm or more from the GaAs channels. The result is that the potential rises gradually so that the transition is adiabatic. Therefore, the rise in current is gradual rather than a modified step. Experiments (Washburn *et al.*, 1988) with 40 nm long gates showed no effects of interference and simulations confirmed that there should not be. Structure was seen, but was believed to be due to fluctuations (Davies and Nixon, 1989) in the barrier height resulting from the discreteness and fluctuations in the charge. In silicon MOSFET structures, the gates can be placed 3 nm from the interfaces, so that more abrupt steps could be made; however, the longest mean free paths are only 50 nm in two dimensions. In one dimension, they may be longer so that for 10-20 nm gates, observation of interference may be feasible in silicon.

When two or more gates are in series, the interference can be reinforced. In the limit of many gates, the structure is periodic and is considered as a superlattice.

A different approach is to interfere waves passing through different paths. The simplest idea might be a ring with two ports (Fowler, 1984; Datta *et al.*, 1985). If the paths are equal, then clearly if a gate covers part of one arm for a length δL, then there will be a periodicity for a condition

$$(k_F - k_F')\delta L = 2\pi n \ , \qquad (4)$$

where k_F and k_F' are the wave vectors of the unperturbed ring and of the region under the gate. Of course, there might well be other interference effects as well, because there could be

reflections at the ports. Therefore, if the circumference of the ring were 2L, another condition might be that $k_F(2L-\delta L)+k_F'\delta L=2\pi n$.

A variation is to make the equivalent of a microwave stub tuner (Büttiker, 1985) with a side arm, so that the wave traveling down theside arm and reflecting back interfaeres with the incident wave. The stub could be tuned, as before, with a separate gate of length δL so that now

$$2\delta L k_F' = 2\pi(n + \Phi) \ . \qquad (5)$$

Another approach would be to gate the entire structure. Then, if the stub length were L, the interference condition would be

$$2Lk_F = 2\pi n \ . \qquad (6)$$

For this device to switch properly, the amplitude of the wave in the side arm would have to be one-half that of the incident wave. Experiments and theory indicate that the electrons in the lowest subband do not tend to deflect into side arms very efficiently, but amodification of the shape to a Y might get around this problem. Sols *et al.* (1989) have simulated this structure, but I am not aware that one has been made.

One might use a ring with unequal paths between and gate the whole structure. The interference condition would be

$$k_F\delta L = 2\pi n \ , \qquad (7)$$

where δL is the difference in path length. This structure has been made by Ford *et al.* (1990) using a MODFET-based technology. The structures, with δL~1 μm, were made by etching GaAs on top of AlGaAs to define the wires, and depositing a Schottky gate over the whole structure. It was expected that the structure could be easily modeled. The resulting structure showed about 10-20% modulation, but the period could not be fit by the model, by a factor of about three. Also, for reasons discussed below, the device did not conduct when only the lowest subband was occupied, so that the interpretation of the results was difficult.

Part of the problem in making such structures is that the lines are not uniform. Non-uniformity can arise either from fluctuations in the doping or from irregular lithography. We believe the former is the primary source of fluctuations, a view supported by simulations of Nixon and Davies (1990). The simplest way to understand this problem is to considerthe number of doping impurities in a square with sides the dimension of the wires. This typically is 50 (50±7). Obviously, the threshold can vary strongly along the wires so that some parts will be non-conducting even though most of the wire conducts.

CONCLUSIONS

At this point, it should be obvious that several factors can act to make it difficult to make a structure that is a good switch in the sense that the device is either "on" or "off". Unless only a single subband is occupied, the interference condition for each is different. Further, the dependence of k_F on gate voltage is different for each subband. This means that the devices would have to be operated in the lowest subband only. Therefore, the maximum possible conductance would be $2e^2/h$ or 80 μS. This is not the only barrier to their usefulness. In addition, k_BT must be small compared to the change in E_F' needed to satisfy the interference condition. (For a gate length of 20 nm, δE_F would be about 4 K. For more realistic lengths, it would go down.) And one would have to match the inputs and outputs perfectly, so that there are not unwanted reflections. Control of, and reproducibility of operating thresholds would be non-trivial. Furthermore, anyone who has looked a conductance fluctuations from real structures in wires, with or without crossings, is aware that "accidental" resonances are very large.

The general conclusions then are that: (1) it will be hard to make good interference structures because of doping fluctuations using a MODFET-based technology, and (2) it is not clear that any devices so far proposed have a foreseeable use. Nonetheless, they do provide a technological challange andmay lead to some more interesting structures and new insights.

REFERENCES

Berggren, K. F., Thornton, T. J., Newson, D. J., and Pepper, M., 1986, Phys. Rev. Letters, 57:1769.

Büttiker, M., 1985, in "SQUID '85", Ed. by Hahlbohm, H. D., and Lubbig, H., Walter de Gruyter and Co., Berlin.

Datta, S., Melloch, M. R., Bandyopadhyay, S., Noren, R., Vaziri, M., Miller, M., and Reifenberger, T., Phys. Rev. Letters, 55:2344.

Davies, J.H., and Nixon, J. A., 1989, Phys. Rev. B, 39:3423.

Ford, C. J. B., Fowler, A. B., Hong, J. M., Knoedler, C. M., Laux, S. E., Wainer, J, and Washburn, S., 1990, Surf. Science, in press.

Fowler, A. B., 1984, U. S. Patent 4,550,330, granted 1985.

Fowler, A. B., and Harstein, A. M., 1985, U. S. Patent 4,672,423, granted 1987.

Kotthaus, J. P., Hansen, W., Pohlmann, H., Wassermeier, M., and Ploog, K., 1988, Surf. Science, 196:600.

Laux, S. E., and Stern, F., 1986, Appl. Phys. Letters, 49:91.

Laux, S. E., and Warren, A. C., 1986, in "Proceedings of the International Electron Device Meeting," 86:567.

Nixon, J. A., and Davies, J.H., 1990, to be published.

Shik, A. Ya., 1985, Phys. Tekh. Poluprovoon, 19:1488 [transl. in Sov. Phys. Semicond., 19:915].

Smith, T. P., III, Arnot, H., Hong, J. M., Knoedler, C. M., Laux, S. E., and Schmid, H., 1987, Phys. Rev. Letters, 59:2802.

Sols, F., Macucci, M., Ravaioli, U., and Hess, K., 1989, Appl. Phys. Letters, 54:350.

Warren, A. C., Antoniadis, D. A., and Smith, H. I., 1986, Phys. Rev. Letters, 59:732.

Washburn, S., Fowler, A. B., Schmid, H., and Kern, D., 1988, Phys. Rev. B, xx:xxx.

DEPHASING AND NON-DEPHASING COLLISIONS IN NANOSTRUCTURES

A. J. Leggett

Department of Physics
University of Illinois
Urbana, Illinois 61801

8

INTRODUCTION

When I was originally asked to give two lectures at this workshop, one on this subject and one on "charge quantization" effects, I felt that there was at best a tenuous link between the two topics. Now I feel that there is a fundamental connection, and that it is absolutely essential to understand what we mean by "dephasing" before we can assess many of the results obtained in the last few years on charge quantization and related effects.

As you know, one of the major advances in mesoscopic physics in the last decade has been the realization that the observation of single-electron interference effects in mesoscopic systems is much less difficult than had naively been thought. To see why there had been thought to be a major problem, let us make a few order-of-magnitude estimates. Suppose for example that we are thinking of trying to observe an effect analogous to the superconducting Meissner effect in a small ring, that is, a nontrivial dependence of the free energy on flux through the ring. The effect only arises because the electron wave function must satisfy the "single-valuedness boundary condition" $\Psi(0) = \Psi(2\pi)$ (where the argument of $\Psi \equiv |\Psi|e^{i\varphi}$ is the angle travelled around the ring); in particular, the *phase* of the wave function must change by exactly $2n\pi$ as we travel once around:

$$\varphi(2\pi) = \varphi(0) + 2n\pi \ . \tag{1}$$

Now if a small magnetic flux is applied through the ring, the expression for the electric current contains a term which depends on the corresponding vector potential:

$$\mathbf{j} = \frac{\hbar}{m}|\Psi|^2 [\nabla\varphi - \frac{e}{\hbar}\mathbf{A}(\mathbf{r})] \tag{2}$$

and the energy is similarly

$$E = \int \frac{\hbar^2}{2m}|\Psi|^2 [\nabla\varphi - \frac{e}{\hbar}\mathbf{A}(\mathbf{r})]^2 \, d\mathbf{r} \ . \tag{3}$$

If we introduce the standard gauge transformation

$$\Psi \rightarrow \Psi' \equiv \Psi \exp\left[\frac{ie}{\hbar}\int \mathbf{A}(\mathbf{r})\cdot d\mathbf{r}\right] \tag{4}$$

and express the current and energy in terms of Ψ', then ($\Psi' \equiv |\Psi| \exp i\varphi'$)

$$\mathbf{j} = \frac{\hbar}{m}|\Psi|^2\nabla\varphi' \ , \quad E = \frac{\hbar^2}{2m}\int |\Psi|^2 (\nabla\varphi')^2 \, d\mathbf{r} \ ; \tag{5}$$

i.e., the same expression in terms of φ' as we had in terms of φ for $\mathbf{A} = 0$. The only effect of the vector potential is to change the single-valuedness boundary condition (SVBC): expressed in terms of φ', it reads

$$\varphi'(2\pi) = \varphi'(0) + 2n\pi + \frac{e}{\hbar}\int_C \mathbf{A}(\mathbf{r})\cdot d\mathbf{r}$$

$$\equiv \varphi'(0) + 2n\pi + \frac{e}{\hbar}\Phi \ . \tag{6}$$

Here Φ is the total flux through the ring (we assume for simplicity that $\mathbf{A}(\mathbf{r})$ does not vary over the thickness of the ring). The effect we are looking for will arise if the total free energy depends nontrivially on Φ; since it is easy to show that the expectation value of the current I around the ring is $\partial F/\partial\Phi$, I will then be finite except for special values of Φ. Except for a few rather pathological cases, the free energy is stationary at $\Phi = 0$ and does not decrease for finite Φ (though see below); thus the only question is whether or not the system can adapt its state to make $F(\Phi)$ equal to $F(0)$. If it cannot, we have a "Meissner-like" effect.

Now in general the system has two ways of doing this: either it can deform the original wave functions, or it can change the occupation of these states. For a noninteracting system the second effect is easily calculated, (cf. e.g. Imry, 1986) and it can be shown that the difference between $F(\Phi)$ and $F(0)$ vanishes exponentially as soon as $kT \geq \Delta E$, where ΔE is a typical spacing of the one-particle levels corresponding to different "winding members" [i.e. different values of n in (1)]: for a single electron moving in a potential which is not too different from a constant ΔE may be estimated as $\sim\hbar^2/mR^2$, while for a degenerate Fermi sea this is multiplied by a factor $\sim k_F R$ (R = radius of ring, k_F = (order of magnitude of) Fermi wave vector); in other words, for the effect to be substantial R must be no greater than the thermal de Broglie wavelength $\lambda_D \sim (\hbar^2)/mkT$). At (say) 10mK this requires a radius of less than ~3000 Å.

Now the problem is (or rather was thought to be) that it was, at least until quite recently, rather difficult to make rings of such small radius without introducing a fairly substantial concentration of chemical and other impurities. In fact, if one were to estimate a typical number of such impurities and hence obtain a mean free path l by the standard kinetic-theory arguments, it turns out that typically we would have $l \ll R$. It was then tempting to argue that if the electron has been scattered many times as it goes once around the ring, it cannot retain any "memory" of its phase, and hence the single-valuedness boundary condition becomes inoperative. In that case one can immediately choose $\varphi'(A) \equiv \varphi(0)$, and automatically get $F(\Phi) = F(0)$.

In retrospect, of course, that argument looks naive. When experiments started to see effects [such as flux-dependence of the resistance (Umbach *et al.*, 1984)—the actual "Meissner" type effect discussed above has been seen only very recently (Levy *et al.*, 1990)] which indicated that the single-valuedness boundary condition is still relevant even under the condition $l \ll R$, it was quickly realized that (as had in fact been predicted earlier) (Büttiker *et al.*, 1983), the mere fact that an electron undergoes repeated scattering does not necessarily mean that all memory of the phase of the wave function is lost. In fact, if one thinks about the actual appearance of the wave function in a ring, be it one- or three-dimensional, containing a random distribution of relatively weak or dilute scattering centers, then it is clear that unless the ring is crossed by a nodal surface (that is, a surface on which the *amplitude* of the wave function goes to zero), then the "winding number", that is, the number of factors of 2π by which the phase changes as we go once around the ring, must be well-defined, and because of the SVBC must be integral. Except for very special values of the external flux (cf. below) it is usually energetically very unfavorable to put in a nodal surface, so the phase of the wave function is indeed well-defined.

Does this mean, then, that <u>all</u> scattering of the electron is irrelevant to the observation of such phase-coherence-dependent effects? Certainly not: it is well established experimentally that a sufficient condition for such effects to vanish is that the circumference of the ring be much greater than the mean free path l_{ph} against scattering of the electron by <u>phonons</u>. Here it should be emphasized that we are talking about the mean free paths which the electron traverses without being scattered at all, irrespective of the direction of scattering. For a reasonably clean

material this is proportional to T^{-3}. By contrast, the "transport mean free path" which enters Drude-type theories of the electrical and thermal conductivity weights the scattering events with the factor by which they reduce the electrical current, namely $(1 - \cos\theta)$, and hence, as is well known (see e.g. Ashcroft and Mermin, 1976) is proportional to T^{-5} at low temperatures, giving the famous Bloch-Gruneisen law. The transport mean free path is not relevant in the present context.

Thus, from an experimental point of view there seems no doubt that collisions with static impurities are "non-dephasing" while collisions with phonons are "dephasing", and in the theory it is now conventional to introduce an "elastic" (or non-dephasing: see below) mean free path l and to contrast it with a "phase breaking" (dephasing) mean free path l_ϕ; one then assumes that the phase coherence of the electron wave function is preserved so long as the path travelled is no less than $\sim l_\phi$. Note that this does not mean that l is irrelevant to phase coherence effects: if it is much shorter than l_ϕ, as will always be the case at sufficiently low temperatures, and is also short compared to relevant distances such as the circumference of the ring, then the motion of the electron (viewed semiclassically) will be diffusive in nature, and the actual path length it travels to get around the ring will not be $2\pi R$ but rather $3(2\pi R)^2/l$, and this should be $\lesssim l_\phi$, so that the condition for phase coherence to persist is the stronger one

$$R \lesssim l_0 \equiv \sqrt{\frac{l l_\varphi}{2}} \ . \qquad (7)$$

PHASE INTERFERENCE

All the above is, of course, by now standard stuff, and can be found in much more detail in various review papers (e.g. Imry, 1986 `). Moreover, there seems little doubt that theoretical calculations based explicitly or implicitly on these ideas will give the right answer. But now let's raise the question: What *exactly* is it that makes collisions with phonons "dephasing" but collisions with static impurities "non-dephasing"? Moreover, how should we generalize these ideas to other types of scattering such as spin-orbit scattering or scattering by localized magnetic moments, other electrons or (in principle) nuclear spins? In particular, is it possible for a particular type of scattering, e.g. by localized magnetic moments, to change its nature from phase-breaking to non-phase-breaking as some external parameter such as the Zeeman field acting on it is varied?

The most obvious difference between collisions with static impurities and collisions with phonons is, of course, that in the second case, but not the first, the electron loses or gains energy, i.e. collisions with impurities are elastic and those with phonons inelastic. There is therefore a temptation to identify "dephasing" with "inelastic" and "non-dephasing" with "inelastic". However, it should be emphasized that the mere fact that a system gains or loses energy in a particular type of interaction or collision process does not in itself mean that it loses all memory of its phase. A beautiful illustration of this point can be found in some experiments conducted in recent years on the neutron interferometer (Badurek *et al.*, 1987). In these experiments a beam of neutrons originally with spin (say y) up is split into two beams which are well separated spatially, and one beam then passes through a cavity which is subject to a strong d.c. external magnetic field and also a weak r.f. field which satisfies the Larmor resonance condition. The strength of the field, and the velocity of the neutrons (hence the time spent in the cavity) is adjusted so that a neutron which enters with spin up emerges with spin down Because the neutron has absorbed a net energy $2\mu H$ from the r.f. field during its passage through the cavity, its kinetic energy after emerging is increased by this amount over the original K.E. it had before entering, and this has been directly verified, for a single beam, in an independent experiment (Alefeld, 1981). Nevertheless, when the two beams are brought back together and the transverse component of spin is measured using stroboscopic detection (this is necessary since, because one beam has shifted its energy in relation to the other, the interference pattern shows beats in time) then the effect of interference of the two beams is clearly and spectacularly seen.

If one treats the r.f. field classically from the start and works entirely in terms of the neutron wave function, then it is tempting to say that the reason that interference can still be seen is that, while the phase of the branch of the wave function corresponding to the neutrons in the upper beam has been changed relative to that of the lower beam, the change is controlled and can be calculated to arbitrary accuracy from a knowledge of the time-dependence of the r.f. field. On the other hand, one would argue that in a typical electron-phonon collision process in a metal, one does not know the initial phase of the phonon field, and therefore the change in phase, though perhaps in some sense definite for each individual event (this may be a question of quantum-mechanical theology!) is nevertheless unknown and randomly distributed. Hence it is fair to characterize such collisions as phase-breaking.

I believe that this picture, while perhaps heuristically helpful in cases like the neutron interferometer, may nevertheless be more misleading than helpful in the case of present interest. The fact is that while in the neutron case the r.f. field, which in the experiment is typically of amplitude very large compared to the zero-point photon amplitudes, may be legitimately regarded as an essentially classical object, in the case of the metal the phonon field is typically only weakly excited relative to the zero-point motion and is by no means semiclassical. A further important difference, which is actually closely related to the above one, is that in the neutron case the r.f. field is dictated by the experimenter and the reaction of the neutron on it is negligible, while the effect of the electron on the phonon field may be very important (as is seen e.g. in treatments of the polaron problem).

A conceptually much more satisfactory approach is to treat both the electron (neutron) and the system with which it interacts as quantum mechanical, and to take the classical limit, if at all, only at the end of the calculation. How does the interferometer experiment look in this picture? Let us think in terms of a wave packet of finite length. Initially, before the upper beam had entered the r.f. cavity, the wave function of the neutron was schematically of the form

$$\Psi = a\Psi_A + b\Psi_B \ , \qquad (8)$$

where A, B correspond to the upper and lower beams respectively and Ψ_A and Ψ_B are time-dependent, while the rf field is in some (also time-dependent) state which we schematically call χ_{in}. Thus the total wave function of the "universe" is schematically of the form

$$\Psi = (a\Psi_A + b\Psi_B)\,\chi_{in} \ . \qquad (9)$$

If at this stage some measurement on the neutron is performed to detect the coherence between the two branches (e.g. by diverting the upper beam so that it does not pass through the r.f. cavity on its way to the screen where the interference is detected) then since the operator in question (call it $\hat{\Omega}$) refers to the neutron alone, it is a unit operator in the space of the rf field, so that its expectation value is

$$\langle\Psi,\hat{\Omega}\Psi\rangle \equiv \left\{(a\Psi_A + b\Psi_B)\cdot\chi_{in} \mid \hat{\Omega} \mid (a\Psi_A + b\Psi_B)\chi_{in}\right\}$$

$$= \left\{(a\Psi_A + b\Psi_B) \mid \hat{\Omega} \mid (a\Psi_A + b\Psi_B)\right\}(\chi_{in},\hat{1}\chi_{in})$$

$$= \left\{(a\Psi_A + b\Psi_B) \mid \hat{\Omega} \mid (a\Psi_A + b\Psi_B)\right\}$$

$$= |a|^2(\Psi_A,\hat{\Omega}\Psi_A) + |b|^2(\Psi_B,\hat{\Omega}\Psi_B) + 2\,\mathrm{Re}\,\{a^*b\,(\Psi_A|\hat{\Omega}|\Psi_B)\}\,. \qquad (10)$$

and it is clear that the behavior of the r.f. field is totally irrelevant, as we expect on physical grounds. Note that it is the last term which contains all the effects of interference. Now consider what happens when the packet representing the upper beam passes through the r.f. cavity. Thereafter, the wave function of the "universe" is schematically of the form

$$\Psi = a\Psi_A\chi_f + b\Psi_B\chi_{in} \qquad (11)$$

where χ_f is the final state of the radiation field; here we took into account that the interaction Hamiltonian by construction cannot induce an interaction between the lower beam and the field. Now if we evaluate the expectation value of $\hat{\Omega}$, which is of course still proportional to the unit matrix in the Hilbert space of the r.f. field, we find:

$$\langle\Omega\rangle \equiv \left\{a\Psi_A\chi_f + b\Psi_B\chi_{in} \mid \hat{\Omega} \mid a\Psi_A\chi_f + b\Psi_B\chi_{in}\right\}$$

$$= |a|^2(\Psi_A,\hat{\Omega}\Psi_A) + |b|^2(\Psi_B,\hat{\Omega}\Psi_B) + 2\,\mathrm{Re}\{a^*b(\chi_f|\chi_{in})(\Psi_A,\hat{\Omega}\Psi_B)\}. \qquad (12)$$

We see that the last term, which expresses the effects of interference, is multiplied by a factor $(\chi_f|\chi_{in})$ proportional to the overlap in Hilbert space of the initial and final states of the radiation field. Now the crucial point is that precisely because the r.f. field is in a highly coherent state, χ_f is essentially identical to χ_{in} <u>despite</u> the fact that an r.f. photon must have been absorbed to supply the energy $2\mu H = \hbar\omega$ necessary to tip over the neutron spin. (We recall that coherent states of the rf field, which show the best quantum-mechanical approximation to classical behavior, are by definition eigenstates of the annihilation operator and near-eigenstates (to order $N^{-1/2}$, N= av. no. of photons) of the creation operator). Were we repeat the experiment using not a coherent r.f. field but rather a thermal (blackbody) radiation field to effect the tipping of the spin (assuming this were experimentally feasible) the situation would be quite different: now the initial state of the field corresponds to a definite number of photons, and the final state is one less, so that one can predict with high confidence that in such a set up the interference pattern (as observed in a measurement of $\langle\sigma_x\rangle$) would vanish.

In the light of this it is very tempting to make the following claim: The distinction between dephasing and non-dephasing collisions lies in whether or not the final wave function of whatever is being collided with is or is not orthogonal to the original wave function. Clearly, since there are varying degrees of orthogonality, this allows a continuum corresponding to differing degrees of dephasing. This point of view has a very long history in discussions of the quantum theory of measurement, where it is usually invoked in particular to demonstrate that after a particle has interacted with a measuring apparatus no observation of phase coherence between the different "branches" which are distinguished by the measurement is possible. The point here is that if the superposition is say of two states Ψ_A and Ψ_B, then in order for us to be able to read off from the final state of the measuring apparatus whether A or B was realized, the final states of the measuring apparatus corresponding to these two possibilities must be mutually orthogonal (indeed, more than that, they must be macroscopically distinguishable), so that $(\chi_A|\chi_B)$ is zero to a very high degree of approximation and all interference effects vanish. In the context of mesoscopic systems, a similar approach has been advocated in a very recent paper by Stern *et al.* (1990) who claim inter alia to be able to derive from it results on the effect of electron-electron interactions which have previously been derived much more tediously by Green's function and similar methods. It should be mentioned that even more recent work by Loss and Mullen (1990) raises serious doubts about the technical validity of some of the results of the former work even in the context to which they are explicitly applied; however here I want to approach the question from a slightly different angle.

If one takes the above point of view, then at first sight the reason for the distinction between collisions with static impurities and with phonons becomes very obvious: in the case of scattering by static impurities the scattering potential is effectively a c-number, so that there is no quantum "environment" to be changed and hence no dephasing. In any interaction involving emission of phonons, on the other hand, the final state of the phonon field is (at least at first sight) automatically orthogonal to the initial state, so that the condition for complete destruction of phase coherence is fulfilled.

At first sight this gives a nice unifying view. However, I believe it runs into a number of difficulties, some possibly trivial but at least one more serious. First, it is unclear how to apply the criterion to some other kinds of scattering. For example, in the case of spin-orbit scattering the electron spin as well as its orbital motion is changed: should one regard the spin degree of freedom as simply a part of the electron wave function, or as a separate degree of freedom and hence part of the "environment"? Secondly, in any reasonable experiment it is actually only <u>one</u> translational degree of freedom of the electron which is of prime interest in the context of phase coherence, typically, in the case of a ring, the degree of freedom corresponding to motion <u>around</u> the ring. The other two ("transverse") degrees of freedom are averaged in the final result. For example, if we consider the "Meissner" effect in a ring of width small compared to its radius, confine the flux for simplicity to the hole and introduce a coordinate z corresponding to motion around the ring and two "transverse" coordinates x and y, then the free energy F as a function of flux Φ can be written in functional-integral form as

$$F(\Phi) = -kT \ln Z(\Phi)\,, \qquad (13)$$

$$Z(\Phi) = \int_0^{2\pi\Phi} dz \int dx \int dy \int_{z_i=z}^{z_f=z+2\pi\Phi} Dz(\tau) \int_{x_i=x}^{x_f=x} Dx(\tau) \int_{y_i=y}^{y_f=y} Dy(\tau) \exp\left[-\int_0^{\beta} \frac{L(\tau)}{\hbar}\,d\tau\right], \qquad (14)$$

$$L(\tau) \equiv \frac{m}{2}(\dot{x}^2 + \dot{y}^2 + \dot{z}^2) - \frac{\dot{z}\Phi}{2\pi R} + V(x,y,z)\,,\ [x \equiv x(\tau), \text{etc.}]\,, \qquad (15)$$

where the integrals over x and y run over the appropriate cross-section of the ring. It is clear from the form of L that the "transverse" coordinates enter the problem only to be integrated over; they do not couple directly to Φ. Thus, it is natural to regard them as part of the "environment". But now what of the phonons? After all, from the point of view of quantum mechanics they are simply extra degrees of freedom, which enter the problem on exactly the same footing as x and y; they do not couple to the flux, and are averaged over in the final expression. Now collisions with static impurities mix the z, x and y degrees of freedom, while electron-phonon collisions mix z, x and y with one another and with the phonon degrees of freedom: what is the fundamental difference? The obvious possibilities which spring to mind are either (a) that there are only 2 transverse degrees of freedom but a great many phonon ones, or (b) that the phonons in practice have a more or less continuous energy spectrum (i.e. a typical $\hbar\omega$ may be small compared to all energies of interest, and moreover is continuously varying) while the transverse motion in a narrow ring is strongly quantized. This clearly needs further consideration.

However, there is, I believe, a much more fundamental difficulty: An essential assumption which is implicit in the above approach to the problem of dephasing is that we can make "measurements" only of operators which are unit operators with respect to the phonon field (or other object collided with). (More precisely, that we cannot measure operators which are nondiagonal with respect to *both* the electron and the phonon field.) Now at first sight this seems obviously true; for all our normal "measurable quantities" such as charge density electric current density etc. indeed refer only to the electron. (To avoid confusion as regards the case of electron-electron interactions, it should be mentioned that in this case what we measure is the *sum* of operators referring to "different" electrons, not the product: so the statement remains true). However, there is a very subtle point involved here, to introduce which I discuss for a moment to a rather simpler situation, namely the neutron interferometer *without* any spin-flipping field (or equivalently the classic Young's-slits experiment). Imagine a neutron wave packet passing through the apparatus: at some time, call it t_0, the neutron is approximately represented by the coherent superposition of two wave packets separated by a distance of the order of several cm. Now, we said that to observe the coherence we had to find some operator $\hat{\Omega}$ which has nonzero matrix elements between Ψ_A and Ψ_B. But it is easy to convince oneself that no simple operator such as x, p or any simple low-order combination of them has this property; and while there indeed formally exist operators which do (such as the operator $\exp(ip \cdot R_{12}/\hbar)$, where R_{12} is the separation between the centers of mass of the two packets) it is clear that in practice we do not have the means of measuring such a thing. So how come that we can ever see coherence between the two beams? The answer is, of course, that we do *not* make the measurement at time t_0. Rather, we wait until some later time t_f where the two wave packets have again approached one another in the neighborhood of the final (detecting) screen In other words, if we use the Heisenberg representation, we are measuring not $\hat{\Omega}(t_0)$ but rather $\hat{\Omega}(t_f)$, which is given by

$$\hat{\Omega}(t_f) \equiv \exp[i\hat{H}(t_f - t_0)]\, \hat{\Omega}(t_0) \exp[-i\hat{H}(t_f - t_0)] \,. \tag{16}$$

This operator may and indeed does (since for a free particle $\hat{H}$ is $\hat{p}^2/2m$) contain very high powers of p, and so can (and does) have nonzero matrix elements between Ψ_A and Ψ_B.

Now let us apply a similar argument to the case of interest. Suppose the system of interest (e.g. electron) has two possible states of interest, Ψ_A and Ψ_B, and that at a certain time t_0, as a consequences of the interaction with some "environment", say the phonons, the wave function of the "universe" (system plus environment) is of the form

$$\Psi = \psi_A\chi_A + \psi_B\chi_B \,, \tag{17}$$

where χ_A and χ_B are to a good approximation mutually orthogonal. Again, if we make a measurement of some purely electronic operator at time t_0 we will see no effects of interference between Ψ_A and Ψ_B. However, nothing guarantees us that we will not subsequently be able to recover the interference, and indeed this will be so unless the quantity

$$\left\langle \Omega(t_f) \right\rangle = \left\langle \psi_A\chi_A + \psi_B\chi_B \,|\, \exp[i\hat{H}(t_f - t_0)]\hat{\Omega}(t_0)\exp[-i\hat{H}(t_f - t_0)] \,|\, \psi_A\chi_A + \psi_B\chi_B \right\rangle \tag{18}$$

vanishes.

It is clear that this will be the case if $\hat{H}$ is "switched off" after time t_0, and this, of course, is what makes the above criterion for dephasing a reasonable one in cases like the neutron interferometer. However, if $\hat{H}$ contains any interaction at all between the system and the environment, we can by no means exclude the possibility that the phase coherence can be recovered. This feature is spectacularly illustrated in the familiar "spin-boson" problem described by the Hamiltonian

$$H = \Delta\sigma_x + \sum_k \left(\frac{m_k}{2}\omega_k^2 x_k^2 + \frac{p_k^2}{2m_k} \right) + \sigma_x \sum_k C_k x_k \tag{19}$$

where we assume for the moment that all the oscillator energies $\hbar\omega_k$ are large compared to Δ. In view of this condition, the adiabatic approximation is valid to high accuracy, and the lower two energy eigenstates are of the approximate form

$$\Psi_\pm = \frac{1}{\sqrt{2}} \left\{ |\uparrow\rangle |\chi_\uparrow\rangle \pm |\downarrow\rangle |\chi_\downarrow\rangle \right\}, \tag{20}$$

where $|\uparrow\rangle$, $|\downarrow\rangle$ are the states $\sigma_2 = \pm$, and $\chi_\uparrow$, $\chi_\downarrow$ are the "shifted" oscillator states, that is, the groundstates of the Hamiltonians

$$H_{\uparrow,\downarrow}(x_k, p_k) = \frac{m_k}{2}\omega_k^2 x_k^2 + \frac{p_k^2}{2m_k} \pm C_k x_k \,. \tag{21}$$

For a reasonable degree of coupling the overlap $(\chi_\uparrow, \chi_\downarrow)$ is exponentially small. If now we start the system at time zero in (say) the state $\sigma_z = +1$, then after a certain time t_0 the wave function of the "universe" will be approximately of the form

$$\Psi = \frac{1}{\sqrt{2}} \left\{ a\,|\uparrow\rangle |\chi_\uparrow\rangle + b\,|\downarrow\rangle |\chi_\downarrow\rangle \right\}, \tag{22}$$

and any operator $\hat{\Omega}$ referring to the system above will have an expectation value

$$\left\langle \hat{\Omega}(t_0) \right\rangle = |a|^2 \left\langle \uparrow|\hat{\Omega}|\uparrow \right\rangle + |b|^2 \left\langle \downarrow|\hat{\Omega}|\downarrow \right\rangle + O(\chi_\uparrow|\chi_\downarrow) \tag{23}$$

so that the interference term is exponentially small. Now at this point we are tempted to throw away this term and argue that we may as well replace the superposition description by an incoherent mixture of $|\uparrow\rangle$ and $|\downarrow\rangle$ with equal weights. If this were legitimate, then we would see no further interesting behavior, in particular no characteristically quantum-mechanical oscillations of the probability density between the states $|\uparrow\rangle$ and $|\downarrow\rangle$. In reality, however, if we wait for a further time t_0, we shall certainly find the system in the state $|\downarrow\rangle$. Thus even exponentially small interference terms may return to haunt us! The situation is made a little less amazing by the observation that the time t_0 is itself of order $(\chi_\uparrow, \chi_\downarrow)^{-1}$, but the point remains. If now we allow the oscillator spectrum to extend down to zero frequency, we will find that the coherence is indeed gradually destroyed, but over a time which corresponds roughly to the lifetime of the upper state (corresponding to the - sign in eqn. (20)) against decay into the lower—which is generally speaking itself proportional to $(\chi_\uparrow, \chi_\downarrow)^{-n}$, $n \geq 1$, [for an exhaustive discussion of this problem, see Leggett *et al.* (1987): in the language of that reference, the above discussion implicitly assumes the oscillator distribution is either "superohmic" or ohmic with $\alpha << 1$]. The crucial point is that one must distinguish carefully between these effects of interaction with the environment which are "adiabatic" in nature and which, therefore, allow one to restore phase coherence by waiting long enough, and those which are genuinely dissipative: only the latter contribute to true dephasing.

Are we then back full circle to our original hypothesis that the crucial distinction is between those processes in which the electron can *irreversibly* lose energy and those in which it cannot? Possibly, but here is a worry here: The above argument implicitly refers to time-dependent behavior, whereas most of the practically important experiments on mesoscopic

systems measure *static* properties. Thus the implicitly semiclassical picture, in which we visualize the electron travelling around the ring in time and making "collisions" en route, may itself be on dubious ground; in any case it would clearly be desirable to have a picture in which time does not enter.

It is time, therefore, to discuss, at least schematically, the actual energy eigenstates of the system. Let's start with the simple and by now relatively well-explored case of a single electron moving on a strictly one-dimensional ring in the presence of some static "random" potential $V(\theta)$. The quantity of interest is the dependence of energy E on the magnetic flux Φ applied through the ring (and hence of the current I circulating in the ring since by a well-known theorem we have $I = -\partial E/\partial\Phi$). As usual, we remove the flux from the Hamiltonian by the gauge transformation

$$\psi \rightarrow \psi' \equiv \psi \exp[\, i\int A\cdot dr \;;$$

the Hamiltonian is then that for a particle moving in the same potential and zero flux, but with a changed boundary condition: the condition $\psi(2\pi) = \psi(0)$ is replaced by $\psi(2\pi) = \psi(0)\exp i\Phi/\Phi_0$, where $\Phi_0 \equiv h/e$ is the "single-particle" flux quantum (to be distinguished from the "Cooper-pair" flux quantum $h/2e$ conventionally used in the theory of superconductivity). It is clear that the energy levels are (a) periodic in Φ with period Φ_0 (note that this does not exclude the possibility of periodicity with a period which is an integral multiple of Φ_0) (b) even functions of Φ (provided the potential terms in the Hamiltonian satisfy the condition of time-reversal invariance). It is also clear that for the special values of flux $\Phi = n\Phi_0/2$, n integral, the wave function ψ is real (or, in the case of degeneracy, can be chosen to be so). The simplest structure of the energy levels which satisfy these conditions is the familiar "band" structure shown in fig. 2: more generally, we could in principle have extra maxima or minima between the ones which are required by the above condition, e.g. the arrangement of fig. 3 is not excluded a priori. It is convenient to visualize this in the spirit of the well-known weak-coupling model of band structure in solids, although it should be emphasized that the role of the independent variable Φ, here an externally controlled c-number, is conceptually completely different from that in the cyrstalline problem (where it is k, the crystal momentum, which is a dynamical variable). For completely free particles ($V(\theta) = 0$) we should have the trivial band structure shown in fig. 4 : At the points $\Phi = (n + 1/2)\Phi_0$ two bands are exactly degenerate and can be split by an arbitrarily small potential $V(0)$ into even and odd combinations: note that the matrix element doing the splitting is simply the lowest nontrivial Fourier component

$$V_1 \equiv \int V(\theta)\, e^{i\theta}\, d\theta \,. \tag{24}$$

Suppose we have N spinless fermions on the ring. Then, assuming that the groundstate is nondegenerate and the bands therefore correspond qualitatively to the form shown in fig. 2, we see that the total energy as a function of Φ is likely to be dominated by the behavior of the last filled "band," lower bands tending to cancel one another out. Thus for N odd we should expect the free energy to have a minimum at $\Phi = 0$ and the system to show diamagnetism, whereas for N even we expect a *maximum* at $\Phi = 0$ and the system should be paramagnetic. If we average the current for given Φ over a large number of rings we expect that with experimentally reasonable control of the parameters N should vary at random from even to odd: the total current would therefore be proportional to $\sqrt{N}$.

If the transverse dimensions of the ring are finite, things are a bit more complicated. The single-electron energies as a function of flux must of course still have the general structure shown in fig. 2, but there will be in general many sub-bands ("channels") corresponding to different possible "transverse" behavior. Since the wave functions corresponding to these different bands are, in the absence of disorder rigorously orthogonal for given Φ, there is in principle nothing to prevent them crossing; and if we then imagine some small perturbation (e.g. due to disorder) which mixes the crossing levels, we get the possibility of a local *minimum* in the energy of a given state at half-integral as well as integral values of Φ. Whether or not this happens should depend crucially on the transverse dimensions, and on the degree of disorder. If it does happen, then clearly we should expect, in a random ensemble of such rings, a net diamagnetism proportional to N rather than to $N^{1/2}$.

Since this is precisely the behavior seen in the recent experiment of Levy *et al.*(1990) it is of interest to inquire how far the above considerations may be generalized to the case of interacting electrons. We will first establish that for spinless electrons with arbitrary electron-electron interactions and an arbitrary external potential the conclusions reached above for the noninteracting case remain at least partially true in the sense that a system of an odd number of particles must be diamagnetic (or at best nonmagnetic) around $\Phi = n\Phi_0$ and a system with an even number must be diamagnetic (etc.) around $\Phi = (n + 1/2)\Phi_0$.

Consider then the Hamiltonian

$$H(\theta_1 \ldots \theta_n) = \sum_i V(\theta_i) + \sum_{ij} U(\theta_i - \theta_j) + \sum_i [p_i - eA(\theta_i)]^2 \frac{1}{2m} \,, \tag{25}$$

where the functions U and V are arbitrary except that they must, of course, be periodic in each θ_i with period 2π. The gauge-transformed wave function $\Psi'(\theta_i \ldots \theta_N)$ must satisfy the two conditions

$$\text{(a)}\ \Psi'(\theta_1 \ldots \theta_i \ldots \theta_j \ldots \theta_N) = -\Psi'(\theta_1 \ldots \theta_i \ldots \theta_j \ldots \theta_N)\,,\ \forall i,j \ \text{(anti-symmetry)}$$

$$\text{(b)}\ \Psi'(\theta_1 \ldots \theta_i \ldots \theta_N) = \Psi'(\theta_1 \ldots \theta_i + 2\pi \ldots \theta_N)\exp(i\Phi/\Phi_0)\,,\ \forall i \ \text{(single-valuedness)}$$

and, of course, Schrödinger's equation with A in the Hamiltonian (25) set to zero. It is clear that for $\Phi = n\Phi_0$ or $(n + 1/2)\Phi_0$ the wave function Ψ' is real (or can be chosen so); for other values of Φ it will in general be complex.

Imagine that we know the wave function Ψ (from now on we omit the prime, but we always mean the gauge-transformed-wave function) for $\Phi = 0$. Consider now the case of finite nonzero Φ. We have to solve the *same* Schrodinger equation as for $\Phi = 0$, but with a different boundary condition, which requires a finite change of phase as we go around the ring rather than zero. One obvious variational possibility is the wave function

$$\Psi_{var}(\Phi{:}\theta_1 \ldots \theta_N) = \exp[i\chi(\theta_1 \ldots \theta_N)]\Psi(0{:}\theta_1 \ldots \theta_N) \tag{26}$$

where χ is a totally symmetric function of its indices and satisfies the conditions

$$\chi(0{:}\theta_1 \ldots \theta_N) = 0\,, \qquad \chi(2\pi{:}\theta_1 \ldots \theta_N) = \frac{\Phi}{\Phi_0} \,. \tag{27}$$

The one- and two-particle potential energies, which depend only on $|\Psi|^2$, are of course the same for (27) as for $\Psi(0)$. The kinetic energy of (27) is however greater by an amount

$$\Delta K = N \int_0^{2\pi} \left|\Psi(0{:}\theta_1 \ldots \theta_N)\right|^2 \left|\nabla_1 \chi(\theta_1 \ldots \theta_N)\right|^2 d\theta_1 \,. \tag{28}$$

In general, ΔK is positive definite. However, if it is possible to find a surface on which $|\Psi(0)|$ is zero (i.e. a nodal surface) *across which the change of χ is not determined by the symmetry,*

(i.e. which does not simply correspond to the point $\theta_i = \theta_j$ where j is some value 2 ... N) then we can simply set χ equal to zero on one side of this surface and equal to Φ/Φ_0 on the other, and this will cast us no kinetic energy ($\Delta K = 0$). We call any such surface a "non-symmetry-dictated nodal surface" (NSDNS) to distinguish it from the "symmetry-dictated nodal surfaces" corresponding to $\theta_i = \theta_j$, $i \neq j$.

Consider now the behavior of the original real wave function $\Phi(0)$ as we take θ_1 from zero to 2π, keeping $(\theta_2 \ldots \theta_N)$ constant. As we go around, we must cross N-1 symmetry-dictated nodal surfaces corresponding to $\theta_i = \theta_j \ldots \theta_N$, and on each of these the wave function change sign. On the other hand, we know that we must arrive at 2π with a + sign, if we are to respect the single-valuedness condition. Thus, for N *even*, we must cross at least one NSDNS, and then by the above variational argument we have

$$E(\Phi) \leq E(0), \quad \forall\Phi \ (\text{N even}), \tag{29}$$

i.e. $\Phi=0$ must be an absolute maximum, or at least a neutral point, of the energy. By an exactly similar argument, for *odd* N $\Phi = \Phi_0/2$ is a maximum:

$$E(\Phi) \leq E(\Phi_0/2), \quad (\text{N odd}). \tag{30}$$

In addition, $E(\Phi)$ must of course be even and periodic in Φ with period Φ_0, just as in the noninteracting case.

The most "natural" inference from this result is that for N=odd the energy has minima at $\Phi=n\Phi_0$ and maxima at $\Phi = (n + 1/2)\Phi_0$ and no other extrema, and similarly that for N = even we have *maxima* at $\Phi=n\Phi_0$ and minima at $\Phi = (n + 1/2)\Phi_0$, all just as for the noninteracting case. However, it should be emphasized that the above argument cannot exclude the possibility that for N = even or N = odd or both we have maxima at *both* integral and half-integral values of Φ/Φ_0, possibly of different heights. Note that if this were true for N both even and odd, we should produce a moment in a system of N_I such rings which is proportional to N_I but, in contrast to the mechanism envisaged above, *paramagnetic* in sign.

The extension of these results to the realistic three-dimensional case seems highly nontrivial, as indeed one might have anticipated from the noninteracting case. The central problem can be seen already by considering the simplest nontrivial case (N=2 in two dimensions). Imagine that we plot the wave function schematically is a function of *relative* coordinates $r_1 - r_2 = r$, $\theta_1 - \theta_2 = \theta$. For $\Phi = 0$ or $\Phi_0/2$, for which the wave function is real, there must exist a symmetry-dictated nodal surface in the (r,θ) plane passing through $r=\theta=0$. However, it need not necessarily intersect the "walls" as in Fig. 5: the behavior shown in Fig. 6 is equally possible. Thus, there is no simple argument analogous to the one-dimensional case, there may be a sophisticated one, but it seems to involve a nontrivial exercise in N-dimensional topology. Needless to say, in the case where the transverse dimensions are very short one can argue that to form the nodal surface as in Fig. 6 would cost too much kinetic energy, and therefore the one-dimensional argument should go through; but the region of applicability of this argument is rather limited.

Let's now go back to the one-particle problem and try to get some feeling for how the one-particle wave function evolves as a function of position in (three-dimensional) space as we change the flux Φ. Of course the details of this process are sensitive to the exact form of the disorder, but there are some general things we can say. We confine ourselves for the moment to the groundstate. Then for $\Phi=0$ the wave function, as we have seen, is real and must undergo a change of $2n\pi$ as we go around the ring: by a simple adaptation of the above argument from eqn. (26) we easily show that n=0 for the groundstate, that is, the winding number is zero. Going now to $\Phi=\Phi_0/2$, we see that the wave function is again real but now has a phase charge of $n\pi$ as we go around. Consequently, there must be a nodal surface intersecting the ring somewhere (Fig. 7). Now, what happens for Φ equal to neither of these values? To visualize this, it is helpful for the moment to consider the limit of very weak disorder, so that we can describe the problem approximately in terms of the eigenfunctions of the homogeneous system. We assume that the transverse dimensions of the ring are small compared to its radius R. Consider then a value of Φ fairly close but not equal to $\Phi_0/2$, say $\Phi_0/2 + \alpha$. The wave function Ψ is approximately a product of the transverse groundstate, say $\chi_0(r)$, times a linear superposition of the two lowest states which meet the boundary condition, namely $\exp[i\theta(2\alpha+1)/2]$ and $\exp[i\theta(2\alpha-1)/2]$, with energies $h^2mR^2(1 \pm \alpha)^2/4$ respectively. That is

$$\Psi = \left[a \exp[i\theta(\tfrac{1}{2} + \alpha)] + b \exp[i\theta(\alpha - \tfrac{1}{2})] \right] \chi_0(r), \tag{31}$$

where the (complex) coefficients a and b are determined by the matrix elements of the scattering potential between the two states. It is clear that Ψ nowhere vanishes, though it may become very small on a particular cross-section of the ring.

Things are a bit more interesting if we consider a fairly highly excited state of the system. The point, now, is that although the minimum transverse splitting may be much larger than the lowest "circumferential" splitting, which as above is $h^2/8mR^2$, it may well be only comparable to the splitting $\Delta E(n,\alpha) \equiv E(N+1,\alpha) - E(n,\alpha) = nh^2/8mR^2$ even for $\alpha << 1$. The interest of this is that we can now form superpositions of the type

$$\Psi = \left[a \exp[in\theta(\tfrac{\alpha}{n} + \tfrac{1}{2})] + b \exp[in\theta(\tfrac{\alpha}{n} - \tfrac{1}{2})] \right] \chi_1(r) \tag{32}$$

and the two branches of the superposition may now be quite close in energy and hence a and b of the same order, even if $\Delta E(n,\alpha)$ is itself large, since this can be compensated by the transverse difference $E_1 - E_0$.

Let us now go back to the "true" wave function by multiplying by $\exp[-i(2\alpha + 1)\theta/2]$

$$\Psi(\theta,r) = [a\,\chi_0(r) + b\,e^{-i\theta}\chi_1(r)]\,e^{in\theta}. \tag{33}$$

We assume for present purposes that a and b may, but need not, be comparable in magnitude. What is the topology of this wave function (or rather the part in brackets)? If b=0, it has winding number zero over the whole cross-section of the ring. If a=0, the winding number is 1 over the whole cross-section. But if a and b are both of order 1, then in general *there will be a vortex* (node) somewhere in the tube: it is found at the point where

$$\frac{\chi_1(r)}{\chi_0(r)} = -\frac{a}{b}e^{i\theta} \equiv |c|\,e^{i(\theta+\delta_0)}, \tag{34}$$

where $\delta_0 \equiv$ phase(a/b). Since χ_1 and χ_0 are real, the vertex, if it exists, will occur at the point $\theta = -\delta_0$, $\chi_1/\chi_0 = |c|$. It is clear that since the transverse groundstate χ_0 is nodeless while χ_1 has a single node, there is always a vortex for arbitrarily *small* values of a/b, but not for arbitrarily *large* values. If the external flux is gradually swept up, then the ratio a/b changes gradually from very large to very small, and at some point the vortex is created at the edge of the ring, moves out to the middle, and stays there as $a/b \to 0$.

Now consider the case where there are very many relevant transverse states in the range of interest. We can write quite generally

$$\Psi = a\chi_+(r) + b\chi_-(r)e^{-i\theta}, \qquad (35)$$

where $\chi_+)(r)$ and $\chi_-(r)$ are both in general *linear combinations* of energy eigenstates. Now the vortex can in principle sweep right across the ring as Φ is varied through $\Phi_0/2$.

It is clear that the case of interaction with phonons can be handled in a very similar way, with r now representing a generalized coordinate. So again, at first sight, there is no particular reason to single out the phonons as qualitatively different from the transverse degrees of freedom. Clearly the problem lies deeper than this.

REFEREENCES

Alefeld, B., Badurek, G., and Rauch, H., 1981, *Z. Phys. B*, 41:231.

Ashcroft,N. W., and Mermin, N. D., 1976, "Solid State Physics," Holt, Rinehart and Winston, New York.

Badurek, G., Rauch, H., and Summhammer, J., 1983, *Phys. Rev. Lett.*, 51:1015.

Buttiker, M., Imry, Y., and Landauer, R., 1983, *Phys. Lett.*, 96A:365.

Imry, Y., 1986, *in* "Directions in Condensed Matter Physics," ed. Grinstein, G., and Mazenko, G., World Scientific, Singapore.

Leggett, A. J., 1973, *Physica Fennica*, 8:125.

Leggett, A. J., Chakravarty, S., Dorsey, A. T., Fisher, M. P. A., Garg, A., and Zwerger, W., 1987, *Rev. Mod. Phys.*, 59:1.

Levy, L. P., Dolan, G., Dunsmuir, J., and Bouchiat, H., 1990, *Phys. Rev. Lett.*

Loss, D., and Mullen, K., 1990, unpublished.

Santhanam, P., Wind, S., and Prober, D. E., 1984, *Phys. Rev. Lett.*, 53:1179.

Stern, A., Aharonov, Y., and Imry, Y., 1990, *Phys. Rev. B.*

Umbach, C. P., Washburn, S., Laibowitz, R. B., and Webb, R. A., 1984, *Phys. Rev. B*, 30:4048.

9

GRANULARITY IN NARROW WIRES: CONDUCTANCE FLUCTUATIONS, DIFFUSE BOUNDARIES AND JUNCTION SCATTERING

T J Thornton, M L Roukes, A Scherer and B P Van der Gaag

Bellcore, 331 Newman Springs Road
Red Bank, NJ07701, USA

INTRODUCTION

Many factors determine the electrical conductivity of a narrow wire. In a weakly disordered system we can begin with the Drude conductivity, $\sigma=ne^2\tau/m$, and add corrections which take account of various aspects of transport not included in this simple expression (see Fig. 1). For instance, disorder induced corrections arise from quantum interference and electron-electron interaction, generally lowering the conductivity. The interference arises because of coherent backscattering of electrons and reduces the diffusion coefficient below the classical result ($D=v_F l/2$ for specular scattering) while the e-e interaction lowers the density of states at the Fermi energy. We can rewrite the Drude expression in terms of the diffusion coeffucient, D, and the density of states, $N(E_F)$, as $\sigma=e^2DN(E_F)$ and provided the corrections are small, $\delta\sigma/\sigma \sim \delta D/D+\delta N/N(E_F)$, so that the two are additive. In one dimension the weak localisation Q.I. (Altshuler and Aronov, 1981) and interaction corrections (Altshuler and Aronov, 1983) can reduce the conductivity by up to 10% and are given by

$$\delta\sigma_{WL} = -\frac{2e^2}{h}\frac{L_\varphi}{L}, \qquad \delta\sigma_{int} = -\frac{2e^2}{h}\sqrt{\frac{hD}{2\pi kT}}. \qquad (1)$$

The expressions for $\delta\sigma_{WL}$ and $\delta\sigma_{int}$ given above are derived for a uniform, isotropic media and apply equally well in two and three dimensions, albeit in a slightly modified form. This school, however, is concerned with the *granular* properties of small systems, i.e. how the impurity distribution, boundaries, voltage probes, etc., affect the transport. In some cases the granular properties can lead to dramatic changes in, say, the magnetoresistance while in others there will be only a small modification to an effect already well developed in an isotropic system. We begin with a short introduction to the ways in which granularity can effect the resistivity of a narrow wire and then present our results on boundary and junction scattering in ion exposed wires.

GRANULARITY IN NARROW WIRES

Many studies of narrow semiconductor wires have used a high mobility 2DEG as the starting material which is then *laterally* confined by various microfabrication techniques. When a wire is made in this way anisotropy can appear for two reasons. The first is due to the random position of impurities and is intrinsic to small, heterojunction wires. The second results from the microfabrication itself. The laterally imposed confinement potential will interact with the random impurities leading to variations in wire width and electron density (Davies and Nixon, 1990). In addition, voltage probes attached to the wire can scatter and perhaps trap the electrons in the junction. The influence of the wire structure on electron transport will be more significant in narrow wires where the granularity is most pronounced.

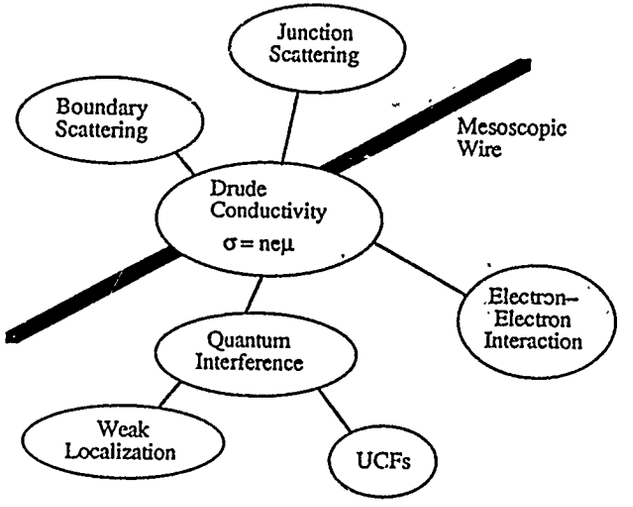


Fig. 1 Schematic representation of some of the mechanisms which determine the conductivity of narrow ballistic wires.

Granularity in a narrow wire will influence the conductivity in a number of ways. If the edges are rough, and scatter diffusely, the extra scattering will reduce the value of τ and lower the conductivity. In this case the magnetoresistance shows a large anomaly which is discussed in detail below. Other effects will arise from the discrete nature of the impurities which modifies the quantum interference leading to sample specific fluctuations (Lee *et al.*, 1987). Conductance switching due to electron trapping by a single impurity in the vicinity of the wire has also been observed (Scopol *et al.*, 1986).

Effects due to voltage probes attached to the wire are rather more varied. In the ballistic regime electron focussing by flared junctions has been suggested (Beenakker and van Houten, 1989) and demonstrated experimentally (Molenkamp *et al.*, 1990). Chaotic scattering within the junction has also been suggested (Beenakker and van Houten, 1989) as has the existence of virtual states bound at the junction of very narrow wires (Kirczenow, 1989). The latter might provide another mechanism for discrete switching of the conductance depending upon whether the bound state is empty or filled. We show below that in ballistic wires the presence of voltage probes significantly increases the resistance of the wire, presumably because of extra scattering within the junction.

CONDUCTANCE FLUCTUATIONS

The weak localisation correction arises from quantum interference between so called time reversed electrons (Bergman, 1983). These are electrons which are scattered back to the origin by a number of elastic collisions before they suffer an inelastic, that is, a phase breaking collision. The time reversed electrons follow identical trajectories but in opposite directions and therefore arrive back at the origin exactly in phase. The resulting constructive interference leads to an increased probability of finding the electron back at the origin which amounts to a reduced conductivity.

The time reversed trajectories are only a small subset of all possible trajectories which give rise to quantum interference but they are unique because of the strict phase relation between the interfering electron waves. Electrons which follow other trajectories can interfere but the resulting phase will vary randomly between 0 and 2π. In a large system many possible trajectories will be sampled and effects due to the random phase interference will average to zero leaving only the contribution from the time reversed trajectories, i.e. the weak localization. In a narrow wire, however, no such self-averaging occurs because the electron paths will be strongly influenced by the *actual* position and nature of the impurities along the wire. The superposition of coherent electron waves will now effect the conductivity according to the specific distribution of impurities. In other words, nominally identical wires will have different conductivity because of the different impurity distribution within each wire. The same applies to the low field magnetoresistance which will be different for each sample. These sample specific fluctuations were first seen in narrow metal wires (Umbach *et al.*, 1984) and have subsequently been observed in narrow Si MOSFETs (Scopol *et al.*, 1986) as well as GaAs (Taylor *et al.*, 1989) and heterojunction wires (Thornton *et al.*, 1987). The amplitude of the conductance fluctuations was shown to be independent of the degree of disorder and therefore, in this sense, universal (Lee *et al.*, 1987). Although appearing random, the fluctuations in, say, the magnetoresistance are reproducible with a well defined rms amplitude and correlation length. For instance, in a dirty metal sample (l_o « W) with L_φ « L_φ « L the amplitude and magnetic and electric correlation lengths of the fluctuations depend on the width and phase coherence length as

$$\delta g = -a\frac{e^2}{h}\left(\frac{L_\varphi}{L}\right)^{-3/2}, \qquad (2)$$

$$B_c = b\frac{\Phi_o}{WL}, \qquad (3)$$

$$E_c = \frac{\pi h D}{2L^2}, \qquad (4)$$

where a and b are constants of order unity which depend upon the exact geometry of the sample (Lee *et al.*, 1987). In a long ballistic wire fluctuations also occur but flux cancellation changes the expressions for δg and B_c given in (2) and (3) (Beenakker and van Houten, 1988).

Conductance fluctuations are a clear example of inhomogeneity in narrow wires arising from the discrete nature of the impurities. To illustrate this phenomena we show data from a split gate wire in Fig. 2. Here the channel is confined between a pair of reverse biased gate electrodes on the surface of a GaAs:AlGaAs heterojunction (Thornton *et al.*, 1986). The channel width can be varied in the range 1000Å <W<1μm by altering the gate potential and the figure shows the conductance of a 15 μm long channel close to pinch-off. The conductance displays aperiodic fluctuations superposed on a background which decays almost linearly with gate voltage. The fluctuations are reproducible even after temperature cycling to 77K and have an amplitude of approximately $2\times10^{-6}\Omega^{-1}$ ie approximately 1/20 of e^2/h. From this result and (2) we estimate the phase coherence length to be approximately 2μm at 0.35K, similar to other estimates.

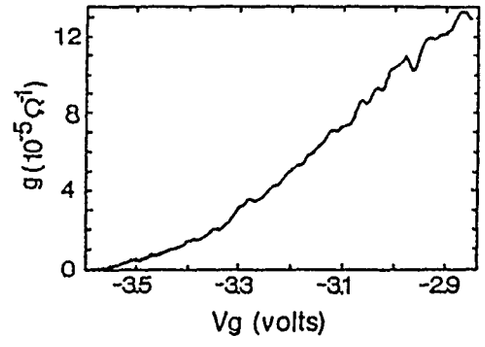


Fig. 2 Conductance fluctuations in the pinch-off characteristics of a split gate wire at 0.35K.

Changing the gate voltage on a split gate device changes both the wire width and carrier density. In the range $-3.2V<V_g<-2.9V$ the density, calculated from the Shubnikov-de Haas oscillations, varies between 1.27 and 1.64 $\times10^{15}$ m^{-2}. In this range of gate voltage the conductance fluctuates significantly when the gate voltage changes by ~10mV corresponding to a change in Fermi energy of 40μeV which is much larger than the 6μeV predicted by (4) (D is estimated to be ~0.2m^2/s). Equations (3) and (4) are the zero temperature results and at finite temperatures L should be replaced by L_φ. Taking account of finite temperature in this way gives an energy correlation of greater than 300μeV. This is now much larger than the thermal energy and the fluctuations scale will be correlated over a few kT (Lee *et al.*, 1987) which is 30μeV and close to the experimental result.

Clearly, the change in electron density as the wire is 'squeezed' is sufficient to explain the observed fluctuations in the conductance. However, we cannot overlook the effect of changing the wire width. Equations (2) to (4) give a statistical description of how the conductance of narrow wires varies from sample to sample or how it fluctuates with changes in magnetic field or Fermi energy. Qualitatively, these findings suggest that anything which alters the nature of a disordered wire will lead to fluctuations in the conductance. For instance, it has been shown that moving a single impurity some fraction of a Fermi wavelength is sufficient to alter the conductance (Feng *et al.*, 1986; Altshuler and Spivak, 1986) . Changing the wire width will have a similar effect if the conducting channel moves relative to the impurities. We estimate that a reduction in gate voltage of 10mV will reduce the wire width by ~100Å thereby moving the channel a significant fraction of the 600Å Fermi wavelength. This may well be sufficient to alter the scattering in the channel and may contribute to the fluctuations shown in (1).

Random fluctuations in conductance have important implications for the device applications of quantum wires. Such applications include Aharonov-Bohm loops and stub-tuners as electron interferometers (Datta, 1989). In these devices quantum interference between electron waves is used to modulate the conductance, switching the current between 'on' and 'off' values by means of a gate electrode. In principle the conductance can be modulated in a well defined and controllable manner. However, the presence of impurities disrupts the ideal characteristics which will vary from device to device.

One way around the problems associated with a granular impurity distribution is to make the wires very narrow, ideally so narrow that all the electrons are in the lowest subband. Conductance modulation in these single subband wires is less sensitive to the impurity distribution, presumably because of the reduced probability of electrons being back-scattered

(Sakaki, 1980). Unfortunately this approach requires wire widths of the order of 100Å which is much smaller than the ~1000Å of most present day systems.

Another way to avoid fluctuations is to make the active region so small that an exiting electron has a negligibly small chance of being coherently scattered back into the active region. A good example of this is the point contact geometry which displays a quantised resistance as shown in Fig. 3. The resistance is given by $h/2Ne^2$ and increases in a step-like fashion as the number of subbands N is reduced (van Wees *et al.*, 1988, Wharam *et al.*, 1988). The width of the point contact is a few thousand Ångstroms and therefore much smaller than the elastic length. An electron which emerges from the point contact has very little chance of returning before suffering an inelastic collision which destroys the phase coherence. Numerical calculations of transport through a narrow constriction show that the quantized steps are remarkably robust and impurities have to be within a few Fermi wavelengths of the opening before they disturb the quantisation (Haanappel and van der Marel, 1989). As a result the pinch off characteristics of point contacts rarely show fluctuations in the conductance. For this reason quantum interference devices which use point contacts as emitters and collectors should, in principle, be free of impurity driven fluctuations.

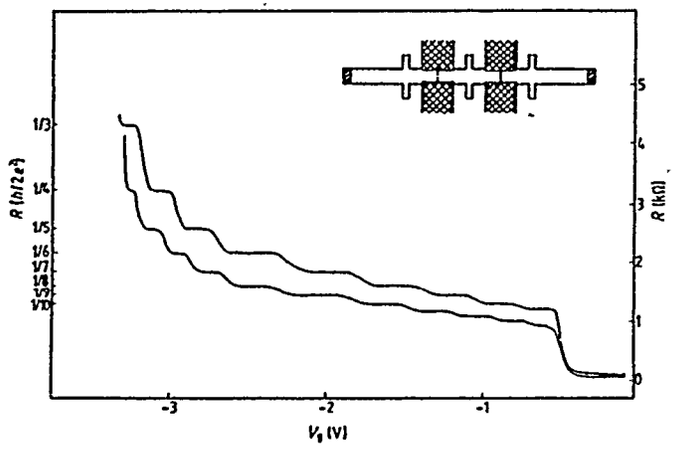


Fig. 3 Quantised resistance of a point contact at T<100mK.

BOUNDARY SCATTERING

The previous section dealt with fluctuations arising from the random impurity distribution within a modulation doped heterojunction and specific examples were taken from wires confined electrostatically between split gates. In this section we are concerned with the diffuse boundary scattering introduced by the lateral confinement of a 2DEG. Although effects due to diffuse boundary scattering have been observed in split gate wires the clearest results are seen in those defined by ion exposure and we describe this technique below.

Lateral Confinement by Ion Exposure

If a modulation doped GaAs:AlGaAs heterojunction is exposed to a beam of ions the conductivity of the 2DEG is drastically reduced and can easily be rendered insulating. This has been widely used as a means of patterning quantum wires and dots. There are basically two approaches; one uses a high energy *focussed* beam which is scanned over the surface and selectively exposes those regions to be turned into insulators (Hiramoto *et al.*, 1987, Hirayama *et al.*, 1988); the other approach uses a broad beam of ions to transfer the pattern of a mask into the underlying 2DEG (Scherer and Roukes, 1989). We have used the latter method which has been shown to have a resolution of a few hundred Ångstroms.

Figure 4 is a schematic cross-section through one of the wires. Once the mask has been patterned by electron beam lithography and lift-off the entire device is exposed to Ne ions of energy ~150 eV. The device resistance is monitored during the exposure to ensure that it receives the optimum dose for pattern transfer. A metal mask doubles as a self aligned gate with which to vary the electron density of the 2DEG in the range $1 - 5 \times 10^{15}$ m^{-2}. A useful property of these wires is the fact that the wire width changes very little over the available range of electron density (Thornton *et al.*, 1990, Roukes *et al.*, 1990). This is in contrast with split gate (Ford *et al.*, 1989) or shallow etched wires (Chang *et al.*, 1989) where the width can vary considerably with electron density. In the following discussion of the magnetoresistance it simplifies the analysis if we assume that the wire width remains constant as the density is varied.

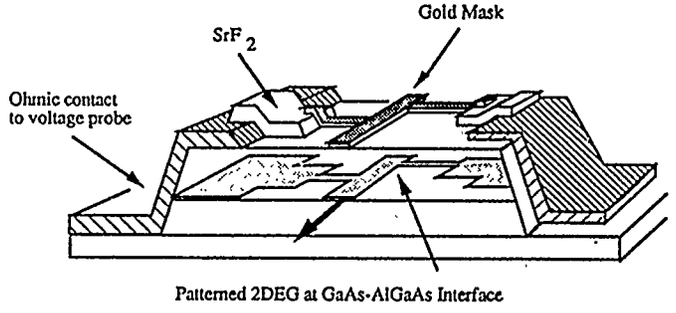


Fig. 4 Schematic cross-section through an ion exposed wire with a self aligned gate. The arrow indicates direction of current flow.

Resistivity and Diffuse Boundary Scattering

In a narrow wire, collisions with the boundaries will be important. This is especially true in wires made from 2DEG material where the transport mean free path in a large sample can exceed 10μm (Foxon *et al.*, 1989). The boundary collisions are a sensitive probe of the roughness at the interface and it turns out that a significant proportion are diffusive, i.e. momentum parallel to the edge is not conserved. The reason for the roughness at the edges is not clear although certainly the random nature of the ion implantation must be significant. Other effects such as ion channeling might also play a role (Laruelle *et al.*, 1990). However, it is clear that in narrow wires diffuse boundary scattering can dominate the resistivity by limiting the mean free path (Thornton *et al.*, 1989) and might well place constraints on the device applications of ballistic transport in narrow wires.

Before discussing how boundary scattering alters the resistivity we must first characterise the roughness of the boundaries. Fuchs (1938) was the first to do this by introducing the parameter p which describes the probabilty of specular scattering and varies between zero and unity corresponding to completely diffuse or completely specular scattering. Although a more sophisticated approach might consider how p varies with incident angle and wavelength, etc., here we consider it as an empirical quantity which describes the *average* properties of an electron interacting with the boundary.

The effects of edge roughness can be incorporated into our description of transport in narrow wires by considering how it effects the mean free time, t, in the Drude conductivity. The simplest way to do this is to invoke Matthiessen's rule ie assume the bulk and edge scattering rates are additive. First, we need to estimate the average distance that an electron will propagate along the wire before suffering a diffuse collision at the boundary. We designate this the boundary scattering length, l_b. Over this distance an electron collides approximately $l_b/W \sim 1/(1-p)$ times with the edges before scattering so that $l_b \sim W/(1-p)$. The effective mean free path is therefore $1/l_{eff}=[1/l_o+(1-p)/W]$ and the total scattering rate will be $1/\tau=hk_F/ml_{eff}$. Inserting this in the Drude expression gives

$$\rho_{eff} = \rho_o l_o \left[\frac{1}{l_o} + \frac{1-p}{W}\right], \qquad (5)$$

where ρ_o is the resistivity of a large 2DEG and $\rho_o l_o = hk_F/ne^2$. From (5) we can see that when $(1-p) > W/l_o$ the dominant electron scattering is from that at the boundaries and for a 1000Å wire with a bulk mean free path of 10 μm this will be true for $p < 0.99$. Our estimates of p are significantly smaller than this and diffuse boundary scattering will dominate the resistivity of proportionately larger wires.

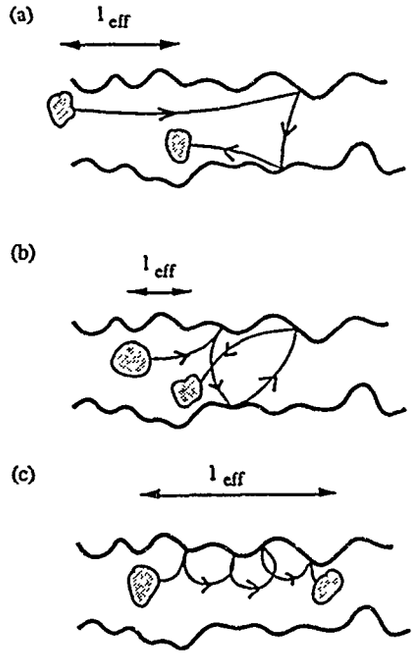


Fig. 5 Electron-boundary scattering in narrow wires with rough edges in a perpendicular magnetic field. (a) When l_o » W the effective mean free path, l_{eff}, is dominated by boundary scattering. (b) At higher fields when $w/r_c \sim 0.55$, more electrons are forced to interact with the walls and the resistivity increases. (c) When $2r_c < W$ electrons are confined to either edge and the backscattering due to the boundaries is quenched. The shaded 'clouds' represent elastic scattering within the wire.

The contribution that boundary scattering makes to the total resistivity can be varied by the application of a perpendicular magnetic field. At low fields the resistivity is actually increased because of the diffuse scattering but can be reduced to the bulk value by a sufficiently large magnetic field. The reasons for this are shown schematically in Fig. 5. At zero and very small magnetic fields (Fig. 5a) the electrons with a large component of momentum parallel to the wire axis interact infrequently with the edges and can contribute significantly to the conductivity even for $p \ll 1$. However as the field is increased these electrons are forced to collide with the edges (Fig. 5b) and for $p < 1$ the extra diffuse scattering will increase the resistivity. The low field positive magnetoresistance eventually saturates when the boundary scattering is at a maximum and any further increase in the field leads to a drop in the resistivity. Eventually, the cyclotron diameter will be smaller than the wire width (i.e., $2hk_F/eB<W$) and an electron scattering off one edge will be confined to the same edge unless it suffers a collision within the bulk of the wire (Fig. 5c). At high enough fields and provided Landau quantization is not important, electrons will be confined to either edge over distances of the order of l_o and the bulk conductivity is recovered. In effect, the back scattering due to the boundaries is *quenched* at high magnetic fields. The suppression of backscattering from impurities within a channel confined by perfectly specular boundaries has been considered by Büttiker (1988). In our samples, $p < 1$ and l_o » W and the backscattering arises mainly from the rough boundaries, with only a small contribution coming from the impurities.

The qualitative behaviour discussed above can be clearly seen in a 1 μm wide wire (Fig. 6). The zero field resistivity is 75 Ω and much larger than the 20 Ω measured in a broad sample at the same carrier density. As the magnetic field is increased a small anomalous peak appears before the resistivity drops rapidly to a value of ~ 25 Ω. After this the resistivity is constant until the onset of Shubnikov - de Haas oscillations. The position of the anomalous resistance, B_{max}, is determined by the ratio of the wire width to the cyclotron radius and calculations by Forsvoll and Holwech (1964) and also by Pippard (1989) have shown that $W/r_c=0.55$ at the maximum. These calculations are based on classical electron trajectories but fully quantum mechanical solutions by Akera and Ando (1990) give a very similar result. We have tested this result by plotting the value of r_c at the maximum against the nominal mask width for many different wires all of which were defined using the optimal ion dose for pattern transfer. A straight line fit to the data in Fig. 7 gives $W/r_c=0.54\pm0.01$. This result implies that in a given wire B_{max} will be proportional to $k_F=(2\pi n)^{1/2}$ as long as the wire width is independent of n. The variation of B_{max} with $\sqrt{n}$ is shown in the inset to Fig. 7 and the straight lines are derived from the expression $B_{max}=0.55h(2\pi n)^{1/2}/eW$ and in each case the best fit gave a vlaue for W which was within 10% of the nominal mask width.

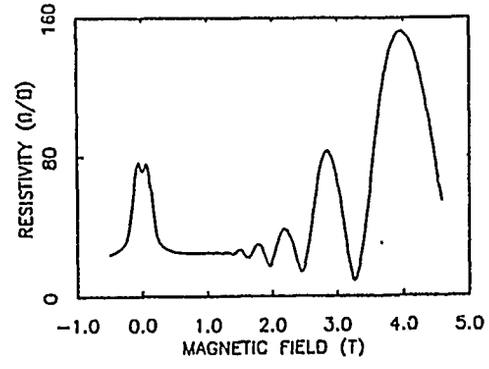


Fig. 6 The magnetoresistance of a 1 μm wide wire demonstrating the complete suppression of the boundary scattering resistance.

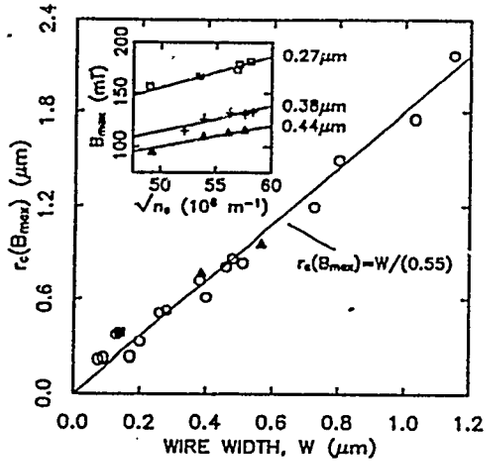


Fig. 7 The cyclotron radius at the peak position, $r_c(B_{max})$, plotted against wire width for many different wires. The variation of B_{max} with carrier density, n, is shown in the inset.

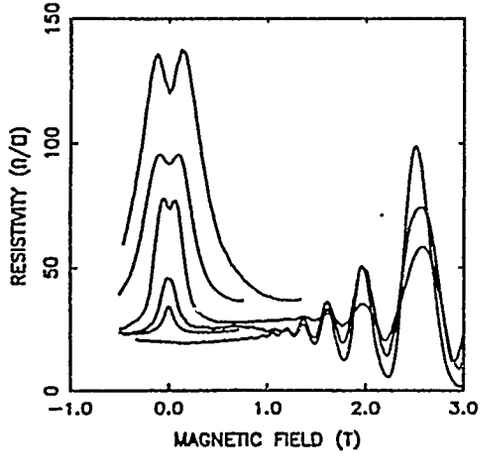


Fig. 8 Magnetoresistance of ion exposed wires with widths 0.42, 0.74, 1.0, 2.0, 3.6, 6.2 μm from top to bottom. The carrier density in each case is in the range $4.2\pm0.1\times10^{15}$m^{-2}.

The magnetoresistance from a number of wires with widths ranging from 7 μm to 1400 Å are shown in Fig. 8. The carrier density, determined from the periodicity of the Shubnikov-de Haas oscillations, is similar in each case so that we can compare resistivities. In the widest wires there is almost no magnetoresistance before the onset of the Shubnikov-de Haas oscillations and because W is larger than l_0 boundary scattering makes a negligible contribution to the resistivity which is similar to that in a broad 2DEG. However, in narrower wires the zero field resistivity increases and a small peak at finite field can be seen for wires of width less than 2 μm. The magnitude of both the zero field resistivity and the amplitude of the peak at B_{max} increase with decreasing wire width and in all cases the additional resistivity is quenched by fields of less than 1 T. If, as we expect, the resistivity remaining after the boundary scattering is quenched corresponds to the resistivity of the bulk then it should not vary with wire width. The data in Fig. 6 appears to contradict this because of a systematic increase in the background resistivity as the wire width is decreased. However, we believe that the background resistivity does in fact represent the bulk value which has increased because of a reduction in the electron mobility. The drop in mobility is most likely due to damage caused by ions which are scattered sideways under the mask and will be most significant for the narrowest wires.

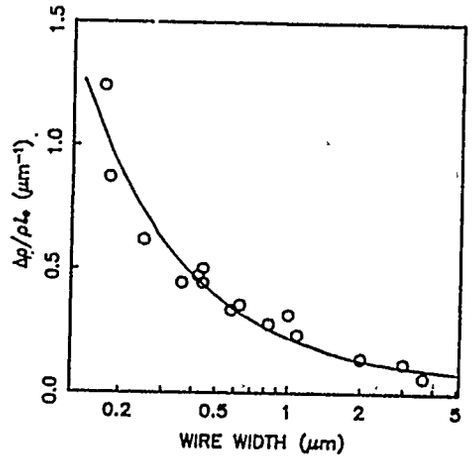


Fig. 9 The increase in the zero field resistivity with decreasing wire width. The solid line is the best fit to (5).

According to (5) the boundary scattering increases the wire resistivity by an amount $\Delta\rho = (1-p)\rho_0 l_0/W$ and in Fig. 9 we plot the extra resistivity as a function of the wire width. All the data was obtained from wires defined by exposure to the optimum dose of 150 eV neon ions and the resistivity was measured at carrier concentrations in the range 3 - 4 $\times10^{15}$m^{-2} corresponding to an approximately constant Fermi wavelength of 350-400 Å. This was necessary to ensure that the electron-boundary interaction, as characterised by p, was the same in each wire. We have obtained a reasonable fit to the data using (5) from which we estimate a value for p of 0.85. Although 85% of the collisions are in fact specular the boundary scattering makes the largest contribution to the resistivity for wires narrower than 1 μm where $l_b \sim 7$ μm and comparable to the bulk mean free path.

The low field magnetoresistance is strongly dependent on the electron density as shown in Fig. 10 for a 1 μm wide wire. The progressive reduction in the relative size of the low field anomaly as the density decreases is due to the rapid reduction in the transport mean free path, $l_0=h\mu\sqrt{(2\pi n)}/e$. As the mean free path decreases the bulk or background resistivity increases while the importance of diffuse boundary scattering decreases and at low carrier density ρ_0 approaches $\rho(0)$. The behaviour of a number of wires is shown in Fig. 11. Here we have plotted both the zero field resistivity, $\rho(0)$, (open symbols) and the value after quenching the boundary scattering contribution to the resistivity (solid symbols). The variation with carrier density is similar in all the wires we have measured and to describe the behaviour we will consider the 1 μm wire marked as circles in the figure. At high densities ρ_0 is less than half of $\rho(0)$ but has a much steeper slope and the two curves eventually coalesce when $n \leq 1.5\times10^{15}$m^{-2}, i.e. when $l_0 \ll W$. The dependence of ρ_0 on n follows an approximate power law behaviour. For all the wires we have studied ρ_0 varies as n^{-a} where a lies in the range 1.8 to 2.3. The variation with mobility, as defined by the Drude result, $\sigma = ne\mu$, is therefore $\mu \sim$

$n^{0.8-1.3}$. The density dependence of mobility is very similar to that observed in wide heterojunction samples (Hirakawa *et al.*, 1985) and confirms our assumption that the background resistivity results from scattering events within the bulk of the wire, the boundary scattering contribution having been quenched by the magnetic field.

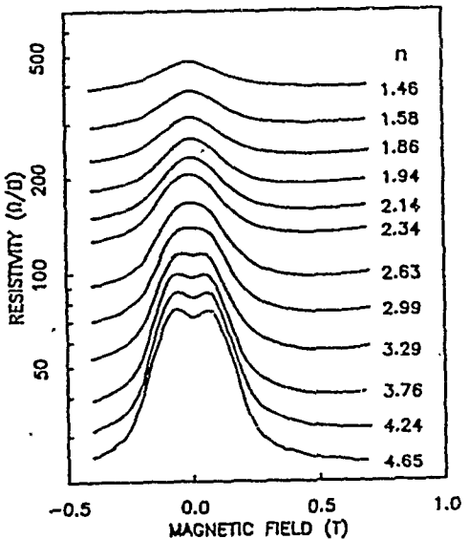


Fig. 10 The low field magnetoresistance of a 1 μm wide wire for different electron densities.

JUNCTION SCATTERING

Up to now we have been concerned with how the random impurity distribution and fluctuations in the confinement potential alter the resistivity of narrow wires. The effects of introducing junctions into the wire, either to inject current or measure voltage, have been tacitly ignored. Unlike the case of impurities and boundary roughness, we have a fair degree of control when it comes to introducing additional junctions to the wire. In quasi-ballistic wires (l_0 « W) several interesting transport anomalies have been observed (Roukes *et al.*, 1987; Timp *et al.*, 1988; Takagaki *et al.*, 1988; Ford *et al.*, 1989) and attributed to ballistic transport within the junctions. This represents another kind of granularity in narrow wires and is unique in that we can consider the effects of junction position and geometry in a systematic and controllable fashion. The effect of junction curvature on various transport anomalies has already been considered by Roukes *et al.* (1990). In this section we consider how the resistance is altered by introducing additional junctions directly into the current path.

From the resistivity peak at B_{max} we can determine the width of the wire with sufficient accuracy that we can estimate the wire resistivity to better than 10%. As a result, the resistivity of different sections of a multi-probe wire are nominally identical as long as the only probes in the vicinity of the wire are those used to measure the voltage. However, when we compare the resistivity of sections which include one or more junctions we observe a sytematic increase in the resistivity. We attribute this to an increase in the wire *resistance* due to the presence of the junction. Typical results, from a 1.0 μm wire at a density of $4.65 \times 10^{15} m^{-2}$, are shown in Fig. 12. The bottom curve shows the low field magnetoresistance of a 4 μm long section of the wire with no probes between those used to measure the voltage. The resistivity at zero field increases as first two probes (L=11.8 μm, middle curve) and then four probes (L=18.5 μm, top curve) are introduced between the voltage probes. The additional probes in the current path lead eventually to large area ohmic contacts which are left floating. At magnetic fields above −0.3 T the junction resistance is suppressed and the difference in resistivity between the three sections can be explained by the uncertainty in the measurement of the width. At zero field, however, the additional resistance is larger than the experimental uncertainty and increases with the number of probes.

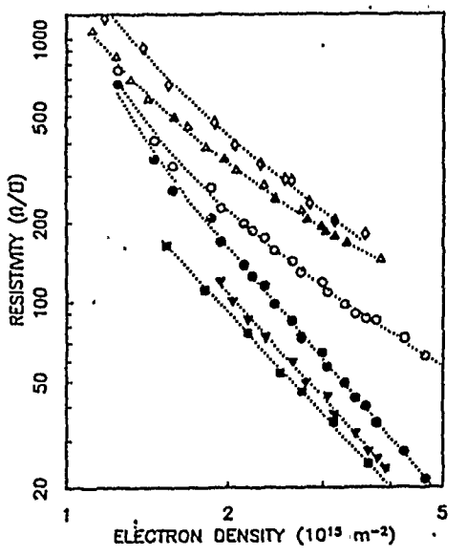


Fig. 11 The density dependence of ρ(0) (open symbols) and ρ_0 (closed symbols). Wire widths are 10 μm (squares), 3.6 μm (solid triangles), 1 μm (circles), 0.74 μm (open triangles) and 0.63 μm (diamonds).

In Fig. 13 we have plotted the average additional resistance per probe as a function of mode number, $k_F W/\pi$. The mode number was varied over a large range by adjusting the carrier density (and therefore k_F) in several wires of different width. The width of the probes and the curvature at the junction was kept fixed in each case. The fairly good overlap between the results from different wires suggests that the junction resistance is not very sensitive to the position of impurities or variations in the confining potential. The additional resistance due to a junction increases quickly with decreasing mode number and can be a significant fraction of the total wire resistance. For instance, the junction resistance of a 10 μm length at the lowest mode numbers shown in Fig. 13 is approximately 25% of the total. This demonstrates the importance of scattering from lithographically defined junctions and further work is needed to determine how the junction resistance varies with curvature and probe width.

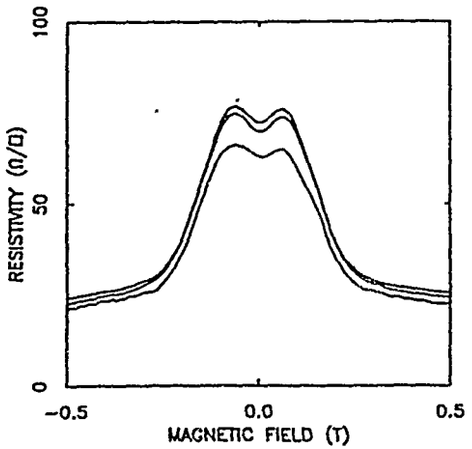


Fig. 12 The low field magnetoresistance of a 1.0 μm wide wire for zero (bottom curve), two (middle curve) and four (top curve) wire junctions between the voltage probes. The electron density is $4.65 \times 10^{15} m^{-2}$.

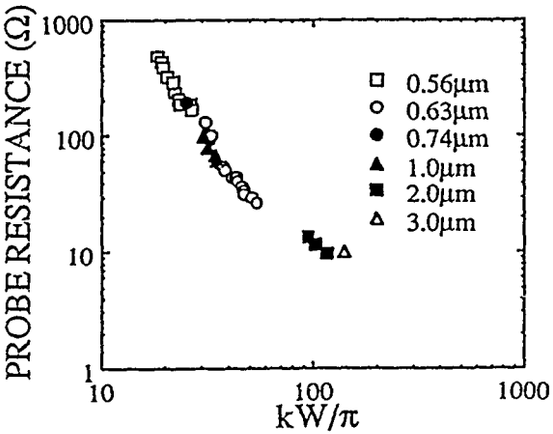


Fig. 13 The average resistance per junction as a function of mode number.

CONCLUSIONS

In this article we have illustrated some of the ways in which the granular nature of narrow, ballistic wires influence their transport properties. In high mobility wires the transport mean free path is large enough that electrons can propagate many microns before being scattered by phonons or impurities. As a result they are sensitive to anisotropies in impurity distribution, confining potential and geometry over distances of the order of l_0, a length scale which can be easily probed by means of a magnetic field. We have shown that these anisotropies can dominate the resistivity and will have to be considered in any device applications of quantum wires.

REFERENCES

Akera, H., and Ando, T., 1990, ICPS to be published.

Altshuler, B. L., and Aronov, A. G., 1981, *JETP Lett.*, 33:501.

Altshuler, B. L., and Aronov, A. G., 1983, *Solid State Comm.*, 46:429.

Altshuler, B. L., and Spivak, B. Z., 1985, *JETP Lett.*, 42:447.

Beenakker, C. W. J., and van Houten, H., 1988, *Phys. Rev. B*, 38:3232.

Beenakker, C. W. J., and van Houten, H., 1989, *Phys. Rev. Lett.*, 63:1857 ; see also "Electronic Properties of Multilayers and Low Dimensional Solids" Ed. by J M Chamberlain, L Eaves and J C Portal (Plenum, London, 1990).

Bergmann, G., 1983, *Phys. Rev. B*, 28:2914.

Buttiker, M., 1988, *Phys. Rev B*, 34:9375.

Chang, A. M., Chang, T. Y., Baranger, H. U., 1989, *Phys Rev Lett.*, 63:996.

Datta, S., 1989, *Superlattices and Microstruc.*, 6:83.

Davies, J. A., and Nixon, J. H., 1990, *Phys. Rev. B*, 41:7929.

Feng, S., Lee, P. A., Stone, A. D., 1986, *Phys, Rev. Lett.*, 56:1960.

Ford, C. J. B., Thornton, T. J., Newbury, R., Pepper, M., Ahmed, H., Peacock, D., Ritchie, D., Frost, J., Jones, G., 1989, *Phys. Rev. B*, 38:8518.

Ford, C. J. B., et al 1989, *Phys. Rev.Lett.*, 62:2724.

Forsvoll, K., and Holwech, L., 1964, *Phil. Mag.*, 9:435.

Foxon, C. T., et al 1989, *Semicond. Sci. Technol.*, 4:582.

Fuchs, K., 1938, *Proc. Camb. Phil. Soc.*, 34:100.

Haanappel E G, van der Marel D 1989 Phys. Rev. B39, 5484.

Hirakawa, K., Sakaki, H.,, Yoshino, J., 1985, *Phys. Rev. Lett.*, 54:1279.

Hiramoto, T., Hirakawa, K., Iye, Y., and Ikoma, T., 1987, *Appl. Phys. Lett.*, 51:1620.

Hirayama, Y., Tarucha, S., Suzuki, Y., and Okamoto, H., 1988, *Phys Rev B*, 37:2774.

van Houten, H., et al 1988, *Surf. Sci.*, 196:144.

Laruelle, F., et al 1990, *Appl. Phys. Lett.*, 56:1561.

Lee, P. A., Stone, A. D., Fukuyama, H., 1987, *Phys. Rev. B*, 35:1039.

Molenkamp, L., W., et al 1990, *Phys. Rev B*, 41:1274.

Pippard, A. B., 1989, "Magnetoresistance in Metals," Chapter 6 and references therein, Cambridge Univ. Press.

Roukes, M. L., et al 1987, *Phys. Rev. Lett.*, 57:3011.

Roukes, M. L.,Thornton, T. J., Scherer, A., Van der Gaag, B. P., 1990, *in* "Electronic Properties of Multilayers and Low Dimensional Solids" Ed. by J M Chamberlain, L Eaves and J C Portal (Plenum, London).

Sakaki, H., 1980, *Jpn. J. Appl. Phys.*, 19:1735.

Roukes, M. L., Scherer, A., Van der Gaag, B. P., 1990, *Phys. Rev. Lett.*, 64:1154.

Scherer, A., and Roukes, M. L., 1989, *Appl. Phys. Lett.*, 55:377.

Scopol, W. J., et al 1986, *Phys. Rev. Lett.*, 56:2865.

Spector, J., Stormer, H. L., Baldwin, K. W., Pfeiffer, L. N., West, K. W., 1990, *in* Proc. 4th Int. Conf. on Modulated Semiconductor Structures (Ann Arbor) to be published in Surface Science.

Takagaki, Y., et al 1988, *Solid State Comm.*, 68:1051.

Taylor, R. P., et al 1988, Proc. 19th ICPS, Warsaw, Ed. by W. Zawadzki, Pub. by IOP Polish Academy of Sciences.

Timp, G., et al 1988, *Phys. Rev. Lett.*, 60:2081.

Thornton, T. J., Pepper, M., Ahmed, H., Davies, G. J., Andrews, D. A., 1986, *Phys. Rev. Lett.*, 56:1181.
Thornton T J, Roukes M L, Scherer A, Van der Gaag B P 1989, Phys. Rev. Lett 63, 2128
Thornton T J, Roukes M L, Scherer A, Van der Gaag B P 1990, in "Science and Technology of 1- and 0- Dimensional Semiconductors" S P Beaumont and C M Sottomayor-Torres eds. Plenum, London, to be published
Umbach, C. P., Washburn, S., Laibowitz, R. B., Webb, R. A., 1984, *Phys Rev B*, 30:4048.
van Wees, B., et al 1988, *Phys. Rev. Lett.*, 60:848.
Wharam, D. A., et al 1988, *J. Phys. C*, 21:L209.

(11)

NON-EQUILIBRIUM QUANTUM DOT TRANSPORT

Mark A. Reed, John N. Randall, and James H. Luscombe

Central Research Laboratories
Texas Instruments Incorporated
P. O. Box 655936, MS 154
Dallas, Texas 75265, USA

INTRODUCTION

Advances in fabrication technology and the understanding of electronic transport in low dimensional structures have led to fascinating discoveries (Heinrich *et al.*, 1988; Reed and Kirk, 1989). Structures such as electron waveguides and quantum point contacts are excellent laboratories to study fundamentals of electronic transport in the quantum ballistic regime. The creation of 3-dimensionally confined systems ("quantum dots") (Hansen *et al.*, 1990; Reed *et al.*, 1988; Van Wees *et al.*, 1989) is especially intriguing since these structures are analogous to semiconductor atoms, with energy levels tunable by the confining potentials. One method for creating these systems is the electrostatic confinement of an existing 2DEG. Unfortunately, electronic transport that reveals quantum size effects in these systems only occurs for low temperatures (<1K) and voltages near equilibrium (typically of the mV level) (Williamson *et al.*, 1990).

A more interesting (and thus, more complex) situation would be the non-equilibrium quantum transport regime. Recently, non-equilibrium electronic transport through quantum dots has been realized (Reed *et al.*, 1988). We present here a study of resonant tunneling through various quantum dot systems, and the bandstructure modeling necessary to understand the experimental electronic transport spectra.

TRANSPORT IN QUANTUM DOTS

The fabrication and measurements of these structures has been previously reported (Reed *et al.*, 1988). The approach essentially embeds a quasi-bound quantum dot between two quantum wire contacts. Creation of dots less than 500 Å is possible, though we will show that the appropriate range for the typical epitaxial structure and process used is in the range 1000 Å - 2500 Å in diameter. A SEM of a collection of these etched structures is seen in Fig. 1.

We have modeled the full screening potential of a quantum dot system taking into account the effects of lateral confinement. The model self-consistently obtains the electrostatic potential in a zero-current theory from Poisson's equation utilizing a Thomas-Fermi approximation for the electron density. The solution of the electrostatic problem then provides the potential responsible for lateral quantization of electron states. The radial bound states in the contacts provide the minima of the emitter and collector subbands. Likewise the discrete quantum well levels are obtained from a solution of the radial Schrodinger equation.

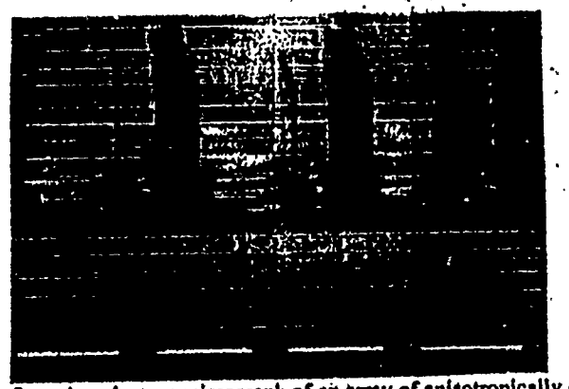


Fig. 1 Scanning electron micrograph of an array of anisotropically etched columns containing a quantum dot. The horizontal white marker is 0.5 micrometer. The smallest diameter column is 200 Å.

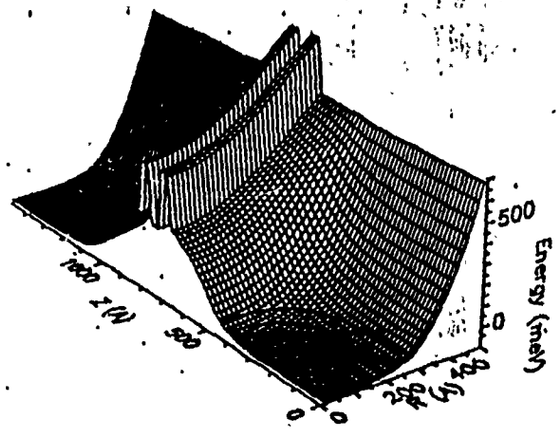


Fig. 2 Self-consistent 3D band diagram of the single quantum dot structure (Reed *et al.*, 1988) at equilibrium. The electron potential energy surface is plotted as a function of radius R and epitaxial z dimensions. The contours in the contact regions are the occupied laterally-defined subbands.

The equilibrium solution to the 3D screening problem of the experimental situation of Reed *et al.* (1988) is displayed in Fig. 2. The electron potential energy surface is plotted as a function of radius and epitaxial dimensions. The energy scale is defined relative to the Fermi energy, thus the potential at the external radius equals 0.7 V. The contours in the contact regions are the three occupied laterally-defined subbands that lie below the Fermi level. For clarity, the quantum dot energy levels are not drawn in Fig. 2.

It is clear that the lateral depletion has a dominant effect on lifting the double barrier structure significantly above the level previously determined only by the z-doping profile. The relavant quantum dot states, determined by solving the radial Schrödinger equation, arise from

the previous quantum well ground state ($n = 1$). The excited state ($n = 2$) quantum dot states are virtual.

It has been suggested (Bryant, 1989) that the observed quantum dot spectrum can be explained as resonances when the quantum dot states are biased through the emitter subband states with increasing device bias. To determine if this mechanism quantitatively explains the spectrum, we solve the 3D self-consistent-screening quantum-dot model at applied bias, to determine the variation of the emitter and quantum dot energy levels with applied voltage.

Figure 3 shows the crossings of the emitter subband levels with the quantum dot levels as a function of applied bias, transposed onto the 1.0 K slice of Reed *et al.*, (1988). There is general agreement between the experimental and predicted peak voltage positions, especially the anomalously large splitting of the first resonance. It should also be noted that the experimental measurement is current, which implies that an integration over the density of emitter states should be done for a strict comparison. The predicted 3'-3 transition appears to be absent in the spectrum, except for a very weak structure at 0.92 to 0.93 V. However, this is expected since the collector barrier becomes sufficiently low that the state becomes virtual. This has an important implication – verification that the observed resonances are due to states localized in the quantum dot.

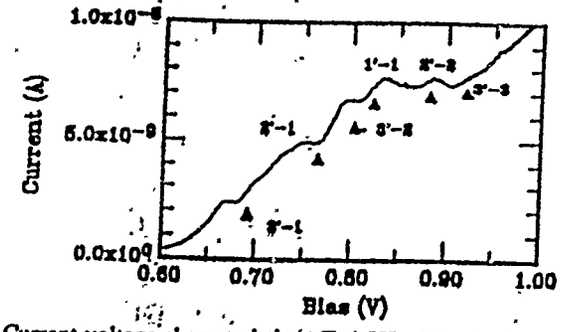


Fig. 3 Current-voltage characteristic at T=1.0K of the single quantum dot with predicted resonant peak positions due to crossing of the laterally confined emitter (n') and dot (n) states.

An additional corroboration of this spectroscopy and the peak indexing is found in the temperature dependence of the quantum dot peaks. When the temperature is raised to > 50 K, the three lowest voltage peaks disappear and the spectrum is dominated by the single 1'-1 transition. This is expected when the subband spacing is $< 3kT$. In the high temperature limit, the structure emulates an unconfined 1D resonant tunneling diode.

The observation of the momentum-nonconserving transitions shows that n is not a conserved quantity in this quantum dot system. This is due to the hourglass topography of the electron energy surface determined partly by the z-dependent doping profile. This absence of a n, n' (radial) selection rule is natural from the radially changing, cylindrically-symmetric geometry that breaks translational symmetry.

QUANTUM MOLECULES

The low temperature limit of these structures has been shown to be due to the thermal distribution of the laterally confined emitter states. An intriguing situation is to filter this distribution with another quantum dot, to achieve an input distribution as (theoretically) narrow as the probed states. The fabrication of such structures is relatively straightforward, in that the coupling of the two dots (i.e., a "molecule") is controlled epitaxially. The starting epitaxial material is a double quantum-well, triple-barrier structure designed to have approximately coincident resonances in one bias direction.

Figure 4 shows the current-voltage-temperature characteristics for this molecule structure (radius of 650 Å). NDR is clearly evident at low temperatures (peak-to-valley current ratios greater than 1.2:1 at $T < 20$ K), which persists up to 200 K. Figure 5 shows a comparison of the peak-to-valley ratios, as a function of temperature, for quantum dots and molecules. The thermal activation processes for the two systems are quite different, and allow the molecule structure to operate near room temperature.

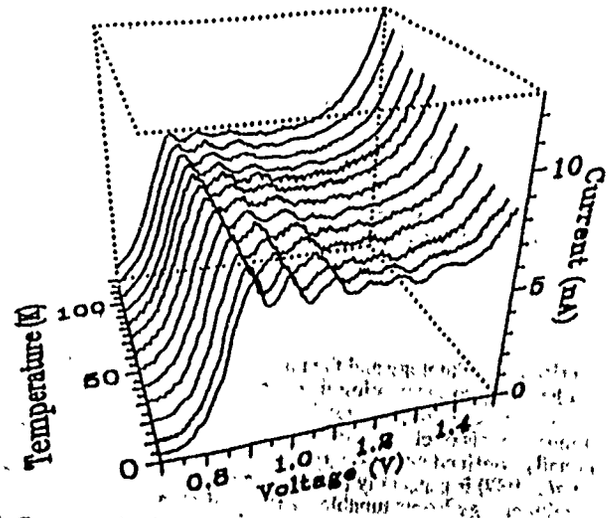


Fig. 4 Current-voltage-temperature characteristics of the quantum molecule.

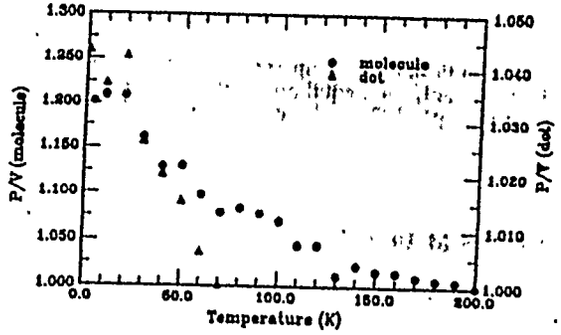


Fig. 5 Comparison of peak-to-valley of quantum dot and quantum molecule structures.

SUMMARY

Quantum dot structures provide not only a unique laboratory for the exploration of quantum transport through nanostructured semiconductors, but also provide a system with possible technological impact on electronic systems. Figure 6 shows a comparison of different electronic quantum systems as a function of the operating temperature and working voltage. The impact of operating quantum dot systems far from equilibrium is apparant.

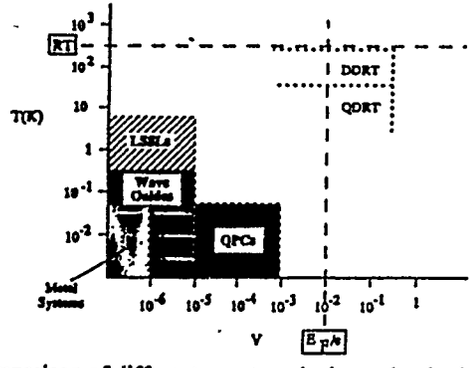


Fig. 6 Comparison of different quantum device technologies as a function of temperature and operating voltage. "LSSLs" stands for lateral surface superlattices; "QPCs" stands for quantum point contacts; "QDRT" stands for quantum dot resonant tunneling; "DDRT" stands for double dot (molecule) resonant tunneling.

We are indebted to our collaborators W. R. Frensley, R. J. Aggarwal, Y.-C. Kao, R. J. Matyi, T. M. Moore, and A. E. Wetsel. We thank R. T. Bate for constant encouragement and support, and R. K. Aldert, E. D. Pijan, D. A. Schultz, P. F. Stickney, and J. R. Thomason for technical assistance. This work was sponsored in part by ONR, ARO, and WRDC.

REFERENCES

Bryant, G. W., 1989, Phys. Rev. B, 39:3145.
Hansen, W. et al, 1990, Appl. Phys. Lett., 56:168.
Heinrich, H. et al, 1988, "Physics and Technology of Submicron Structures," Springer-Verlag, New York.
Reed, M. A. et al, 1988, Phys. Rev. Lett., 60:535.
Reed, M. A., and Kirk, W. P., 1989, "Nanostructure Physics and Fabrication," Academic Press, San Diego.
Van Wees, B. J. et al, 1989, Phys. Rev. Lett., 62:2523.
Williamson, J. G. et al, 1990, Phys. Rev. B, 41:1207.

12

TUNNELING BETWEEN CONSTRAINED DIMENSIONALITY SYSTEMS

E. Gornik, J.Smoliner, F.Hirler

Walter Schottky Institut
TU München, D-8046 Garching, Germany

INTRODUCTION

Systems of reduced dimensionality became a topic of increased interest during the last few years. To induce a one-dimensional system, a two-dimensional electron gas can be constrained to a thin stripe ("quantum wire") by etching processes or electrostatic confinement. This confinement results in an additional set of quantized states. The knowledge of these one dimensional (1D) subband energies is one of the basic requirements to understand the physical properties of quantum wires such as the quenching of the quantum hall effect (Roukes *et al.*, 1987), mobility modulations (Ismail *et al.*, 1989) or boundary scattering (Thornton *et al.*, 1989). One effect widely used to determine the 1D-energy levels is the magnetic depopulation of the 1D-subbands. This was done first by (Berggren *et al.*, 1986) on a split-gate field effect transistor structure. Assuming a parabolic electrostatic confinement, the influence of an additional magnetic field can be analyzed analytically. It was shown (Berggren *et al.*, 1986), that the additional magnetic field increases the 1D subband spacing, so that 1D-subbands are shifted above the Fermi energy if the magnetic field is increased. Consequently, these subbands are depopulated, resulting in an oscillating behavior of the magneto resistance. In contradiction to the 2D-case, a plot of the oscillation index versus inverse magnetic field (Landau-plot) is not linear and saturates at low magnetic fields. By fitting the experimental results, both the 1D electron concentration n_{1D} and the subband energies are determined. From this, the widths of the conducting channels can be calculated. Since the simple oscillator model is not always sufficient to explain the experimental results, modified potentials, density of state effects, and the influence of lifetime broadening due to scattering processes were also taken into account (Berggren *et al.*, 1988; Rundquist *et al.*, 1989). In addition, self consistent calculations were performed (Laux *et al.*, 1986,1988). Alternative to the method presented above, the subband energies can be extracted from Dingle-plots (Lakrimi *et al.*, 1989) or the nonlinear behavior of the magneto-conductance at extremely low magnetic fields (Thornton *et al.*, 1986). Far infrared transmission measurements were also used to study the properties of 1D-systems. Brinkop (1988) has observed far infrared absorption lines due to intersubband resonances in GaAs-GaAlAs quantum wires. The discrepancies between the far infrared data and the results obtained by magneto-transport measurements were attributed to large depolarization effects. Similar experiments were performed on InSb (Alsmeier *et al.*, 1988; Hansen *et al.*, 1987), where, due to the small effective mass, large 1D-subband spacings can be achieved easily. Demel *et al.*(1988,1989) detected plasmon modes in multi-layered quantum wires. Vertical transport through low-dimensional structures was also used to demonstrate the existence of quantized states. Quantum dots fabricated on double-barrier structures, showed a series of negative differential resistance regions due to resonant tunneling processes between the 0D-states (Randall *et al.*, 1988).

In this paper, we analyze the dependence of the electron concentration, the effective wire width, and the 1D-subband energies in quantum wires on an external gate voltage. Using magneto-transport experiments and tunneling spectroscopy, we determine the 1D-subband energies and study the influence of a magnetic field applied perpendicular to the sample. In most recent experiments, evidence of tunneling processes between quantum wires and a two-dimensional electron gas is found.

SAMPLE PREPARATION

The samples consist of an unintentionally p-doped GaAs layer grown on a semiinsulating substrate ($N_A < 1 \cdot 10^{14}$ cm^{-3}, followed by an undoped spacer (d=100Å) having an aluminum concentration of 35 %, and doped GaAlAs (d=100Å $N_D=1.2\cdot10^{18}$ cm^{-3}. The additional GaAs cap layer was also highly n-doped (d=200Å, $N_D=1.2\cdot10^{18}$cm^{-3}). From Shubnikov-de Haas measurements the 2D electron concentration n_{2D} was determined to be $3.6\cdot10^{11}$ cm^{-2}. In these samples, the 2D-channel is located very close to the surface, which is favorable for quantum wire fabrication by shallow mesa etching. The mobility (μ=80,000 cm^2/Vs), however, is therefore rather low, which is also due to the small spacer thickness.

Prior to quantum wire fabrication, bar-shaped mesas having an area of 1.5 mm x 0.3 mm were etched into the sample, and ohmic AuGe contacts were alloyed. Holographic photoresist-gratings having a period of 1000 nm, 450 nm, and 300 nm were fabricated on the mesas using a HeCd laser. The gratings were chemically wet etched into the sample. This procedure is controlled simply by measuring the sample resistance. As a last step, an aluminum gate is evaporated. A schematic view of the sample is shown in Fig.1.

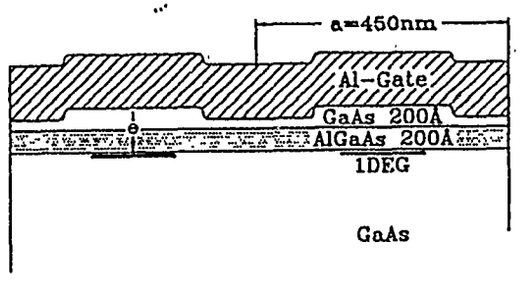


Fig. 1 Schematic view of a typical sample. The tunneling process between the gate and the 1D-wire is also indicated.

MAGNETO-TRANSPORT EXPERIMENTS

To characterize the samples, the differential magneto resistance (dR/dV_G) was measured as a function of the magnetic field and the applied gate voltage V_G. The measurements were carried out using a drain-source current of 1 μA and a modulation of V_G by 5 mV. Note that the modulation voltage has to be small in order to avoid an undesired filling of 1D-subbands. This method has an advantage in that the influence of the ungated region is completely eliminated. Figure 2a shows a Landau plot of the observed magneto resistance oscillations for a sample having a grating period of 450 nm at two different gate voltages. Both curves show a clear deviation from a linear behavior at low magnetic fields, demonstrating the existence of one-dimensional subbands. For a gate voltage of V_G=–50 mV, the deviation from the linear behavior is considerably larger than in the V_G=+100mV case. This is consistent with the fact that at more negative gate voltages, the subband energy differences -E are larger due to the stronger confining potential. In Figure 2b, a Landau plot is shown for

samples having grating periods of 1000 nm, 450 nm and 300 nm at V_G=–50 mV. As expected, the deviation from the linear behavior increases with decreasing grating period.

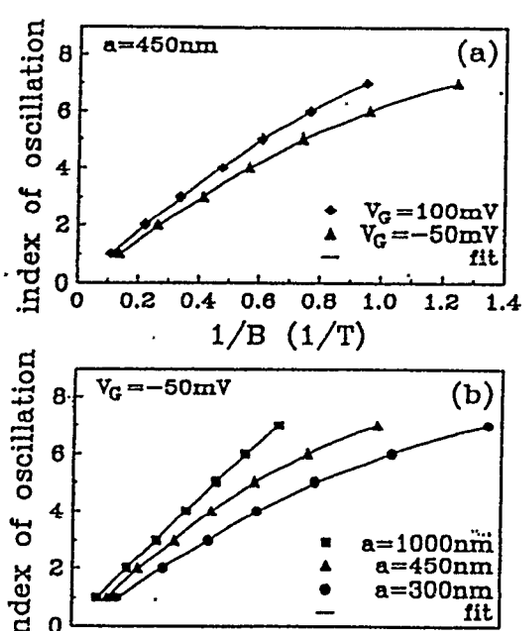


Fig. 2 (a) Landau plot for sample 1531/1b (a=450 nm) at V_G=–50 mV and V_G=100 mV. (b) Landau plot for samples having different grating periods. To show all data in one plot, the curves were scaled individually in 1/B.

To extract n_{1D}, -E, and the widths w of the conducting channels from the magneto transport data, a simple harmonic oscillator potential is assumed,

$$V(x) = \frac{1}{2} m^* \Omega_o^2 x^2 \quad , \tag{1}$$

resulting in equidistant 1D-subband energy levels in an external magnetic field given by

$$E_n = h\Omega(n+\tfrac{1}{2}) + \frac{h^2 k_y^2}{2m^*(B)} \quad , \tag{2}$$

where

$$\Omega^2 = \Omega_o^2 + \omega_c^2 \quad , \tag{3}$$

$$m^*(B) = \left(\frac{\Omega}{\Omega_o}\right)^2 m^* \quad , \tag{4}$$

and ω_c is the cyclotron frequency. The subbands are separated by -E= Ω.

The positions of the observed magneto resistance oscillations due to the depopulation of the 1D-subbands in the magnetic field were calculated using a two-parameter (n_{1D}, Ω_o) fit, taking the 1D-density of states into account. After Berggren (1988), the effective width w of the 1D-channels is calculated from $n_{1D}=n_{2D}\cdot w$. Calculating n_{1D} and n_{2D} from the corresponding density of states we obtain

$$w = 2\pi (n_{1D})^{1/3} \quad . \tag{5}$$

Note that w can also be determined from the Fermi energy in the quantum wire, E_F, using

$$E_F = eV_G\left(\frac{w}{2}\right) . \tag{6}$$

The resulting dependence of the widths of the 1D-channels as a function of the applied voltage is shown in Fig.3(a). The channel widths decreases slowly at more negative voltages. This can be explained by the fact that the strength of the confining potential is influenced by V_G. For the 1000 nm sample, 1D-behavior can only be observed if V_G exceeds -150 mV. As expected, from the increasing channel widths, the subband energy differences are decreasing at more positive values of V_G (Fig.3(b)). According to Poisson's equation, the electron concentration n_{1D} is rising with increasing V_G (Fig.3(c)). The different slopes of the curves are consistently explained if the variation of the channel width on the applied gate voltage is taken into account. Although the 1D electron concentration is changed drastically by an applied gate voltage (Fig.3(c)), the 1D subband energy differences stay almost unaffected. From this, we conclude that the confining potential is mainly space-charge determined. Thus, self-consistent effects can be neglected justifying the assumption of a harmonic oscillator potential.

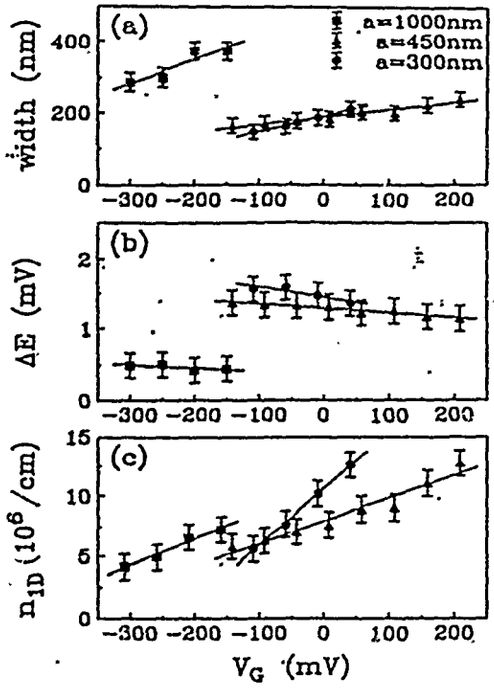


Fig. 3 Dependence of (a) w, (b) ΔE, and (c) n1D on the applied gate voltage for grating periods of 300 nm, 450 nm, and 1000 nm.

TUNNELING EXPERIMENTS

To measure the 1D-subband energies directly, tunneling experiments were performed, where the electrons tunnel between the 1D-channels and the metallic Schottky gate contact, which is indicated schematically in Fig.1. All dI/dV measurements were carried out using a 4-terminal conductance bridge (Christanell *et al.*, 1988) at a modulation frequency of 22 Hz to reduce parasitic capacitance effects and a modulation voltage of 0.25 mV to achieve a high resolution. Figure 4 shows the dI/dV curves for an a=450 nm sample at various magnetic fields and the dI/dV curve for an a=300 nm sample at B=0 T. A series of equidistant peaks is evident in all curves, where at B=0 T, the period of the observed structures is larger for the 300 nm sample. With increasing magnetic field, the period of the dI/dV structures is also increased. At negative gate voltages, the structures are more pronounced than at positive bias, reflecting the influence of V_G on the confining potential. At very low gate voltages, peaks are difficult to resolve due to the extremely low tunneling current. At too high V_G values, however, the structures in dI/dV are washed out since the tunneling resistance is no longer much larger than the resistance of the quantum wire. Tunneling experiments were also carried out on the a=1000 nm samples, but in this case, we were not able to detect any structures in dI/dV.

All these effects are well explained by the harmonic oscillator model, assuming that equidistant structures in the tunneling conductance will occur each time, the Fermi level in the metal contact crosses a subband in the quantum wires, which is shown schematically in the insert of Fig.4.

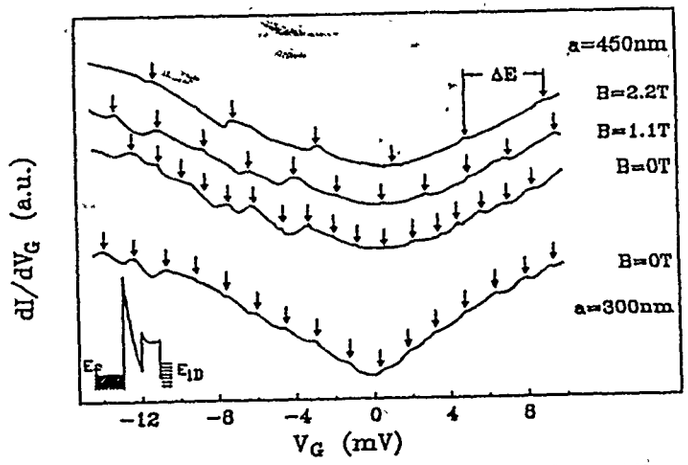


Fig. 4 Typical dI/dV characteristics for an a=300 nm and an a=450 nm sample at various magnetic fields. The expected subband resonance spacings are indicated by arrows. All curves are scaled to fit on the same plot.

Although the observed peaks in dI/dV are weak, the linewidth of the structures is only in the order of 0.5 meV. In analogy to tunneling processes into 2D-subbands, where dI/dV is proportional to the density of states (Smoliner *et al.*, 1988), we conclude that the small linewidth is a consequence of the sharp peaks in the density of states in the 1D-subbands. At higher voltages, the structure in dI/dV is washed out, because in this range, the tunneling resistance becomes comparable to the quantum wire resistance. In the small voltage range, where the potential distribution under the gate contact is homogeneous, the Fermi energy in the quantum wires varies only by 1.2 meV. As typical subband energies are also in the order of 1 meV, this rules out any possibility that the observed series of dI/dV oscillations are due to the systematic filling or depopulation of 1D-subbands by the applied gate voltage.

From magneto-transport results (Fig.3(b)), it is obvious that the influence of V_G on the subband spacings is relatively small. Therefore, the 1D-subband energy differences -E are approximately constant in the small range, where V_G is varied for the tunneling experiments. Thus, the subband spacing -E is equal to -E=e-V_G. A comparison between the (B=0 T) subband energy differences obtained by tunneling spectroscopy ($-E^{450nm}$=1.3±0.2 meV, $-E^{300nm}$=1.56±0.2 meV) and the magneto transport data $-E^{450nm}$=1.31 ± 0.18 meV, $-E^{300nm}$=1.47±0.18 meV) shows that they agree very well for both samples.

MAGNETO-TUNNELING EXPERIMENTS

To check the harmonic oscillator model, a magnetic field was applied perpendicular to the sample. As can be seen in Fig.4, the spacings between the structures in dI/dV are drastically increased, while the structures tend to wash out at higher magnetic fields. The increased 1D-subband energy differences are due to the magnetic contribution to the electrostatic potential, as shown in (1). As parasitic magneto-resistance effects under the gate contact cannot be compensated by the 4-terminal technique, this might be the reason why the

resonance peaks are washed out at higher magnetic fields. However, calculating the subband energies by the harmonic oscillator model and comparing them with the experimental results (Table I), one can see that they agree very well. This proves that both the confining electrostatic potential and the influence of an additional magnetic field is well described by a simple harmonic oscillator model.

Table I

Comparison between the experimental and theoretically calculated 1D-subband energy differences at various magnetic fields (a=450 nm)

B	-E (exp.)	-E (calc.)
0T	1.3±0.2meV	1.3 meV
1T	2.3±0.3meV	2.3 meV
2T	4.0±0.4meV	4.0 meV

TUNNELING PROCESSES BETWEEN QUANTUM WIRES AND A 2DEG

Most recent experiments were carried out on samples, where independent contacts can be achieved to an accumulation layer and an inversion layer which are only separated by a thin barrier. In these samples, we have studied a number of interesting 2D-2D tunneling processes, such as transitions between different Landau levels and the conservation of the canonical momentum (Smoliner *et al.*, 1989). In order to gain information about tunneling processes between a multiple quantum wire system and a two-dimensional electron gas, we have nanostructured these samples. In detail, the samples consist of an unintentionally p-doped GaAs layer grown on a semiinsulating substrate ($N_A<1\cdot10^{15}$ cm^{-3}), followed by an undoped spacer (d=50Å) ,doped GaAlAs (d=45Å, $N_D=4\cdot10^{18}$ cm^{-3}), another spacer (d=100Å) and n-doped GaAs (d=800Å, $N_D=1.2\cdot10^{15}$ cm^{-3}). An additional GaAs cap layer was highly n-doped (d=150Å, $N_D=6.4\cdot10^{18}$ cm^{-3}). Thus, we have a system where an accumulation layer and an inversion layer are separated by a barrier of only 200Å. From Shubnikov-de Haas measurements, it is found that only one subband is occupied both in the inversion layer and the accumulation layer. The electron concentrations are $n_s^{inv}=6.1\cdot10^{11}$ cm^{-2} and $n_s^{acc}=5.8\cdot10^{11}$ cm^{-2}, respectively. Ohmic contacts to the inversion layer were formed using a AuGe alloy. To structure the accumulation layer into quantum wires, 300 nm holographic photoresist gratings were made in the area of the tunneling contact. Then, the wires were wet chemically etched into the sample, in order to deplete the accumulation layer in the uncovered regions. After removing the photoresist, AuGe is evaporated on the total area of the tunneling contact. The ohmic contacts to the quantum wires were induced by a shallow diffusion process. A schematic view of the sample is shown in Fig.5(a), the resulting bandstructure is shown in Fig.5(b). Note, that in the etched areas, a leakage current can flow directly from the AuGe contacts into the 2D channel. However, these areas are surface depleted, so that the leakage current will be small compared to the tunneling current between the wires and the inversion layer.

In Fig.6, curve 1 shows the dI/dV characteristics of the unstructured sample, where large peaks are clearly observed in the considered voltage range. As shown elsewhere, these peaks are due to resonant tunneling processes between the two 2D-channels, which occur each time a subband in the accumulation layer matches a subband in the inversion layer. The peak at positive voltages is due to tunneling processes from the lowest subband in the inversion layer E_0^{inv} into the subband in the accumulation layer E_0^{acc}. The peaks at negative voltages are due to transitions from the accumulation layer E_0^{acc} into the higher subbands of the inversion layer E_0^{inv}.

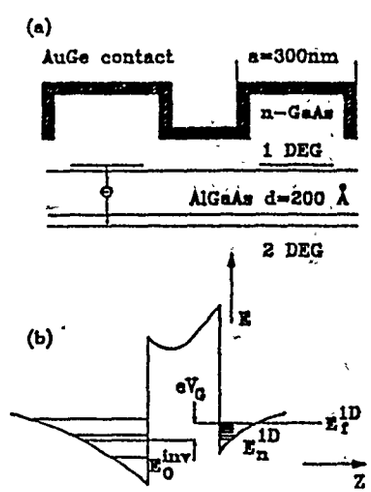


Fig. 5 (a) Schematic fiew of the sample. (b) The bandstructure of the 1D-2D tunneling sample.

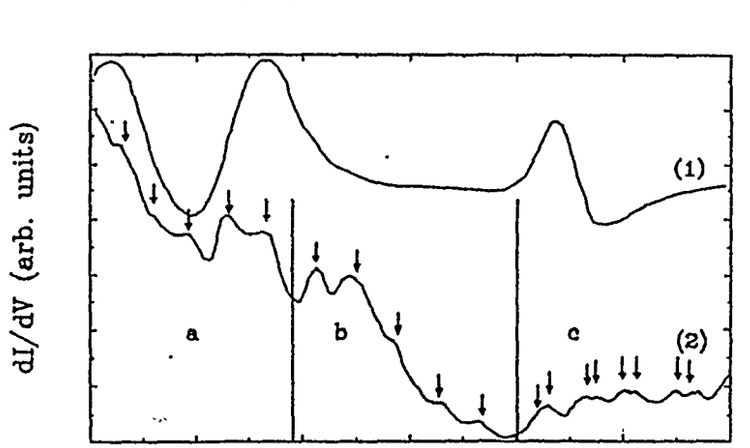


Fig. 6 The dI/dV characteristics of an unstructured sample (curve 1) and for the 1D-2D tunneling structure (curve 2).

The dI/dV characteristics of the structured sample, however, look completely different. As shown in curve 2, a large number of peaks can now be observed in the voltage range of the 2D-2D subband resonances. We attribute these structures to resonant tunneling processes between the 1D-subbands and the 2D-subbands in the inversion layer. The experimental results can be explained consistently if one assumes that in the 1D-channels, the subband spacing is -E=3.4±0.2 meV, and that 5 subbands are occupied in the 1D-channels. Then, the

peaks in the range of V_G=0,...,-22 mV (range b) are due to tunneling processes from E_5^{1D}...E_0^{1D} into the first subband in the inversion layer E_1^{2D}, the peaks between V_G=-22mV and V_G=-40mV (range a) are due to tunneling processes from E_5^{1D} ...E_0^{1D} into E_2^{2D}. At higher negative voltages, the structures in dI/dV cannot be identified unambiguously, due to the small 2D subband energy differences, tunneling from the 1D-subbands into different 2D subbands can occur simultaneously. At positive voltages (range c), we observe transitions between positive voltages (range c) we observe transitions between E_0^{inv} into the 1D subbands E_0^{1D}...E_4^{1D}. Note that in this case, the observed structures are always pairs of two peaks. This effect, however, can be explained if one assumes that structures in dI/dV in forward bias occur as E_0^{inv} matches a 1D subband or as E_F^{inv} matches a 1D-subband. This assumption is consistent with the momentum conservation rules, since resonant tunneling from 2D states into 1D states is allowed over the whole range from E=E_0^{inv}...E_0^{inv}+E_F^{inv}. Thus, structures in dI/dV will occur when E_0^{inv} crosses a 1D-subband, which means that tunneling into this subband is no longer possible, and when E_F^{inv} crosses a 1D subband, which means that electrons can tunnel into a new, additional 1D-subband.

In principle, the observed effects could also be due to 2D-2D tunneling processes between areas of different electron densities into the subbands of the lower channel. To check this, we have prepared a test structure, where the wires are etched only slightly into the sample. On this special sample, the dI/dV characteristic is completely different from the 1D-2D case, and the observed peaks were unambiguously identified with tunneling processes from the etched areas into the inversion layer, which have a lower electron concentration, and the non-etched areas, having a higher electron concentration. If one etches too deeply into the sample, considerable density modulation will also occur in the lower channel, resulting in a smearing out of the observed 1D-2D tunneling peaks.

Compared to conventional 1D structures, the measured subband energies are relatively large. This, however, may be due to the complicated space charge distribution in our sample. To check the results, calculations of the potential distribution by solving Poisson's equation in two dimensions are being tried. Further, magneto-tunneling experiments are also being performed to study transitions from the 1D-magneto-hybrid states into the Landau-levels in the 2D channel.

SUMMARY

In summary, we have analyzed the influence of a gate voltage and an additional magnetic field on the 1D-subband energies in quantum wires fabricated on GaAs-AlGaAs high electron mobility transistor structures. For comparison, we also have measured the 1D-subband energies directly using tunneling spectroscopy. Excellent agreement was found between the tunneling experiments and the values obtained from magneto-transport measurements, showing that both the confining potential and the influence of an additional magnetic field is well described by a simple harmonic oscillator model. In most recent experiments, we also found evidence of tunneling processes between quantum wires and a two-dimensional electron gas, where resonances between the 1D-subbands and the subbands of the 2DEG are well resolved as a series of peaks in derivative of the tunneling current.

ACKNOWLEDGEMENTS

This work was partially supported by the "Bundesministerium für Forschung und Technologie."

REFERENCES

Alsmeier, J., Sikorsky, Ch., and Merkt, U., 1988, Phys. Rev. B 37:4314.

Berggren, K. F., and Newson, D. J., 1986, Semicond. Sci. Technol. 1:327.

Berggren, K. F., Thornton, T. J., Newson, D. J., and Pepper, M., 1986, Phys. Rev. Lett. 57:1769.

Berggren, K. F., Roos, G., and van Houten, H., 1988, Phys. Rev. B, 17:10118.

Brinkop, F., Hansen, W., Kotthaus, J. P., and Ploog, K., 1988, Phys. Rev. B, 37:6547.

Christanell, R., and Smoliner, J., 1988, Rev. Sci. Instrum., 59:1290.

Demel, T., Heitmann, D., Grambow, P., and Ploog, K., 1988, Phys. Rev. B, 38:12732.

Demel, T., Heitmann, D., Grambow, P., and Ploog, K., 1989, Superlattices and Microstructures, 5:287.

Hansen, W., Horst, M., Kotthaus, J. P., Merkt, U., Sikorsky, Ch., and Ploog, K., 1987, Phys. Rev. Lett. 58:2586.

Ismail, K., Antoniadis, D. A., and Smith, H. I., 1989, Appl. Phys. Lett., 54:1130.

Lakrimi, M., Grassie, A. D. C., Hutchings, K. M., Harris, J.J., and Foxon, C. T., 1989, Semicond. Sci. Technol., 4:313.

Laux, S. E., and Stern, F., 1986, Appl. Phys. Lett., 49:91.

Laux, S. E., Frank, D. J., and Stern, F., 1988, Surface Science, 196:101.

Randall, J. N., Reed, M. A., Moore, T. M., Matyi, R. J., and Lee, J. W., 1988, J. Vac. Sci. Technol., B6:302.

Randall, J. N., Reed, M. A., , Matyi, R. J., and Moore, T. M., 1988, J. Vac. Sci. Technol., B6:1861.

Roukes, M. L., Scherer, A., Allen, S. J., Jr., Craighead, H. G., Ruthen, R. M., Beebe, E. D., and Harbison, J. P., 1987, Phys. Rev. Lett., 59:3011.

Rundquist, H. J., 1989, Semicond. Sci. Technol., 4:455.

Smoliner, J., Gornik, E., and Weimann, G.. 1988, Appl. Phys. Lett., 52:2136.

Smoliner, J., Gornik, E., and Weimann, G., 1989, Phys. Rev. B, 39:12937.

Smoliner, J., Demmerle, W., Berthold, G., Gornik, E., Weimann, G., and Schlapp, W., 1989, Phys. Rev. Lett., 63:2116.

Thornton, T. J., Roukes, M. L., Scherer, A., and van de Gaag, B. P., 1989, Phys. Rev. Lett., 63:2128.

Thornton, T. J., Pepper, M., Ahmed, H., Andrews, D., and Davies, G. J., 1986, Phys. Rev. Lett., 56:1198.

(13)

ADIABATIC TRANSPORT IN THE FRACTIONAL QUANTUM HALL EFFECT REGIME

C.W.J. Beenakker

Philips Research Laboratories
5600 JA Eindhoven
The Netherlands

INTRODUCTION

The Quantum Hall effect (QHE) is the phenomenon that the Hall conductance G_H is quantized in units of e^2/h, as expressed by the formula

$$G_H = \frac{p}{q}\frac{e^2}{h}\ , \tag{1}$$

(p and q being mutually prime integers). The *integer* QHE ($q = 1$) was discovered 10 years ago by Von Klitzing, Dorda, and Pepper (1980) in the two-dimensional electron gas (2DEG) confined to a Si inversion layer. The *fractional* QHE ($q > 1$ and odd) was first observed by Tsui, Störmer, and Gossard (1982) in the 2DEG at the interface of a GaAs-AlGaAs heterostructure. Microscopically the two effects are entirely different. The integer QHE, on the one hand, can be explained satisfactorily in terms of the states of noninteracting electrons in a magnetic field (the Landau levels). The fractional QHE, on the other hand, exists only because of electron-electron interactions (Laughlin, 1983a). Phenomenologically, however, the integer and fractional QHE are quite similar. In an *unbounded* 2DEG this similarity is understood from Laughlin's (1983b) general argument that: (1) The Hall conductance shows a plateau as a function of magnetic field (or Fermi energy) whenever the quasi-particle excitations in the bulk of the 2DEG are localized by disorder, and (2) The value of G_H on the plateau is precisely an integer multiple p of ee^*/h, where $e^* = e/q$ is the quasi-particle charge. (The product ee^* appears because one e is needed to change the unit of conductance from Amps per eV to Amps per V.) Theory and experiment on the QHE in an unbounded 2DEG have been reviewed in the books by Prange and Girvin (1987) and by Chakrahorty and Pietiläillen (1988).

In the past few years, a variety of experiments have uncovered a novel phenomenology of the QHE on short length scales. For example, in small sub-micron-size samples the QHE can occur in the absence of disorder (Wharam *et al.*, 1988; van Wees *et al.*, 1988) and can show deviations from precise quantization (Chang *et al.*, 1988). An anomalous quantization of the Hall conductance has been observed (van Wees *et al.*, 1989) in samplrs which are large but which contain a pair of closely spaced current and voltage contacts: Quantization of G_H then occurs at multiples of e^2/h determined by the properties of the contacts, rather than of the bulk 2DEG. Indeed, it has been possible in such an experiment to measure the fractional QHE in a 2DEG which by conventional measurements shows the integer effect (Kouwenhoven *et al.*, 1990a).

These anomalies are not easily understood within the conventional description of the QHE, which determines the quantized value of G_H from the charge of a quasi-particle excitation localized in the bulk of the 2DEG. One needs a description which can be applied to small samples without disorder and which explicitly includes the properties of the current and voltage contacts. For the integer QHE the Landauer-Büttiker (1988a) formalism provides such a description . A central concept in this formulation is the concept of an *edge channel*, which is the collection of states at the Fermi energy within a given Landau level. These states are extended along the edges of the 2DEG whenever the Fermi level lies between two Landau levels in the bulk. Many of the anomalies in the integer QHE can be understood as resulting from the absence of local equilibrium at the edge, which in turn is a consequence of the reduction of scattering between edge channels in a strong magnetic field (van Wees *et al.*, 1989a; Komiyama *et al.*, 1989; Alphenaar *et al.*, 1990). On short length scales the electron transport becomes fully *adiabatic*, i.e. without inter-edge-channel scattering. Edge channels in the integer QHE are defined in one-to-one correspondence with bulk Landau levels. This single-electron description is not applicable to the fractional QHE, which is fundamentally a many-body effect. In this article we review recent work towards a generalization of the concept of adiabatic transport in edge channels, with the aim of providing a unified description of anomalies in the integer and fractional QHE.

We first summarize, in the following section, the Landauer-Büttiker formalism for the integer QHE. Our generalization (Beenakker, 1990) to the fractional QHE is described in the third section and then applied to experiment. Two open problems are addressed in later sections. One is the question: "What charge does the resistance measure?" The second refers to an alternative generalized Landauer formula proposedby MacDonald (1990). We will argue that the appearance of both "electron" and "hole" channels in this formula implies a novel limitation to the accuracy of the fractional QHE. Much of the material in the present article is based on a review with a wider scope written in collaboration with H. van Houten (1990).

INTEGER EDGE CHANNELS

The Landauer-Büttiker formalism (Landauer, 1957, 1988; Büttiker, 1986, 1988b) is a linear response formalism which expresses the conductance (a nonequilibrium property) in terms of an equilibrium Fermi level property of the conductor. This property consists of transmission probabilities, between current and voltage contacts of propagating modes, which are a rational function of the Fermi energy. In a strong magnetic field in the QHE regime the propagating modes are extended along the edges of the conductor, because all Fermi level states in the bulk are localized. The Landauer-Büttiker formalism thus describes the integer QHE in terms or properties of edge states. We review this description in the presecnt section.

We restrict the discussion to the case of a smoothly varying potential energy landscape $V(x, y)$ in the 2DEG. The smoothness criterion is that V should vary by less than the Landau level separation $\hbar\omega_c = heB/m$ over a magnetic length $l_m = (h/eB)^{1/2}$ (which plays the role of the wavelength in a strong magnetic field B). In such a smooth potential the quantized cyclotron motion energy $(n+1/2)\hbar\omega_c$ (being the energy of the nth Landau level, $n = 1, 2, ...$) is a constant of the motion. The total energy E_F of an electron at the Fermi level is the sum of this Landau level energy and the energy E_G from the electrostatic potential,

$$E_G = E_F - (n-\tfrac{1}{2})\,\hbar\omega_c\ . \tag{2}$$

(The spin-splitting of the Landau levels by the Zeeman energy is ignored here, for simplicity.) The constancy of the Landau level index n for smooth V implies that the motion of the electron is along the equipotential $V(x, y) = E_G$. Classically, the center of the cyclotron orbit is guided along equipotentials by the combined effects of the Coulomb and Lorentz forces. Hence the name guiding center energy for E_G. The drift velocity v_{drift} of the orbit center (known as the guiding center drift) follows by balancing the Coulomb and Lorentz forces,

The wavefunctions of states at the Fermi level have an appreciable amplitude within l_m of the equipotentials at E_G. One can distinguish between extendedstates near the sample boundaries, and localized states encircling potential maxima and minima in the bulk, as illustrated in Fig. 1. The extended states with the same Landau level index n are referred to collectively as the nth edge *channel*. The edge channel with the smallest index n is closest to the sample boundary, because it has the largest E_G according to (2). This is seen more clearly in a crosssectional plot of $V(x, y)$ (Fig. 2). Notice that if the peaks and dips of the potential in the bulk have amplitudes below $\hbar\omega_c/2$, then only states with the highest Landau level index can exist in the bulk at the Fermi level.

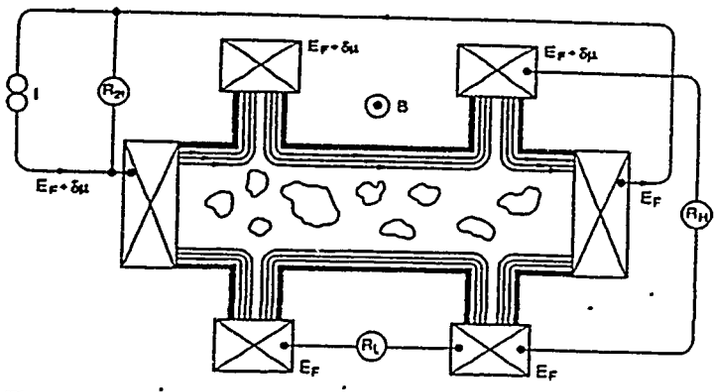


Fig. 1 Measurement configuration for the two-terminal resistance R_{2t}, the four-terminal Hall resistance R_H, and the longitudinal resistance R_L. The edge channels at the Fermi level are indicated, arrows point in the direction of motion of edge channels filled by the source contact at chemical potential $E_F + \delta\mu$. The current is equipartitioned among the edge channels at the upper edge, corresponding to the case of local equilibrium (Beenakker and van Houten, 1990).

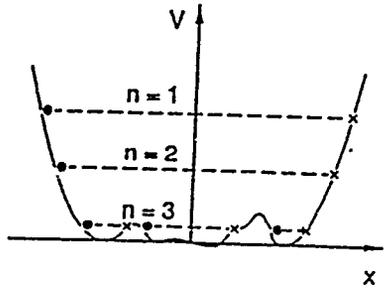


Fig. 2 Crosssection of the electrostatic potential $V(x, y)$, along a line perpendicular to the Hall bar in Fig. 1. The location of the states at the Fermi level is indicated by dots and crosses (depending on the direction of motion). The value of E_G for each n is indicated by the dashed line (Beenakker and van Houten, 1990).

The simplicity of the guiding center drift along equipotentials has been originally used in percolation theory (Kaziarinov and Luryi, 1982; Luryi and Kazarinov, 1983; Luryi, 1987; Iordansky, 1982; Trugman, 1983) of the QHE, soon after its experimental discovery. In this theory the existence of edge states is ignored, and the Hall resistance is expressed in terms of properties of extended states in the bulk of the sample. Since in equilibrium all Fermi level states in the bulk are localized in general, the percolation theory requires for its applicability a threshold electric field to create extended bulk states (it is thus not a linear response theory). A description of the QHE based on extended edge states and localized bulk states, as in Fig. 1, was first put forward by Halperin (1982), and further developed by several authors (Büttiker, 1988a; MacDonald and Streda, 1984; Apenko and Lozovik, 1985; Streda *et al.*, 1987; Jain and Kivelson, 1988). With the exception of Büttiker, these authors assume local equilibrium at the edge. In the presence of a chemical potential difference $\delta\mu$ between the edges, each edge channel can be shown to carry a current $(e/h)\delta\mu$, and thus to contribute e^2/h to the Hall conductance. The *equipartitioning* of current among the edge channels is characteristic for a local equilibrium. The total number of edge channels N at the Fermi level is equal to the number of bulk Landau levels below the Fermi level (because of the one-to-one correspondence between edge channels and bulk Landau levels). In this case of local equilibrium one thus has the usual integer QHE, $R_H = h/Ne^2$ with $R_H = 1/G_H$ the Hall resistance (we disregard for convenience of notation the two-fold spin degeneracy of each Landau level).

The Hall resistance R_H is a four-terminal resistance meaning that the voltage contacts are distinct from the contacts through which the current is passed. For the two-terminal resistance R_{2t} the voltage is measured between the current-carrying contacts. In the case of local equilibrium these two resistances are the same $R_H = R_{2t} = h/Ne^2$, as can be seen in Fig. 1. One can also read from Fig. 1 that the (four-terminal) longitudinal resistance R_L vanishes. The distinction betwren a longitudinal and Hall resistance is topological: A four-terminal resistance measurement gives R_H if current and voltage contacts alternate along the boundary of the conductor, and R_L if that is not the case. There is no need to further characterize the contacts in the case of local equilibrium at the edge.

If the edges are not in local equilibrium the measured resistance depends on the properties of the contacts. Büttiker (1988a) has developed the formalism to treat anomalies in the integer QHE due to the absence of local equilibrium when measured with non-ideal contacts. To illustrate this formalism we consider a situation in which the edge channels at the lower edge are in equilibrium at chemical potential E_F, while the edge channels at the upper edge are not in local equilibrium. The current at the upper edge is then not equipartitioned among the N edge channels. Let f_n be the fraction of the total current I which is carried by states above E_F in the nth edge channel at the upper edge, $I_n = f_n I$. The voltage contact at the lower edge measures a chemical potential E_F, regardless of its properties. The voltage contact at the upper edge, however, will measurec a chemical potential which depends on how it couples to each of the edge channels. The transmission probability T_n is the fraction of I_n which is transmitted through the voltage probe to a reservoir at chemical potential $E_F + \delta\mu$. The incoming current

$$I_{\rm in} = \sum_{n=1}^{N} T_n f_n I \ , \quad \text{with} \quad \sum_{n=1}^{N} f_n = 1 \ , \tag{4}$$

has to be balanced by an outgoing current

$$I_{\rm out} = \frac{e}{h}\delta\mu \sum_{n=1}^{N} T_n \tag{5}$$

of equal magnitude, so that the voltage probe draws no net current. [In (5) we have applied a sum rule to identify the total transmission probabilities of outgoing and incoming edge channels (Beenakker and van Houten, 1990)]. The requirement $I_{in} = I_{out}$ determines $\delta\mu$ and hence the Hall resistance $R_H = \delta\mu/eI$.

$$R_H = \frac{h}{e^2}\left(\sum_{n=1}^{N} T_n f_n\right)\left(\sum_{n=1}^{N} T_n\right)^{-1} . \qquad (6)$$

The Hall resistance has its regular quantized value $R_H = h/Ne^2$ only if *either* $f_n = 1/N$ *or* $T_n = 1$, for $n = 1, 2, \ldots N$. The first case corresponds to local equilibrium (the current is equipartitioned among the edge channels), the second case to an ideal contact (all edge channels are fully transmitted).

A non-equilibrium population of edge channels is generally the result of *selective backscattering*. Because edge channels at opposite edges of the sample move in opposite directions, backscattering requires scattering from one edge to the other. Selective backscattering of edge channels with $n \geq n_0$ is induced by a potential barrier across the sample, if its height is between the guiding center energies of edge channel n_0 and $n_0 - 1$ (recall that the edge channel with a larger index n has a smaller value of E_G). Selective backscattering can also occur naturallly in the absence of an imposed potential barrier. The edge channel with the highest index $n = N$ is selectively backscattered when the Fermi level approaches the energy $(N - 1/2)\hbar\omega_c$ of the Nth bulk Landau level. The guiding center energy of the Nth edge channel then approaches zero, and backscattering either by tunneling or by thermally activated processes becomes effective—but for that edge channel only, which remains almost completely decoupled from the other $N-1$ edge channels over distances as large as 250 μm (although on that length scale the edge channels with $n < N - 1$ have equilibrated to a large extent) (Komiyama *et al.*, 1989; Alphenaar *et al.*, 1990; van Wees *et al.*, 1989b; van Houten *et al.*, 1990).

We conclude this section by emphasizing that the edge channel formulation of the QHE by no means implies that the current flows within a few magnetic lengths of the edge. [This assumption would be untenable experimentally (Beenakker and van Houten, 1990).] The flow lines in Fig. 1 only show the location of the extended states at the equilibrium Fermi level. A determination of the spatial current distribution, rather than just the total current, requires consideration of all states below the Fermi level which acquire a net drift velocity because of the Hall field. Within the range of validity of a linear response theory, however, knowledge of the current distribution is not necessary to know the resistance.

FRACTIONAL EDGE CHANNELS

In this section we show, following Beenakker (1990), how the concept of an edge channel can be generalized to the fractional QHE, in the case of a smoothly varying electrostatic potential. This is the case of relevance for experiments on adiabatic transport in the fractional QHE. Our result is phrased in terms of a generalized Landauer formula, in which the edge channels contribute with a fractional weight. Hence the name "fractional" edge channels is used. The different generalization of the Landauer formula proposed by MacDonald (1990) is discussed in a later section.

Consider first the equilibrium state of the system. If the electrostatic potential energy $V(x, y)$ varies slowly in the 2DEG, the equilibrium density distribution $n(x, y)$ follows by requiring the local electrochemical potential $V(\mathbf{r}) + du/dn$ to have the same value μ at each point $\mathbf{r}$ in the 2DEG. Here du/dn is the chemical potential of the *uniform* 2DEG with density $n(\mathbf{r})$. It is a remarkable fact that the internal energy density $u(n)$ of a uniform interacting 2DEG in a strong magnetic field has downward cusps at densities $n = \nu_p Be/h$ corresponding to certain fractional filling factors ν_p. The chemical potential du/dn thus a discontinuity (an energy gap) at $\nu = \nu_p$, with du_p^+/dn and du_p^-/dn the two limiting values as $\nu \to \nu_p$. The size of the gap is the cyclotron energy $\hbar\omega_c$ when ν_p is an integer, and of the order of the Coulomb energy $e^2/\varepsilon l_m$ when ν_p is a fraction (ε is the dielectric constant). An order of magnitude for the energy gap is 10 meV at $B = 6$ T. As noted by Halperin (1983), when $\mu - V$ lies in the energy gap the filling factor is pinned at the value ν:

$$n = \nu_p eB/h \ , \ \text{if} \ \frac{du_p^-}{dn} < \mu - V < \frac{du_p^+}{dn} \ ,$$

$$\frac{du}{dn} + V(\mathbf{r}) = \mu \ , \ \text{otherwise} . \qquad (7)$$

Note that $V(\mathbf{r})$ itself depends on $n(\mathbf{r})$ and thus has to be determined selfconsistently from (7) taking the electrostatic screening in the 2DEG into account. We do not need to explicitly solve for $n(\mathbf{r})$ but can identify the edge channes from the following general considerations.

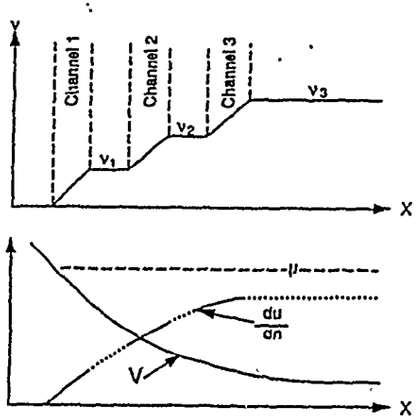


Fig. 3. Schematic drawing of the variation in filling factor ν, electrostatic potential V, and chemical potential du/dn, at a smooth boundary in a 2DEG. The dashed line in the bottom panel denotes the constant electrochemical potential $\mu = V + du/dn$. The dotted intervals indicate a discontinuity (energy gap) in du/dn, and correspond in the top panel to regions of constant fractional filling factor ν_p which spatially separate the edge channels. The width of the edge channel regions shrinks to zero in the integer QHE, since the compressibility χ of these regions is infinitely large in that case (Beenakker, 1990).

At the edge of the 2DEG the electron density decreases from its bulk value to zero. Equation (7) implies that this decrease is stepwise as illustrated in Fig. 3. The requirement on the smoothness of V for the appearance of a well-defined region at the edge in which ν is pinned at the fractional value ν_p, is that the change in V within the magnetic length l_m is small compared to the energy gap $du_p^+/dn - du_p^-/dn$. This ensures that the width of this region is large compared to l_m which is a necessary (and presumably sufficient) condition for the formation of the fractional QHE state. Depending on the smoothness of V, one thus obtains a series of steps at $\nu = \nu_p$ ($p = 1, 2, \ldots P$), as one moves from the edge toward the bulk. The series terminates in the filling factor $\nu_p = \nu_{bulk}$ of the bulk assuming that in the bulk the chemical potential $\mu - V$ lies in an energy gap. The regions of constant ν at the edge form bands extending along the conductor. These *incompressible bands* [in which the compressibility $\chi = (n^2 d^2u/dn^2)^{-1} = 0$] alternate with bands in which $\mu - V$ does not lie in an energy gap. The latter compressible bands (in which $\chi > 0$) may b identified as the *edge channels* of the transport problem as will be discussed below. To resolve a misunderstanding we note that the particular potential and density profile illustrated in Fig. 3 (in which the edge

channels have a non-zero width) assumes that the compressibility of the edge channels is not infinitely large—but that the analysis given below is independent of this assumption.

The conductance is calculated by bringing one end of the conductor in contact with a reservoir at a slightly higher electrochemical potential $\mu + \Delta\mu$. We are concerned with the linear response current, so that the electrostatic potential landscape $V(r)$ is kept at its equilibrium form. The resulting change in in electron density is

$$\Delta n = \left.\frac{\delta n}{\delta \mu}\right|_V \Delta\mu = -\left.\frac{\delta n}{\delta V}\right|_\mu \Delta\mu \ . \tag{8}$$

where δ denotes a functional derivative. In the second equality in (8) it has been assumed that n is a functional of $\mu - V$, by virtue of (7). In a strong magnetic field, this excess density moves along equipotentials with the guiding-center-drift velocity given by (3). The component v_{drift} of the drift velocity in the y direction (along the conductor) is

$$v_{drift} = \hat{y} \cdot \left(\frac{\nabla V \times B}{eB^2} \right) = -\frac{1}{eB}\frac{\partial V}{\partial x} \ . \tag{9}$$

The current density $j = -ev_{drift}\Delta n$ becomes simply

$$j = -\frac{e}{h}\Delta\mu \frac{\partial \nu}{\partial x} \ . \tag{10}$$

It follows from (10) that the incompressible bands of constant $\nu = \nu_p$ do not contribute to j. The reservoir injects the current into the compressible bands at one edge of the conductor only (for which the sign of $\partial\nu/\partial x$ is such that j moves away from the reservoir). The edge channel with index $p = 1, 2, \ldots P$ is defined as that compressible band which is flanked by incompressible bands at filling factors ν and ν_{p-1}. The outermost band from the center of the conductor, which is the $p = 1$ edge channel, is included by defining formally $\nu_0 = 0$. The arrangement of alternating edge channels and compressible bands is illustrated in Fig. 4a. Note that different edges may have a different series of edge channels at the same magnetic field value, depending on the smoothness of the potential V at the edge (which, as discussed above, determines the incompressible bands that exist at the edge). This is in contrast to the situation in the integer QHE, where a one-to-one correspondence exists between edge channels and bulk Landau levels. In the fractional QHE an infinite hierarchy of energy gaps exists, in principle, corresponding to an infinite number of possible edge channels—of which only a small number (corresponding to the largest energy gaps) will be realized in practice.

The current $I_p = (e/h)\,\Delta\mu\,(\nu_p - \nu_{p-1})$ injected into edge channel p by the reservoir follows directly from (10), on integration over x. The total current I through the conductor is

$$I = \sum_{p=1}^{P} I_p T_p \ ,$$

if a fraction T_p of the injected current I is transmitted to the reservoir at the other end of the conductor (the remainder returning via the opposite edge). For the conductance $G = eI/\Delta\mu$ one thus obtains the generalized Landauer formula for a two-terminal conductor (Beenakker, 1990)

$$G = \frac{e^2}{h} \sum_{p=1}^{P} T_p \, \Delta\nu_p \ , \tag{11}$$

which differs from the usual two-terminal Landauer formula by the presence of the fractional weight factors $\Delta\nu_p = \nu_p - \nu_{p-1}$. In the integer QHE, $\Delta\nu_p = 1$ for all p, so that (11) reduces to the Landauer formula with unit weight factors.

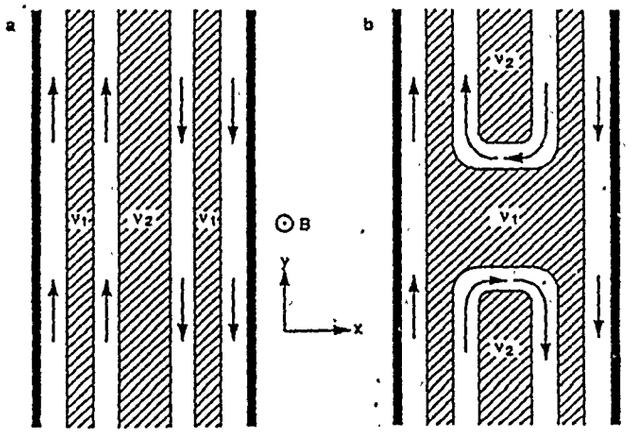


Fig. 4 Schematic drawing of the incompressible bands (hatched) of fractional filling factor ν_p, alternating with the edge channels (arrows indicate the direction of electron motion in each channel). (a) A uniform conductor; (b) A conductor containing a barrier or reduced filling factor (Beenakker, 1990).

A multi-terminal generalization of (11) for a two-terminal conductor is easily constructed, following Büttiker (1986, 1988b):

$$I_\alpha = \frac{e}{h}\nu_\alpha\mu_\alpha - \frac{e}{h}\sum_\beta T_{\alpha\beta}\mu_\beta \ , \tag{12a}$$

$$T_{\alpha\beta} = \sum_{p=1}^{P} T_{p,\alpha\beta}\,\Delta\nu_p \ . \tag{12b}$$

Here I_α is the current in lead α, connected to a reservoir at electrochemical potential μ_α, and with fractional filling factor ν_α. Equation (12b) defines the transmission probability $T_{\alpha\beta}$ from reservoir β to reservoir α (or the reflection probability, for $\alpha = \beta$), in terms of a sum over the generalized edge channels in lead β. The contribution from each edge channel $p = 1, 2, \ldots P\beta$ contains the weight factor $\Delta\nu_p = \nu_p - \nu_{p-1}$ and the fraction $T_{p,\alpha\beta}$ of the current injected by reservoir β into the pth edge channel of lead β which reaches reservoir α. Apart from the fractional weight factors, the structure of (12) is the same as that of the usual Büttiker formula (1986, 1988b).

Applying the generalized Landauer formula (11) to the ideal conductor in Fig. 4a, where $T_p = 1$ for all p, one finds the quantized two-terminal conductance

$$G = \frac{e^2}{h}\sum_{p=1}^{P}\Delta\nu_p = \frac{e^2}{h}\nu_p \ . \qquad (13)$$

The four-terminal Hall conductance G_H has the same value, because each edge is in local equilibrium. In the presence of disorder this edge channel formulation of the fractional QHE is generalized in an analogous way as in the integer QHE, by including localized states in the bulk. In a smoothly varying disorder potential these localized states take the form of circulating edge channels, as in Fig. 1. In this way the filling factor of the bulk can locally deviate from ν_p without a change in the Hall conductance, leading to the formation of a plateau in the magnetic field dependence of G_H. In a narrow channel, localized states are not required for a finite plateau width, because the edge channels make it possible for the chemical potential to lie in an energy gap for a finite magnetic field interval. The Hall conductance then remains quantized at ν_p (e^2/h) as long as $\mu - V$ in the bulk lies between du_p^+/dn and du_p^-/dn .

EXPERIMENTS

We now apply the generalized Landauer formula (11) to some recent experiments on adiabatic transport in the fractional QHE regime. Consider first a conductor containing a potential barrier. The potential barrier corresponds to a region of reduced filling factor $\nu_{P,\min} = \nu_{\min}$ separating two regions of filling factor $\nu_{P,\max} = \nu_{\max}$. The arrangement of edge channels and incompressible bands is illustrated in Fig. 4b. We assume that the potential barrier is sufficiently smooth that scattering between the edge channels at opposite edges can be neglected. All transmission probabilities are then either zero or one: $T_p = 1$ for $1 < p < P_{\min}$, and $T_p = 0$ for $P_{\min} < p < P_{\max}$. Equation (11) then tells us that the two-terminal conductance is

$$G = \frac{e^2}{h}\nu_{\min} \ . \qquad (14)$$

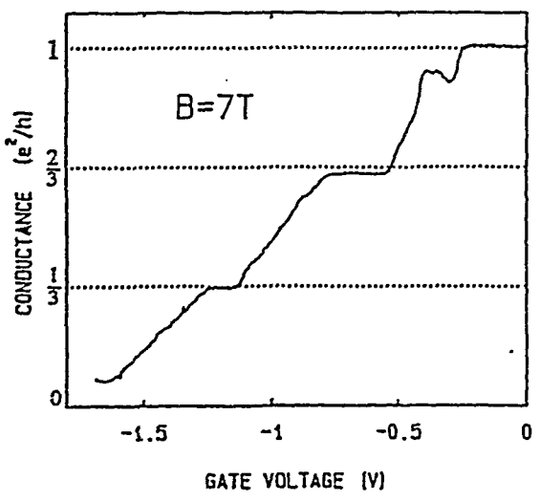


Fig. 5 Two-terminal conductance of a constriction containing a potential barrier as a function of the voltage on the split gate defining the constriction at a fixed magnetic field of 7 T. The conductance is quantized according to (14) (Kouwenhoven *et al.*, 1990b).

In Fig. 5 we have reproduced experimental data by Kouwenhoven *et al.* (1990b) on the fractionally quantized two-terminal conductance of a constriction containing a potential barrier. The constriction (or point contact) is defined by a split gate on top of a GaAs–AlGaAs heterostructure. The conductance in Fig. 5 is shown for a fixed magnetic field of 7 T as a function of the gate voltage. Increasing the negative gate voltage increases the barrier height, thereby reducing G below the Hall conductance corresponding to $\nu_{\max} = 1$ in the wide 2DEG. The curve in Fig. 5 shows plateaus corresponding to $\nu_{\min} = 1$, 2/3, and 1/3 in (14). The 2/3 plateau is not exactly quantized, but is too low by a few percent. The constriction width on this plateau is estimated (Kouwenhouven *et al.*, 1990) to be $W = 500$ nm, which is a factor of 50 larger than the magnetic length at $B = 7$ T. It would seem that scattering between fractional edge channels at opposite edges (necessary to reduce the conductance below its quantized value) can only occur via states in the bulk for this large ratio of W/l_m.

Timp *et al.* (1989) have measured the four-terminal Hall conductance in a narrow cross geometry ($W = 90$ nm). They find, in addition to quantized plateaus near 1/3, 2/5, and 2/3 × e^2/h, also a plateau-like feature around $1/2 \times e^2/h$. (This even-denominator fraction is special because it is not observed as a Hall plateau in a bulk 2DEG.) Notice, however, that the 500 nm wide constriction of Fig. 5 has a conductance which is featureless at e^2/h. A narrower constriction ($W = 150$ nm) studied by Kouwenhoven *et al.* (1990b) shows more fluctuations on the plateaus at 1/3 and $2/3 \times e^2/h$, but no plateau-like feature at $1/2 \times e^2/h$. The origin of the difference between these two experiments remains to be understood .

A four-terminal measurement of the fractional QHE in a conductor containing a potential barrier can be analyzed by means of (12). The longitudinal resistance R_L of the barrier (measured by two adjacent voltage probes, one at each side of the barrier) is given by

$$R_L = \frac{h}{e^2}\left(\frac{1}{\nu_{\min}} - \frac{1}{\nu_{\max}}\right) \ . \qquad (15)$$

This result follows from (12) provided *either* the edge channels transmitted across the barrier have equilibriated with the extra edge channels available outside the barrier region, *or* the voltage contacts are ideal, i.e. they have unit transmission probability for all fractional edge channels. In the case of the integer QHE, (15) (with ν integeer) was derived some time ago by Van Houten *et al.* (1988) and (independently) by Büttiker (1988a), and was found to be in agreement with experiments (van Houten *et al.*, 1988; Haug *et al.*, 1988; Washburn *et al.*, 1988). Chang and Cunningham (1989) have measured R_L in the fractional QHE, using a 1.5 μm wide 2DEG channel with a gate across a segment of the channel. Contacts to the gated and ungated regions allowed $\nu_{\min}$ and $\nu_{\max}$ to be determined independently. Equation (15) was found to hold to within 0.5% accuracy.

Adiabatic transport in the fractional QHE has been demonstrated by the selective population and detection of fractional edge channels, achieved by means of barriers in two closely separated current and voltage contacts. The geometry is illustrated in Fig. 6a. It is essentially the same as the geometry employed by van Wees *et al.* (1989a) for the selective population and detection of Landau levels in the integer QHE. Figure 6b illustrates the arrangement of fractional edge channels and incompressible bands for the case that the chemical potential lies in an energy gap for the bulk 2DEG (at $\nu = \nu_{\text{bulk}}$), as well as for the two barriers (at ν_I and ν_V for the barrier in the current and voltage lead, respectively). Adiabatic transport is assumed over the barrier, as well as from barrier I to barrier V (for the magnetic field direction indicated in Fig. 6). Equation (12) for this case reduces to

$$I = \frac{e}{h}\nu_I\mu_I \ , \quad 0 = \frac{e}{h}\nu_V\mu_V - \frac{e}{h}\min(\nu_I, \nu_V)\,\mu_I \ , \qquad (16)$$

so that the Hall conductance $G_H = eI/\mu_V$ becomes

$$G_H = \frac{e^2}{h} \max(\nu_I, \nu_V) \leq \frac{e^2}{h} \nu_{bulk} \,. \qquad (17)$$

The quantized Hall plateaus are determined by the fractional filling factors of the current and voltage leads, not of the bulk 2DEG.

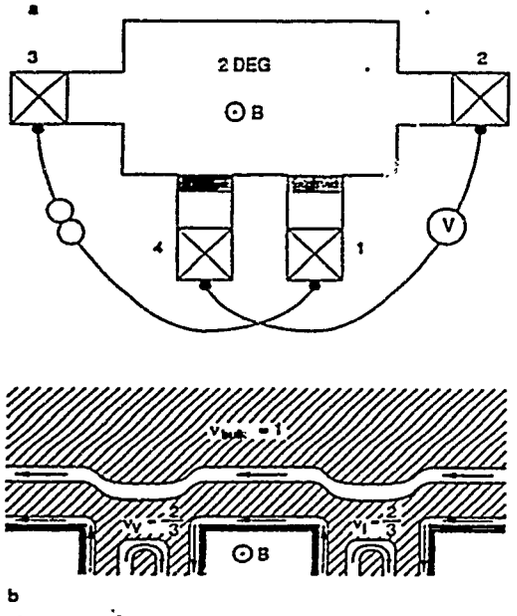


Fig. 6 (a) Schematic drawing of the experimental geometry of Kouwenhoven *et al.* (1990a). The crossed squares are contacts to the 2DEG. One current lead and one voltage lead contain a barrier (shaded) of which the height can be adjusted by means of a gate (not drawn). The current I flows between contacts 1 and 3, the voltage V is measured between contacts 2 and 4. (b) Arrangement of incompressible bands (hatched) and edge channels near the two barriers. In the absence of scattering between the two fractional edge channels one would measure a Hall conductance $G_H = I/V$ which is fractionally quantized at $2 \times e^2/h$ although the bulk has unit filling factor (Beenakker, 1990).

Kouwenhoven *et al.* (1990) have demonstrated the selective population and detection of fractional edge channels in a device with a 2 μm separation of the gates in the current and voltage leads. The gates extended over a length of 40 μm along the 2DEG boundary. In Fig. 7 we reproduce one of their experimental traces. The Hall conductance is shown for a fixed magnetic field of 7.8 T as a function of the gate voltage (all gates being at the same voltage). As the barrier heights in the two leads are increased, the Hall conductance decreases from the bulk value e^2/h to the value $2 \times e^2/h$ determined by the leads—in accord with (17). A more general formula for G_H valid also in between the quantized plateaus is shown in this latter work to be in quantitative agreement with the experiment.

OPEN PROBLEMS

What charge does the resistance measure?

The fractional quantization of the conductance in the experiments discussed above is understood as a consequence of the fractional weight factors in the generalized Landauer formula (11). These weight factors $\Delta\nu_p = \nu_p - \nu_{p-1}$ are *not* in general equal to e^*/e, with e^* the fractional charge of the quasiparticle excitations of Laughlin's incompressible state. The reason for the absence of a one-to-one correspondence between $\Delta\nu_p$ and e^* is that the edge channels themselves are not incompressible. The transmission probabilities in (11) refer to charged "gapless" excitations of the edge channels, which are not identical to the charge e^* excitations above the energy gap in the incompressible bands [the latter charge might be obtained from thermal activation measurements as suggested by Clark *et al.* (1988)]. It is an interesting and (to date) unsolved problem to determine the charge of the edge channel excitations. Kivelson and Pokrovsky (1989) have suggested performing tunneling experiments in the fractional QHE regime for such a purpose, by using the charge decpendence of the magnetic length $(\hbar/eB)^{1/2}$ (which determines the penetration of the wave function in a tunnel barrier, and hence the transmission probability through the barrier). Alternatively, one could use the h/e periodicity of the Aharanov-Bohm magnetoresistance oscillations as a measure of the edge channel charge. Simmons *et al.* (1989) find that the characteristic field scale of quasiperiodic resistance fluctuations in a 2 μm wide Hall bar increases from 0.016 T ± 30% near $\nu = 1, 2, 3, 4$ to 0.05 T ± 30 % near $\nu = 1/3$. This is suggestive of a reduction in charge from e to $e/3$, but not conclusive since the area for the Aharonov-Bohm effect is not well-defined in a Hall bar.

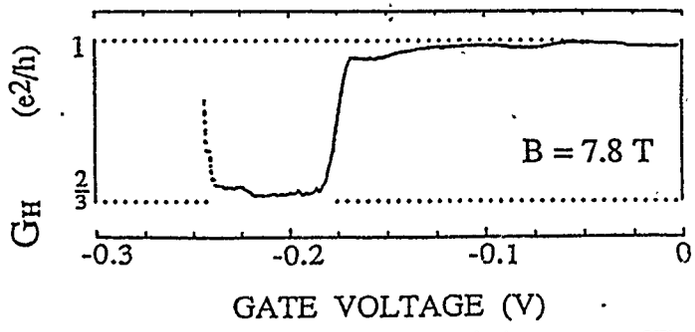


Fig. 7 Anomalously quantized Hall conductance in the geometry of Fig. 6, in accord with (17) ($\nu_{bulk} = 1$, $\nu_I = \nu_V$ decreases from 1 to 2/3 as the negative gate voltage is increased). The temperature is 20 mK. The rapidly rising part (dotted) is an artifact due to barrier pinch-off (Kouwenhoven *et al.*, 1990a).

Electron and hole channels

MacDonald (1990) has, independently of Beenakker (1990), proposed a different generalized Landauer formula for the fractional QHE in a smooth electrostatic potential. The difference with (11) is the weight factors in MacDonald's formula can take on both positive and negative values—corresponding to electron and hole channels, respectively. In the case of local equilibrium at the edge, the sum of weight factors is such that the two formulations give identical results. The results differ in the absence of local equilibrium, if fractional edge channels are selectively populated and detected. For example, MacDonald predicts a negative longitudinal resistance in a conductor at filling factor $\nu = 2/3$ containing a segment at $\nu = 1$. Another implication, as we understand it, is that the two-terminal conductance G of a conductor at $\nu_{max} = 1$ containing a potential barrier at filling factor ν is reduced to $1/3 \times e^2/h$ if $\nu_{min} = 1/3$

[in accord with (14)], but remains at $1 \times e^2/h$ if $\nu_{min} = 2/3$. That this is not observed experimentally (see Fig. 5) could be due to inter-edge channel scattering, as argued by MacDonald. The experiment by Kouwenhoven *et al.* (1990a) (Fig. 7), however, is apparently in the adiabatic regime, and was interpreted in Fig. 6 in terms of an edge channel of weight 1/3 at the edge of a conductor at $\nu = 1$. In MacDonald's formulation, the conductor at $\nu = 1$ has only a single edge channel of weight 1. This would have to be reconciled with the experimental observation of quantization of the Hall conductance at $2/3 \times e^2/h$. What is needed is a theory which allows one to introduce edge channels not only for the case of a smooth potential at the edge [considered in Beenakker (1990) and MacDonald (1990)], but also for an abrupt confinement. Such a theory exists for the integer QHE but not yet for the fractional effect.

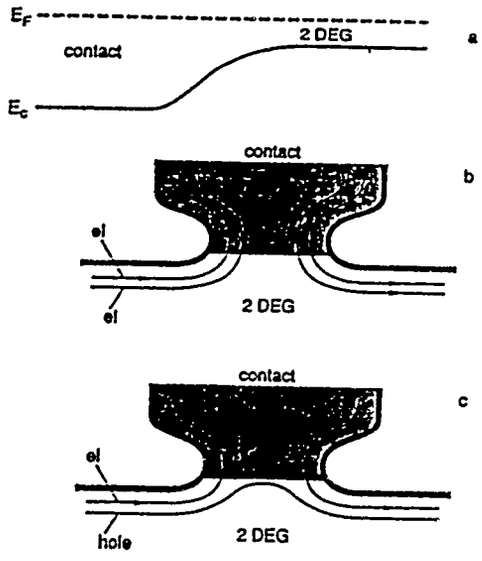


Fig. 8 (a) Schematic drawing of the bottom of the conduction band E_c and the Fermi energy E_F at the transition from a low-density to a high-density region in a 2DEG. (b,c) Top view of a 2DEG near a contact modecled hy a high-density region (shaded) as in (a). The contact is ideal (i.e. fully transmitting) for electron channels (b), but not for hole channels (c). The arrows indicate the current-carrying edge channels. This figure illustrates why a contact is effective in establishing local equilibrium among electron channels, but not among electron and hole channels. In case (c) one would measure anomalies in the Hall conductance, due to the abscence of local equilibrium.

The presence of both positive and negative weights in a generalized Landauer formula has an interesting implication for the accuracy of the fractional QHE. As we discussed above for the integer QHE, accurate quantization of the Hall resistance requires either a local equilibrium at the edge, or ideal contacts (i.e., contacts which fully transmit all available edge channels). The reduction of interedge channel scattering in strong magnetic fields leads to deviations from local equilibrium (i.e., the current is not equipartitioned among the edge channels). Ideal contacts then become necessary for accurate quantization. A contact is essentially a region with a high electron density connected to the low-density electron gas. see Fig. 8a. An ideal contact is realized by a smooth increase in density in the contact region, so that the edge channels in the 2DEG are transmitted adiabatically into the contact. The contact then includes a local equilibrium by redistributing the current among the edge channels. This is illustrated in Fig. 8b by means of arrows, which indicate the current-carrying edge channels: One incoming edge channel carries the current, whereas both outgoing edge channels are populated by the contact. These considerations for the integer QHE carry over completely to the fractional edge channels described above. However, if both electron and hole channels are present in the 2DEG, then the situation is different. A hole channel is reflected on approaching a region with a smoothly increasing electron density (MacDonald, 1990). In other words, a contact can not be "ideal" (i.e. fully transmitting) for both electron and hole channels. As shown in Fig. 8c, the contact is then not able to redistribute the current among the edge channels. The accuracy of the fractional QHE would thus be limited by the extent to which inelastic scattering is effective in establishing a local equilibrium between electron and hole channels—regardless of the ideality of the contacts. This conclusion is of importance not only for adiabatic transport in the fractional QHE, but for other situations as well in which coexisting electron and hole channels are believed to occur. One example is the integer QHE in a periodic potential discussed by MacDonald. Another is the integer QHE in parallell conducting electron and hole gases, presellt in certain semiconductor hetcrostructures.

REFERENCES

Alphenaar, B. W., McEuen, P. L., Wheeler, R. G., and Sacks, R. N., 1990, *Phys. Rev. Lett.*, 64:677.

Apenko, S. M., and Lozovik, Yu. E., 1985, *J. Phys. C*, 18:1197.

Beenakker, C. W. J., 1990, *Phys. Rev.Lett.*, 64:216.

Beenakker, C. W. J., and van Houten, H., 1990, "Quantum Transport in Semiconductor Nanostructures," *in* "Solid State Physics," Ed. by H. Ehrenreich and D. Turnbull, (Academic Press, New York, in press).

Büttiker, M.,1986, *Phys. Rev. Lett.*, 57:1761.

Büttiker, M., 1988a, *Phys. Rev. B*, 38:9375.

Büttiker, M.,1988b, *IBM J.Res. Dev.* 32:317.

Chakraborty, T., and Pietiläinen, P., 1988, "THe Fractional Quantum Hall Effect," (Springer, Berlin).

Chang, A. M., 1990, *Solid State Comm.* (to be publishcd).

Chang, A. M., and Cunningham, J. E., 1989, Solid State Comm., 72:651.

Chang, A. M., Timp, G., Cunningham, J. E., Mankiewich, P. M., Behringer, R. E., and Howard, R.E., 1988, *Solid State Comm.*, 76:769.

Clark, R. G., Mallett, J. R., Haynes, S. R., Harris, J. J., and Foxon, C. T., 1988, *Phys. Rev. Lett* ., 60:1747.

Fontein, P. F., Kleinen, J .A., Hendriks, P., Blom, F. A. P., Wolter, J. H., Lochs, H. G. M., Driessen, F. A. J. M., Giling, L. J., and Beenakkecr, C. W. J., 1990, submitted to *Phys. Rev. B.*

Halperin, B. I., 1982, *Phys. Rev. B*, 25:2185.

Halperin, B. I., 1983, *Helv. Phys. Acta*, 56:75.

Haug, R. J., MacDonald, R. J., Streda, P., and von Klitzing, K., 1988, *Phys. Rev. Lett.*, 361:2797.

Iordansky, S. V., 1982, *Solid State Comm.* 4:1.

Jain, J. K., and Kivelson, S. A., 1988, *Phys. Rev. B*, 37:4276.

Kazarinov, R. F., and Luryi, S., 1982, *Phys. Rev. B*, 25:7626.

Kivelson, S. A., and Pokrovsky, V. L., 1989, *Phys. Rev. B*, 40:1373.

Komiyama, S., Hirai, H., Sasa, S., and Hiyamizu, S., 1989, Phys. Rev. B, 40:12566.

Kouwenhoven, L.P., van Wees, B. J., van der Vaart, N. C., Harmans, C. J. P. M., Timmering, C. E., and Foxon, C. T., 1990a, *Phys. Rev.Lett.*, 64:685.

Kouwenhoven, L. P., van Wees, B. J., van der Vaart, N. C., Harmans, C. J. P. M., Timmering, C. E., and Foxon, C.T., 1990b, unpuhlished.

Landauer, R., 1957, *IBM J. Res.Dev.*, 1:223.

Landauer, R., 1988, *IBM J. Res.Dev.*, 32:306.
Laughlin, R. B., 1983a, *Phys. Rev. Lett.*, 50:1395.
Laughlin, R. B., 1983b, *Phys. Rev. B.*, 27:3383.
Luryi,S., 1987, *in* "High Magnetic Fields in Semiconductor Physics," G. Landwehr, ed. (Springer, Berlin).
Luryi, S., and Kazarinov, R. F., 1983, *Phys. Rev. B*, 27:1386.
MacDonald, A. H., 1990, *Phys. Rev. Lett.*, 64:220.
MacDonald, A. H., and Streda, P., 1984, *Phys. Rev. B*, 29:1616.
Prange, R. E., and Girvin, S. M., eds., 1987, "The Quantum Hall Effect," (Springer, New York).
Simmons, J. A., Wei, H. P., Engel, L. W., Tsui, D. C., and Shayegan, M., 1989, *Phys. Rev. Lett.*, 63:1731.
Streda, P., Kucera, J., and MacDonald, A. H., 1987, *Phys. Rev. Lett.*, 59:1973.
Timp, G., Behringer, R. E., Cunningham, J. E., and Howard, R. E., 1989, Phys. Rev. Lett., 63:2268.
Trugman, S.A., 1983, *Phys. Rev. B*, 27:7539.
Tsui, D. C., Störmer, H. L., andGossard, A. C., 1982, *Phys. Rev. Lett.*, 48:1559.
van Houten, H., Beenakker, C. W. J., van Loosdrecht, P. H. M., Thornton, T. J., Ahmed, H., Pepper, M., Foxon, C. T., and Harris, J. J., 1988, *Phys. Rev. B*, 37:8534.
van Houten, H., Beenakker, C. W. J., and van Wees, B. J., 1990, "Quantum Point Contacts," *in* "Semiconductors and Semimetals," M.A. Reed, ed. (Academic Press, New York, in press).
van Wees, B. J., Kouwenhoven, L. P., van Houten, H., Beenakker, C. W.J ., Mooij, J. E., Foxon, C. T., and Harris, J. J., 1988, *Phys. Rev. B*, 38:3625.
van Wees, B. J., Willems, E. M. M., Harmans, C. J. P. M., Beenakker, C. W. J., van Houten, H., Williamson, J. G., Foxon, C. T., and Harris, J.J., 1989a, *Phys. Rev. Lett.* 62:1181.
van Wees, B. J., Willems, E. M. M., Kouwenhoven, L. P., Harmans, C. J. P. M., Williamson, J. G., Foxon, C. T., and Harris, J. J., 1989b, *Phys. Rev. B* 39:8066.
Washburn, S., Fowler, A. B., Schmid, H., and Kern, D., 1988, *Phys. Rev.Lett.*, 61:2201.
Wharam, D. A., Thornton, T. J., Newbury, R., Pepper, M., Ahmed, H., Frost, J. E. F., Hasko, D. G., Peacock, D. C., Ritchie, D. A., and Jones, G. A. C., 1988, *J. Phys. C*, 21:L209.
von Klitzing, K., Dorda, G., and Pepper, M., 1980, *Phys. Rev. Lett.*, 45:494.

14

NOISE IN SMALL AND ULTRA-SMALL GEOMETRIES

Lino Reggiani and Tilmann Kuhn*

Dipartimento di Fisica e Centro Interuniversitario di Struttura della Materia
Universita' di Modena, Via Campi 213/A, 41100 Modena, Italy

INTRODUCTION AND GENERAL DEFINITIONS

The aim of this lecture is to investigate the effect of scaling down the geometrical dimensions on the electronic noise properties of a two-terminal device of length L. Under stationary conditions, the basic quantities of interest are the current spectral density, $S_I(\omega)$, and the correlation function of the instantaneous total current fluctuation, $C_I(t)$. The former is defined, through the power spectrum theorem, by (Van der Ziel 1986):

$$\frac{1}{2\pi}\int_0^{\infty} S_I(\omega)\, d\omega = \overline{\delta I^2} \quad , \tag{1}$$

where

$$\overline{\delta I^2} = \overline{I^2} - \bar{I}^2$$

is the variance of the total current fluctuation $\delta I(t) = I(t) - I_{av}$, $I(t)$ being the instantaneous total current (which includes both conduction and displacement contribution) as measured in the outside closed circuit, with the bar and I_{av} both indicating the time averaged current.

The correlation function is defined as:

$$C_I(t) = \overline{\delta I(0)\delta I(t)} \quad . \tag{2}$$

The Wiener-Khintchine theorem states that (McQuarrie, 1976):

$$S_I(\omega) = 2\int_{-\infty}^{\infty} C_I(t)\, e^{i\omega t}\, dt \quad . \tag{3}$$

Equation (1) is normally used to obtain $S_I(\omega)$ experimentally by measuring the power dissipated by the device per unit band-width, while the correlation function in (2) represents the result of a theoretical analysis. Therefore, (3) establishes the basis for a comparison between theory and experiment. For completeness we remark that the same formulae hold when the current $I(t)$ is replaced by the voltage $V(t)$ as measured between the two terminals of the device.

At equilibrium we have the two following theorems: Within a field (or Schrödinger) approach, which associates the current fluctuations to fluctuations of carrier velocity due to the their Brownian like motion, the Nyquist theorem states that (Nyquist, 1928):

$$S_I(0) = 4\, k_B T\, G \quad , \tag{4}$$

k_B the Boltzmann constant, T the bath temperature, and G the conductance of the two terminal device. For Ohmic devices G is independently obtained from Ohm's law. Conversely, (4) can be taken as a generalized definition of G once $C_I(t)$ is given, as can be seen in the Kubo formula (Toda *et al.*, 1983; Kubo *et al.*, 1983). Within a particle (or Heisenberg) approach, which associates the current fluctuations to fluctuations in carrier number due to the Poissonian randomness of their arrival at the collecting electrode, the Schottky theorem states that (Schottky, 1918):

$$S_I(0) = 2qI_{eq} \quad , \tag{5}$$

q being the absolute value of the charge of the carrier and I_{eq} the equivalent current of a saturated vacuum diode. The appearance of the charge of the carriers in this formula reflects the fact that the current noise is indeed a consequence of the granular nature of the electric charge and therefore the noise can be used to measure the value of the elementary charge. We remark that, for devices which differ from a vacuum diode, the microscopic interpretation of I_{eq} remains an open problem.

EQUIVALENT FORMULATION OF NYQUIST AND SCHOTTKY THEOREMS

For the case of an Ohmic conductance, the Nyquist formula can be deduced from Schottky's theorem by using the generalized Einstein relationship (Kubo *et al.*, 1983):

$$G = \frac{q^2}{L^2} N_{av} D \frac{\partial \ln(N_{av})}{\partial \mu_0} = \frac{q^2}{L^2}\frac{D}{k_B T}\overline{\delta N^2} \quad , \tag{6}$$

N_{av} being the average number of charge carriers inside the device, D their diffusion coefficient (which in general depends on N_{av}) and μ_0 the chemical potential. Here, we have used the identity for the number fluctutations in a Fermi gas

$$\overline{\delta N^2} = N_{av}\, k_B T \frac{\partial \ln(N_{av})}{\partial \mu_0} \quad , \tag{7}$$

which reduces in the classical limit to

$$\overline{\delta N^2} = N_{av} \quad . \tag{8}$$

Then, from (4) to (7) we obtain

$$I_{eq} = 2qN_{av}\frac{Dk_B T}{L^2}\frac{\partial \ln(N_{av})}{\partial \mu_0} = \frac{2q\overline{\delta N^2}}{\tau_d} \quad , \tag{9}$$

where we define a so called "average diffusion transit time" τ_d, given by

$$\tau_d = \frac{L^2}{D} \quad , \tag{10}$$

which represents the average time required by a carrier, injected from one of the terminal

contacts, to cross the device through a diffusive motion being finally collected by the opposite contact. For the limiting case of a non-degenerate gas, we can write

$$\tau_d = \frac{L^2}{D} = \frac{L^2}{l\sqrt{\frac{k_B T}{m}}} = \tau_c \frac{L^2}{l^2} , \qquad (11)$$

where τ_c is the average collision time (obtained from the definition of mobility $\mu = q\tau_c/m$) and l is the mean free path. Obvioulsly, in this case the spectral density is a function of the geometry (L) and the bulk properties (l, τ_c) of the device.

Taking advantage of the above definitions, we introduce the general concept of an *average transit time*, τ_T, which also includes the possibility of ballistic motion. In terms of τ_T, the current spectral density can be given the general form

$$S_I(0) = \frac{4q^2\overline{\delta N^2}}{\tau_T} . \qquad (12)$$

Then, in the case of a diffusive motion, which corresponds to the condition $L >> l$, it is:

$$\tau_T = \tau_d . \qquad (13)$$

In the case of a ballistic motion, which corresponds to the condition $L << l$, it is:

$$\tau_T = \tau_b = \frac{L}{v_b} , \qquad (13b)$$

where τ_b is the "average ballistic transit time" and v_b is the "average ballistic velocity" of the injected carriers in the direction of the two contacts. The spectral density now is a function of geometry (L) and of the properties of the contacts acting as carrier injectors (v_b).

SIZE EFFECTS ON THERMAL NOISE AND CONDUCTANCE

A deeper physical insight of the average times and velocities introduced above can be obtained from the knowledge of the correlation function in (2). Stanton and Wilkins (1985) have calculated the current correlation function for a classical gas in a finite one-dimensional system. For a Fermi gas the situation is much more complicated because of the difficulty in modeling contacts in a quantum mechanical treatment. With the assumption that there are no correlations between the carriers in the system and in the contacts, we obtain a generalized formula for the current correlation function in a one-dimensional system of length L at thermal equilibrium (Kuhn *et al.*, 1990)

$$C_I(t) = \frac{2q^2}{L\pi\hbar}\sqrt{\frac{2}{m}} \int_0^{\frac{L^2 m}{2t^2}} \varepsilon^{1/2} f(\varepsilon)[1 - f(\varepsilon)]\left[1 - \sqrt{\frac{2t^2}{L^2 m}}\right] d\varepsilon , \qquad (14)$$

where $f(\varepsilon)$\$ is the Fermi-Dirac distribution function.

The shape of the correlation function depends only on two dimensionless parameters. One is the degree of degeneracy as determined by the factor $\delta = \mu_0/k_B T$. The result of Stanton and Wilkins is obtained from (14) by neglecting the occupation factor ($1-f$) or, equivalently, in the classical limit $\delta \rightarrow -\infty$. On the other hand, the limit $\delta \rightarrow \infty$ corresponds to the complete degenerate case. The second parameter $\alpha = \tau_L/\tau_c$ relates a characteristic time for ballistic transport τ_L to the relaxation time τ_c. The characteristic time τ_L can be defined by replacing the energy in the term $(2t^2\varepsilon/L^2m)^{1/2}$ in (14) by a characteristic energy of the carrier system. This characteristic energy is different for the case of a classical and a degenerate gas. The former case is determined by the thermal energy and the latter one by its chemical potential. Thus, we define $\tau_L^{class} = L\,(m/2k_BT)^{1/2}$ for $\mu_0 < k_BT$ and $\tau_L^{degen} = L\,(m/2\mu_0)^{1/2}$ for $\mu_0 > k_BT$. Then, the value of α determines the type of the transport: For $\alpha >> 1$ the correlation function is governed by the exponential term in front of the integral in (14) and the tranport is diffusive. For $\alpha << 1$, this exponential term is approximately constant on the time-scale where the integral decays and the transport is ballistic.

Figure 1 shows the normalized correlation functions for the ballistic regime ($\alpha = 0$) in a linear plot. While in the classical limit the long time decay is according to a power law, in the degenerate limit there exists an upper value of the time, given by τ_L, above which the correlation function is identically zero. In this latter case the correlation function has the characteristic triangular shape well-known from the shot noise of an injector at constant velocity (Lecoy, 1990). In our case the carriers are emitted at all velocities up to the Fermi velocity v_F. Due to the occupancy factor in (14), however, only those carriers at the Fermi edge contribute to the noise.

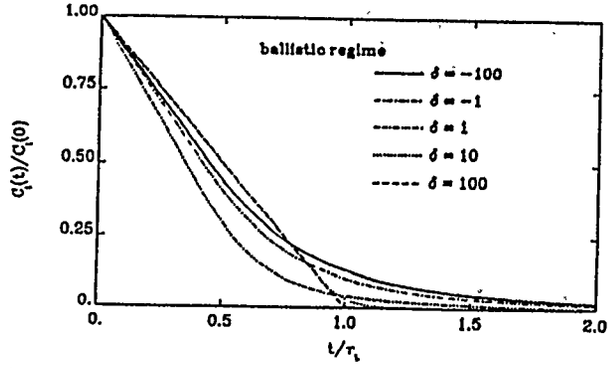


Fig. 1 Normalized current correlation function in the ballistic regime showing the transition from the classical (solid line) to the completely degenerate limit (dashed line).

In Fig. 2 the normalized correlation functions are plotted for different values of the parameter α covering the range from the diffusive to the ballistic regime in an Arrhenius plot. Figure 2a shows the classical limit ($\delta \rightarrow -\infty$) and Fig. 2b the degenerate limit ($\delta \rightarrow \infty$). As can be seen from (14), in the diffusive regime the exponential term dominates and the results are the same for the classical and the degenerate case. In the ballistic limit, however, there are large differences as already seen in Fig. 1.

The corresponding normalized spectral densities are shown in Fig. 3. In the diffusive regime we obtain the Lorentzian shape independent of the degree of degeneracy. The transition to the ballistic regime in the classical limit (Fig. 3a) manifests itself in a blue-shift of the corner frequency as well as in a slight deviation from the Lorentzian shape. In the degenerate limit (Fig. 3b), an analogous blue-shift of the corner frequency is observed, the deviation from the

Lorentzian shape, however, is much more pronounced. The spectral density now exhibits an oscillatory behaviour which, in the ballistic limit, leads to a discrete set of frequencies $\omega = 2\pi n v_F/L$ (n being an integer number) where there is no noise. The appearance of L demonstrates that these are geometrical resonances for carriers at the Fermi edge.

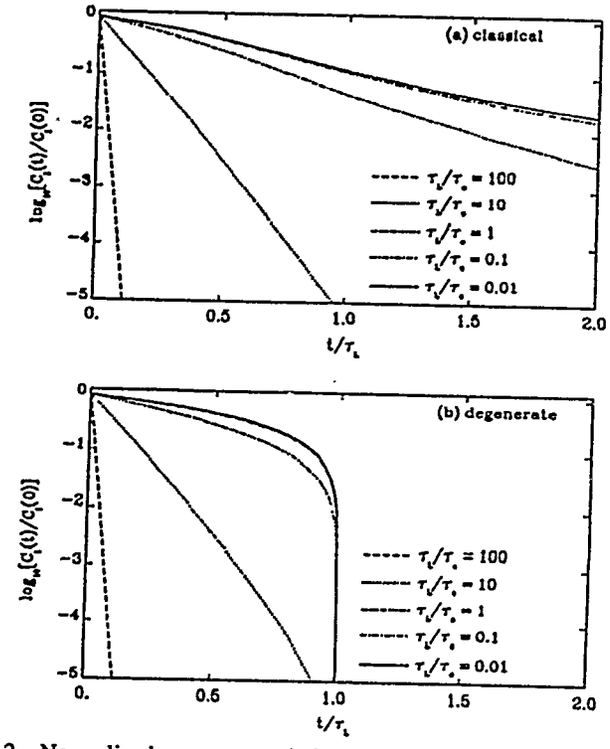


Fig. 2 Normalized current correlation functions showing the transition from the diffusive (dashed lines) to the ballistic regime (solid lines). (a) Classical limit ($\delta \to -\infty$), (b) degenerate limit ($\delta \to \infty$).

From the correlation function of (14) we can provide a better physical interpretation of the parameters previously introduced. First, we notice that the inititial value of the correlation function is directly related to the density of the one-dimensional Fermi gas by (Kuhn *et al.*, 1990)

$$C_I(0) = \frac{q^2 N_{av}}{L^2}\frac{k_B T}{m} . \qquad (15)$$

Therefore, we obtain the general expression for the conductance

$$G = \frac{q^2 N_{av}}{mL^2}\int_0^\infty \frac{C_I(t)}{C_I(0)}\,dt , \qquad (16)$$

and for the average transit time τ_T

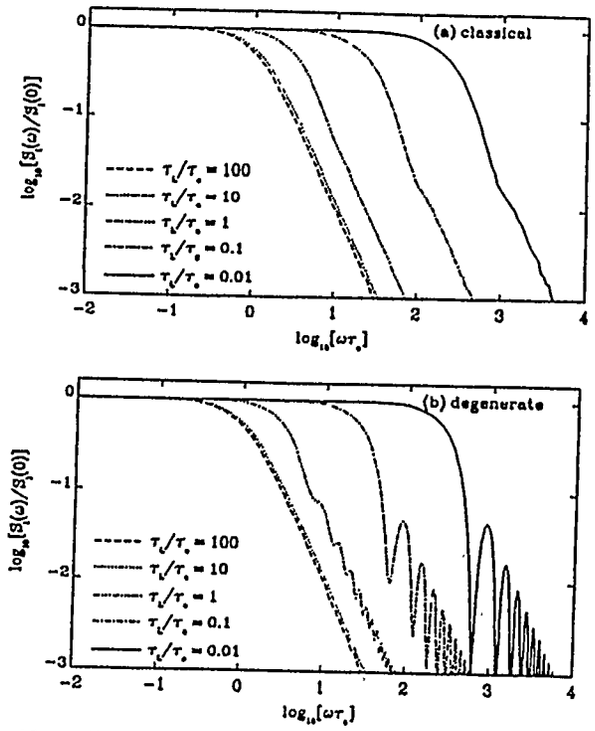


Fig. 3 Normalized current spectral densities as function of frequency corresponding to the correlation functions in Fig. (2). (a) Classical limit ($\delta \to -\infty$), (b) degenerate limit ($\delta \to \infty$).

$$\frac{1}{\tau_T} = \frac{1}{L^2}\left(\frac{\partial \ln(N_{av})}{\partial \mu_0}\right)^{-1}\int_0^\infty \frac{C_I(t)}{C_I(0)}\,dt . \qquad (17)$$

Figure 4 illustrates τ_T in units of τ_c as a function of the sample length measured in units of the mean free path l given by $l^{class} = \tau_c\,(k_B T/m)^{1/2}$ in the classical limit and $l^{degen} = \tau_c\,(2\mu_0/m)^{1/2}$ in the degenerate limit. The transition from ballistic to diffusive transport is clearly evidenced by the change from L to L^2 behavior as expected from (11) and (13b). In the diffusive regime the two limiting cases give exactly the same result, while in the ballistic regime there is a small difference by a factor of $(\pi/2)^{1/2}$. Thus, we conclude that the definition of an average transit time from the low frequency spectral density is a useful concept, because it reproduces the ballistic and the diffusive limits while at the same time provides an interpolation formula between these limits.

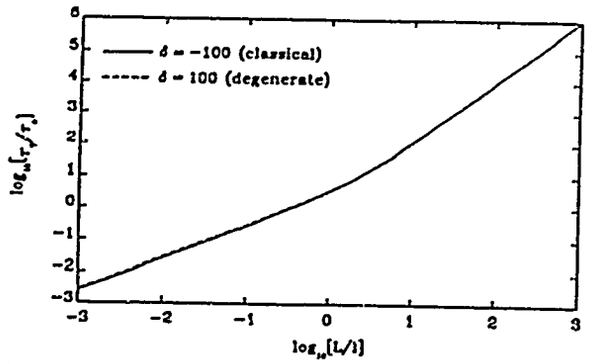


Fig. 4 Average transit time as a function of the sample length for the classical and the completely degenerate case. The change from an L to an L^2 law evidences the transition from ballistic to diffusive transport. Only in the ballistic regime there is a small difference between the two cases.

Under the diffusive condition (14), (16) and (17) may be combined to give:

$$S_I(0) = \frac{4q^2 k_B T N_{av}}{mL^2}\tau_c \tag{18}$$

and

$$G = \frac{q^2 N_{av}}{mL^2}\tau_c \ , \tag{19}$$

independent of the degree of degeneracy. Under this condition, for a given one-dimensional carrier concentration $n_{av} = N_{av}/L$ the low-frequency spectral density $S_I(0)$ and therefore G are proportional to L^{-1}. For the average transit time we obtain in the classical limit

$$\tau_T^{class} = \frac{mL^2}{k_B T \tau_c} \tag{20}$$

as expected from the general considerations in (11). In the degenerate limit we obtain

$$\tau_T^{degen} = \frac{mL^2}{2\tau_c \mu_0} = \frac{L^2}{\tau_c v_F^2} \ , \tag{21}$$

where we have introduced the Fermi velocity $v_F = (2\mu_0/m)^{1/2}$. Using (10) the diffusion coefficient is now given by

$$D = v_F^2 \tau_c \ , \tag{22}$$

demonstrating that the diffusion of a degenerate gas is governed by the carriers at the Fermi energy.

Under the ballistic condition (16) and (17) can be evaluated analytically in the two limiting casses of a classical and a completely degenerate system. The low frequency spectral density now reads, respectively:

$$S_I^{class}(0) = \frac{4q^2 N_{av}}{L}\left(\frac{k_B T}{2\pi m}\right)^{1/2} \ , \tag{23}$$

and

$$S_I^{degen}(0) = \frac{8q^2 k_B T}{h} \ . \tag{24}$$

Under this condition, for a given one dimensional carrier concentration n_{av} the value of $S_I(0)$ in both cases is independent of L and in the degenerate limit even independent of $\bar n$. The average transit times are given, respectively, by:

$$\tau_T^{class} = L\sqrt{\frac{2\pi}{k_B T}} \ , \tag{25}$$

and

$$\tau_T^{degen} = L\frac{2m}{\mu_0} = \frac{2L}{v_F} \ . \tag{26}$$

Again, we recover the results from the general considerations above. Indeed, using (13b) we find that the ballistic velocity v_b really is the average velocity of the injected carriers in the classical $[v_b = (k_B T/2\pi m)^{1/2}]$ as well as in the degenerate case $(v_b = v_F/2)$.

The conductance in the ballistic regime for the classical and completely degenerate regime is given, respectively, by:

$$G^{class} = \frac{q^2 N_{av}}{L}\left(\frac{1}{2\pi m k_B T}\right)^{1/2} \ , \tag{27}$$

and

$$G^{degen} = \frac{2q^2}{h} \ . \tag{28}$$

Thus, the conductance defined by the Nyquist formula for a degenerate Fermi gas under ballistic conditions has just the value known from the universal conductance fluctuations (Feng *et al.*, 1986).

We remark that for a classical system the present results hold for any spatial dimension of the problem. On the other hand, in the degenerate limit these results only hold for a quasi-one-dimensional sytem, that is, the two perpendicular directions have to be quantized. As is well known, in this case the conductance as measured directly from the current-voltage characteristic can be quantized (Van Wees *et al.*, 1988).

CONCLUSIONS

We have presented an equivalent formulation of the Nyquist and Schottky picture for the study of noise in small and ultrasmall geometries. The low frequency spectral densitiy has been expressed either in terms of a conductance G or an average transit time τ_T. In any case,

the basic physical quantity is the current autocorrelation function $C_I(t)$. We have provided an explicit form of $C_I(t)$ adequate to a Fermi gas at equilibrium for arbitrary values of the degeneracy and for systems of arbitrary but finite length. Using this function we have studied the transition from the diffusive to the ballistic regime. This transition implies the following aspects for the current spectral density: (1) The cut-off frequency increases with decreasing sample length. (2) The high-frequency decay becomes non-Lorentzian, especially in the case of a degenerate system, where we obtain an oscillatory behaviour. This leads in the ballistic limit to a discrete set of angular frequencies $\omega = 2\pi n v_F/L$ where the spectral density vanishes. These frequencies correspond to geometrical resonances for the carriers at the Fermi edge. At this frequency the motion is totally correlated and no noise occurs. (3) While in the diffusive regime the low frequency spectral density is proportional to L^{-1}, in the ballistic regime it becomes independent of L. (4) The temperature dependence of $S_I(0)$ in the diffusive regime is governed by the temperature dependence of the scattering rate. In the ballistic regime, however, it becomes a universal function depending only on the degree of degeneracy. (5) The conductance for the degenerate limit in the ballistic regime is given by the value of the universal conductance fluctuations.

The present analysis has been restricted to the case of thermal equilibrium. Already in this case the definition of the relevant quantities for small systems is a complicated task due to the nonlocality of quantum mechanics. From the point of view of an application to ultrasmall devices, however, it will also be necessary to generalize the results to the case of external electric fields, where, as it is known from bulk studies (Reggiani, 1985; Weissman, 1988; Gurevich, 1989), much more complicated shapes of the spectral density can occur, induced by the interactions between different time-scales relevant for the problem.

ACKNOWLEDGMENTS

This work has been partially supported by the CEE ESPRIT II BRA 3017 project and the Centro di Calcolo of the Modena University.

(a) Permanent address: Institut für Theoretische Physik, Universität Stuttgart, Pfaffenwaldring 57, 7000 Stuttgart 80, Federal Republic of Germany.

REFERENCES

Feng, S., Lee, P. A., and Stone, A., D., 1986, Sensitivity of the conductance of a disordered metal to the motion of a single atom: Implication for 1/*f* noise, ıÈ, 56:1960.

Gurevich, V. L., 1989, Nonequilibrium carrier noise and its effects in microstructures, *Solid State Electron.*, 32:1749.

Kubo, R., Toda, M., and Hashitsume, N., 1983, "Statistical Physics II", Springer Series in Solid-State Sciences Vol. 31, Springer-Verlag, Berlin.

Kuhn, T., and Reggiani, L., 1990, to be published.

Lecoy, G., 1990, "Lecture notes", Montpellier, unpublished.

McQuarrie, P. A., 1976, "Statistical Mechanics", Harper and Row, New York.

Nyquist, H., 1928, Thermal agitation of electric charge in conductors, *Phys. Rev.*, 32:110.

Reggiani, L., 1985, "Hot Electron Transport in Semiconductors", Topics in Applied Physics Vol. 58, Springer-Verlag, Berlin.

Schottky, W., 1918, Uber spontane Stromschwankungen in verschiedenen Elektrizitätsleitern, *Ann. der Physik*, 57:541.

Stanton, C. J., and Wilkins, J. W., 1985, Non-equilibrium current fluctuations in finite size semiconductors, *Physica*, 134B:255.

Toda, M., Kubo, R., and Saitô, N., 1983, "Statistical Physics I", Springer Series in Solid-State Sciences Vol. 30, Springer-Verlag, Berlin.

Van der Ziel, A., 1986, "Noise in Solid State Devices and Circuits", Wiley and Sons, New York.

Van Wees, B. J., Van Houten, H., Beenakker, C. W. J., Williamson, J. G., Kouwenhoven, L. P., Van der Marel, D., and Foxon, C. T., 1988, Quantized conductance of point contacts in a two-dimensional electron gas, *Phys. Rev. Lett.*, 60:848.

Weissman, M. B., 1988, 1/*f* noise and other slow, nonexponential kinetics in condensed matter, *Rev. Mod. Phys.*, 60:537.

16

NONEQUILIBRIUM GREEN FUNCTION TECHNIQUES APPLIED TO HOT-ELECTRON QUANTUM TRANSPORT

Antti-Pekka Jauho

Physics Laboratory, University of Copenhagen, H.C.Ørsted Institute
Universitetsparken 5, DK-2100 Copenhagen Ø, Denmark

INTRODUCTION

Transport in semiconductors has traditionally been described with the help of the Boltzmann equation (BE) (Conwell, 1967) (or by some of its simplifications, such as the drift-diffusion equation). Powerful Monte Carlo techniques have been developed to solve the BE (for a review see e.g. Jacoboni and Reggiani, 1983), and combined with Poisson equation solvers these methods form the basis for the theoretical analysis and design of modern semiconductor microdevices. As the characteristic length scales continue to decrease, however, quantum effects begin to dominate the transport, and consequently the semiclassical BE cannot be used as a starting point. Examples of such quantum effects are space quantization (the characteristic length scales in one or several directions are such that plane waves are no longer appropriate wave functions for describing the charge carriers), or ballistic transport (the charge carriers experience no, or only few, collisions within the active region of the device); here we point out the difference between quantum ballistic transport, such as underlies resistance quantization in quantum point contacts (van Wees *et al.*, 1988; Wharam *et al.*, 1988) and ballistic peaks observed in the distribution function in Boltzmann equation studies of microstructures (Baranger and Wilkins, 1984). Other effects beyond the conventional Boltzmann picture include the influence of field on collision processes, the so-called intra-collisional field effect (ICFE) (Barker, 1973; Levinson, 1970), and collisional broadening (CB), or, in other words, effects due to finite quasiparticle lifetimes. A common theme for all these quantum effects is that the *phase coherence* of the charge carriers is maintained longer than some characteristic length scale (for example, the inelastic mean free path), or time scale (the collision duration time is not negligible).

A number of theoretical approaches have been developed to describe quantum transport. Methods based on the Landauer formula (Landauer, 1970) have been particularly useful in analyzing transport in situations where inelastic collisions in the 'interesting' part of the system are infrequent (dissipation and loss of phase coherence takes place in the contacts, or measuring probes). The Feynman path integral method has been applied to high field transport in semiconductors (Thornber, 1978; Mason and Hess, 1989), but it has not yet gained widespread use. A quantum Langevin equation approach has been developed by Hu and O'Connell (1987, 1988, 1989). Wigner distribtution function methods have also been reported (Frensley, 1987; and Kluksdahl *et al.*, 1989; the latter reference contains extensive references to other papers employing the Wigner distribution function). In fact, the Wigner function bears a close connection to the nonequilibrium Green's function methods which form the actual topic of this review. Quite recently, Rammer (1990) has developed a *single-particle* diagrammatic density matrix technique which appears promising. Datta (1989) has derived a steady state kinetic equation for a system where the inelastic interaction is due to impurities with internal degrees of freedom, and shown how the current response can be written in a form similar to the Landauer equation.

THEORETICAL BACKGROUND

The nonequilibrium Green's functions were developed simultaneously, and independetly by Baym and Kadanoff (1962), and by Keldysh (1965). These two formalisms are equivalent. An elegant demonstration can be found in the review article by Langreth (1976). Several review articles focusing on different aspects have recently appeared: Danielewicz (1984) (application to nonequilibrium nuclear collisions), Rammer and Smith (1986) (quasi-classical Green functions; applications to degenerate Fermi systems); and Jauho (1990) (derivation of quantum kinetic equations for model systems). We refer the reader to these articles for technical details and additional references. Here we try to elucidate the main physical content of the various formulations, and only sketch the often rather lengthy derivations.

In equilibrium (and consequently also in linear response theory, which only involves equilibrium quantities) one can prove the *fluctuation-dissipation theorem* which connects the causal (or time-ordered) Green's function to the retarded (or advanced) Green's function. This relation is of great importance: Wick's theorem, and hence the diagrammatic perturbation techniques can be proved only for the causal function while physical observables are obtained from response functions which are related to retarded functions, or correlation functions. Thus, equilibbrium theory is in a way short-circuited: it is not necessary to develop calculational schemes for *both* causal *and* retarded functions. This fortunate situation does not hold in nonequilibrium: one has to develop a theory which contains the two types of Green's functions as independent objects. This was the bad news, the good news is that nonequilibrium theory can be written *formally* in a form which appears entirely equivalent to equilibrium theory for the causal Green's functions (and thus the diagrammatic perturbation theory exists). The price that has to be paid is the time-labels of the Green's function no longer are real: they reside on a complex path which goes from $-\infty$ to $+\infty$, and back to $-\infty$. As shown by Craig (1968) and Langreth (1976) the Baym-Kadanoff and Keldysh formalisms correspond to slightly different choices of this contour in the complex-time plane. Complex-time integrals are not convenient to work with, and the set of rules which tells how to extract real time quantities out of objects defined on a complex contour is known as 'analytic continuation' or 'Langreth theorems' (Langreth and Wilkins, 1972; Langreth, 1976). These rules can be summarized as follows: Consider the following integral defined on the complex-time contour (this is the generic type of integral one encounters in the diagrammatic perturbation theory):

$$C(t_1,t_{1'}) = \int_c d\tau\, A(t_1,\tau)\, B(\tau,t_{1'}) \qquad (1)$$

Here the functions A and B are Green's functions or self-energies familiar from the Dyson equation. If t_1 is on the first part of the contour, and $t_{1'}$ on the second, we call C a 'lesser than' function (because t_1 is smaller than $t_{1'}$ in the contour sense) and denote it by $C^<$. According to the Langreth rules (1) reads on the real axis as

$$C^<(t_1,t_{1'}) = \int_{-\infty}^{\infty} dt (A_r(t_1,t)\, B^<(t,t_{1'}) + A^<(t_1,t)\, B_a(t,t_{1'})) \qquad (2)$$

A similar result can be derived for the 'greater than' function. Rules analogous to (2) can also be given in the case where two functions are multiplied together without integration over intermediate variables (Langreth, 1976). As remarked above, the Dyson equation has the same formal appearance as in equilibrium: the only difference is that the internal time integrations now run along the complex contour. Thus self-energies can be viewed as known functionals of the nonequilibrium Green's functions. Apply now (2) to the the Dyson equation defined on the contour (here we suppress the integrations):

$$(G_0^{-1} - U)\,G = 1 + \Sigma G$$

$$G\,(G_0^{-1} - U) = 1 + G\Sigma \qquad (3)$$

which become

$$(G_0^{-1} - U)\,G^< = \Sigma_r G^< + \Sigma^< G_a$$

$$G^<(G_0^{-1} - U) = G_r\Sigma^< + G^<\Sigma_a \qquad (4)$$

Subtracting these equations, and after some manipulations, one can derive the (generalized) Baym-Kadanoff equation (GKB):

$$[G_0^{-1} - U, G^<] - [\Sigma, G^<] - [\Sigma^<, G] = \frac{1}{2}\{\Sigma^<, G^>\} + \frac{1}{2}\{\Sigma^>, G^<\} \qquad (5)$$

Here we define

$$\Sigma \equiv \frac{1}{2}(\Sigma_a + \Sigma_r)\ ,\ G \equiv \frac{1}{2}(G_a + G_r) \qquad (6)$$

and the symbol U contains the driving forces: heterojunction conduction-band-edge potentials, and other one-body potentials, e.g. self-consistent Hartree terms. To obtain a closed set of equations the GKB equation must be supplemented with Dyson equations for the retarded and advanced Green's functions:

$$G_{r,a} = G^0_{r,a} + G^0_{r,a}\,U\,G_{r,a} + G^0_{r,a}\Sigma_{r,a}G_{r,a} \qquad (7)$$

The various Green functions are defined as (we use a shorthand notation: $1 \equiv (\mathbf{x}_1, t_1)$ etc)

$$G^<(1,1') = i\langle \psi^\dagger(1')\,\psi(1)\rangle \qquad (8)$$

$$G_r(1,1') = -i\theta(t_1 - t_{1'})\langle\{\psi(1), \psi^\dagger(1')\}\rangle \qquad G_a(1,1') = i\theta(t_{1'} - t_1)\langle\{\psi(1), \psi^\dagger(1')\}\rangle \qquad (9)$$

Brackets indicate commutators and curly brackets mean anticommutators. A product of two terms implies integration or matrix multiplication over intermediate variables.

The GKB equations are exact (given that a self-energy can be defined), and form the starting point for the theory. Many different applications exist in the literature, and it is beyond the scope of these lectures to give an extensive review. The linear response (in the driving fields) studied with the aid of the GKB equations is on a firm basis, and has been reviewed by Mahan (1984, 1987). Equivalence of the GKB equations and Kubo formula (within linear response) has recently been demonstrated (Chen and Su, 1989). An interesting recent application is the study of strongly disordered systems within the GKB formalism: Hershfield and Ambegaokar (1986) and Strinati *et al.* (1989) have derived quantum-kinetic equations to treat localization effects. The main task of the present review is to discuss *nonlinear* applications, and before embarking on this effort we discuss some general features of the GKB equations.

In general, the GKB equation and the Dyson equation are coupled equations: for example, the retarded self-energy may depend on the correlation function $G^<$. This, however, leads to enormous complications, and usually one attempts to split the calculation into two steps: first the retarded/advanced Green functions are determined, and then used as an input to the GKB equation.

The GKB equation can be given the following physical interpretation. From the definition of the correlation function $G^<$ it follows that the Wigner-distribution function can be extracted from it:

$$f^W(p,r,T) = -iG^<(p,\tau=0,R,T) = -i\int \frac{d\omega}{2\pi} G^<(p,\omega,R,T) \qquad (10)$$

Here we introduce the center-of-mass coordinates: $R = (x + x')/2$, $r = x - x'$; p is Fourier transform of r. Analogous definitions hold for the temporal variables. The Wigner function is the quantum mechanical generalization of the semiclassical Boltzmann distribution: in particular, physical observables such as density, or current density are obtained as its moments. Some care should be excercized, however — the Wigner function is not positive semidefinite, and no probability interpretation can be given. Thus, the GKB equation governs the *distribution* of particles — it is a quantum kinetic equation which generalizes the Boltzmann equation. Its structure is also suggestive: The first term on the left-hand side gives rise to a driving term, the second and third terms are renormalization terms, and the right side is the quantum collision term with the characteristic gain-loss structure. A demonstration of how GKB reduces to the BE can be found in Langreth (1976) or Jauho (1990). A final point to note is that the initial value for the distribution used in GKB shoud be chosen to be consistent with the Dyson equation (Barker, 1987; Kluksdahl *et al.*, 1989).

The GKB equations, when integrated over ω (or, equivalently, by setting $t = 0$); can be written as (Reggiani *et al.*, 1987)

$$E\cdot\frac{\partial}{\partial k} f^W(k) = -\int (\Sigma^> G^< + G^<\Sigma^> - \Sigma^< G^> - G^>\Sigma^<)\ , \qquad (11)$$

where the terms on the right-hand side have the structure

$$\int BC \equiv \int_0^\infty du\, B\left(k - \frac{1}{2}Eu, u\right) C\left(k - \frac{1}{2}Eu, -u\right)\ . \qquad (12)$$

These equations are still exact (with the same restrictions as the GKB equation), and they are valid for spatially homogenous and constant electric fields. The structure of (11) illustrates one of the main difficulties in constructing a quantum-kinetic theory for the Wigner function — the equation is not closed but involves the full correlation function $G^<$. Below, we return to this point. Incidentally, when examining (11)-(12), one may wonder what happened to the renormalization terms in the GKB equation (5) (the second and the third term on the left-hand side). These terms are, in fact, contained in (11)-(12): their effect is to produce the shifted momentum labels in the integrand, and to restrict the integration from 0 to ∞ rather than for the full range as is the case in the GKB equation. Linearized forms of (11)-(12) coincide with the transport equations of Hänsch and Mahan (1983).

Consider now the Dyson equation for the retarded/advanced Green's functions. We recall from equilibrium Green's function theory the definition of the *spectral density* $A = i(G_r - G_a)$. Many important objects (e.g. density of states, scattering rates, quasiparticle lifetimes) require the knowledge of the spectral density, and thus the solution of the Dyson

equation, apart for serving as an input to the GKB equation, is of independent interest. Below we analyze several model spectral densities.

The GKB and Dyson equations are often too complicated to be solved directly in the nonlinear case, and during recent years many research groups have developed simplifications and approximation schemes which allow further progress. We now turn to these applications.

QUANTUM KINETIC EQUATIONS FOR THE WIGNER FUNCTION

There are a number of reasons for trying to develop quantum kinetic equations for the Wigner function f_W, rather than trying to solve directly for the correlation function $G^<$. First, f_W depends on one less variable than $G^<$, and is therefore hopefully a less complicated object. (In fact, numerical studies show that for spatially inhomogenous systems f_W is an exceedingly complex object (Kluksdahl *et al.*, 1989). Second, most physical observables do not require the knowledge of the full correlation function (but some, such as the many body energy, do!) and the knowledge of f_W suffices. Next, since f_W has many properties of a true distribution function, its equation of motion can be expected to resemble the semi-classical BE so much that a part of the vast body of experience on interpreting and solving the BE can perhaps be carried over to the quantum case. Finally, many other semiconductor quantum transport equations have been proposed (e.g. Barker and Ferry, 1979; Levinson, 1970; Seminozhenko, 1982) and it would be desirable to obtain independent rederivations and/or generalizations of these results.

In order to reduce the GKB equation to an equation governing the Wigner function some assumptions must be made. The approach chosen by many groups has been to make an Ansatz which directly expresses $G^<$ in terms of f_W. The relation (10) between the Wigner function and the correlation function provides an important sum rule (in ω-space), or a boundary condition (in t-space) which any guess for $G^<$ must satisfy. In *equilibrium* the exact relation

$$G^<(p,\omega) = i\, A(p,\omega)\, f_{FD}(\omega) \tag{13}$$

holds (here f_{FD} is the Fermi-Dirac distribution). The above relation is one way of expressing the fluctuation-dissipation theorem. In non-equilibrium no such theorem, in general, is known, and it is clear that any guess trying to relate the correlation function and the Wigner function in nonequilibrium conditions can have only restricted range of validity. Early papers (Barker, 1981; Jauho and Wilkins, 1982) employed a direct generalization of (13) to nonequilibrium:

$$G^<(p,\omega,R,T) = i\, A(p,\omega,R,T)\, f^W(p,R,T) \tag{14}$$

Here A is the nonequilibrium spectral function obtained from the solution of the nonequilibrium Dyson equation (7). This assumption satisfies the sum rule (10), and for vanishing fields it reduces to the exact equilibrium result *provided that* the spectral function A is *strongly peaked* at $\omega = \varepsilon(p)$. This is equivalent to making the quasi-particle approximation. Using (13) in the kinetic equation [either (5) or (11)] leads to a closed equation for $f_W$which one may then attempt to solve for various types of interactions. The weak point of this approach is that guessing an Ansatz does not provide any means of estimating its limits of validity. Even worse, it was soon found (Jauho and Wilkins, 1984) that the *Ansatz* (13) led to a collision integral which differed from the one derived with the density matrix method (Barker and Ferry, 1979; Levinson, 1970). The resolution of this paradox came in two steps: It first was realized that the theory should be formulated in a gauge invariant manner [the crucial ideas occur already in the early work of Langreth (1966), and in the linear theory of Mahan and Hänsch (1983), but it took some time before they were adopted to semiconductor high-field transport (Sarker, 1985; Khan *et al.*, 1987; Reggiani *et al.*, 1987)]. The second step was completed when Vinogradov (1986), and Lipavsky *et al.* (1986) gave the first *systematic* derivations of an *Ansatz* of the type of (14). With these improvements perfect agreement was found between the Green's function methods, and the earlier density matrix results. For uniform and steady fields the replacement of (14) reads

$$G^<(k,\tau) = iA(k,\tau)\, f_W(k - \tfrac{1}{2}E\tau)\ . \tag{15}$$

Note that here one uses the kinematical momentum k rather than the canonical momentum p, and that the expression is in t-space in contrast to (14) which is in ω space. The point of the derivations leading to (15) is that the correlation function $G^<$ is written as a time-diagonal piece, and a correction term which has an integral-equation structure. The integral equation is not a perturbative expansion in a small coupling constant but rather an expansion in the various relaxation times — the quasi-particle life-time and the characteristic decay time for correlations. While this approach appears promising its full implications are not yet fully understood. For example, is it possible, in principle, to generalize the procedure for space- or time-inhomogenous systems (Rammer, 1990)? Clearly more work is required here.

We conclude this section by giving the quantum kinetic equation obtained with the *Ansatz* (15). We consider nondegenerate carriers, driven by a uniform and steady driving field of arbitrary strength, and the electron-phonon interaction is treated within the self-consistent Born approximation (Khan *et al.*, 1987)·

$$E\frac{\partial}{\partial k} f_W(k) = \sum_q \int_0^\infty d\tau \,[\, P(k - E\tau,\, k - q - E\tau;\, \tau)\, f_W(k - q - E\tau)$$

$$- P(k + q - E\tau,\, k - E\tau;\, \tau)\, f_W(k - E\tau)]\ , \tag{16}$$

where

$$P(k + q, k;\tau) = 2\pi|M_q|^2 \sum_{\eta=\pm 1} \left[N_q + \tfrac{1}{2}(1+\eta)\right] \times$$

$$\mathrm{Re}\,\frac{1}{\pi}\left[A\left(k + q + \tfrac{1}{2}E\tau, \tau\right) A\left(k + \tfrac{1}{2}E\tau, -\tau\right) e^{-i\omega_q \eta \tau}\right] \tag{17}$$

Setting $E = 0$ in the collision integral and using free spectral densities, $A(k,t) = \exp(-i\varepsilon(k)t)$, recovers the BE. Equations (16)-(17) generalize the Barker-Ferry equation (1979) — the interacting nonequilibrium spectral densities allow, in principle, a systematic treatment of interference effects between driving fileds and scattering.

The mathematical structure of (16)-(17) differs crucially from the BE: the additional integral appearing on the right-hand side makes it unsuitable for Monte Carlo simulation schemes. Further, the quantities P in (16)-(17) are not positive definite. Thus additional simplifications are called for. In the next section we describe some of the suggested approximation schemes. Very little is known about (16)-(17) formal properties, and we mention only a few points where our understanding is incomplete: conservation laws, convergence, stability, existence of solutions, (ir)reversibility, and consistency within a given order of perturbation theory.

FURTHER APPROXIMATIONS

As mentioned above, (16)-(17) appear unsuitable for a direct numerical evaluation. This is because of the retardation, or memory effects in the collision integral. Rather than approximating (16)-(17) directly, it is advantageous to integrate both sides, and after some manipulations one finds [here we follow Khan *et al.* (1987)]

$$f^W(k) = \sum_q \int_0^\infty dt\left\{\widetilde{W}(k - Et, k - q - Et; t)\, f^W(k - q - Et) - \widetilde{W}\, f^W\right\}, \qquad (18)$$

where

$$\widetilde{W}(K + q, K; t) = \int_0^t dt'\, P(K + q, K, t') \qquad (19)$$

In the second term on the right-hand side of (18) one should make the replacement $k \rightarrow k + q$. Equation (18) bears a striking similarity to the integrated BE: the only formal difference is the explicit time-dependence in $\widetilde{W}$. Further simplification is possible if $\widetilde{W}$ approaches its asymptotic value on a time scale faster than any other relevant time scale, and can be replaced by its limiting value $\widetilde{W}(t \rightarrow \infty)$. A condition for this is that $P(t)$ be a short-ranged function of t. Khan *et al.* (1987) argue that this is indeed the case (see their Appendix C), while other groups are content in passing to the asymptotic limit ('completed collisions limit') phenomenologically (Kim *et al.*, 1987 ; Reggiani *et al.*, 1987, 1988a). According to the analysis of Khan *et al.* the asymptotic limit becomes better as the the field gets stronger. These conclusions depend sensitively on the form of the *Ansatz* used to relate $G^<$ and f_W. Further, it has been shown recently, within linear response, that the completed collision limit overestimates the effects of collisional broadening (Suhrke, 1989). These ambiguities stress the importance of further work on the points raised in the previous section.

Summarizing, the completed collisions limit gives rise to a Boltzmann type of transport equation with the energy conserving delta function replaced by a 'joint spectral density' $K(k + q, k)$:

$$\delta(\varepsilon(k + q) - \varepsilon(k) - \eta\omega_q) \rightarrow$$

$$K(k + q, k) \equiv \int_0^\infty d\tau\, \mathrm{Re}\left\{\frac{1}{\pi} A\left(k + q + \tfrac{1}{2}E\tau, \tau\right) A\left(k + \tfrac{1}{2}E\tau, -\tau\right) e^{-i\eta\omega_q\tau}\right\} \qquad (20)$$

In recent years several groups have performed Monte Carlo simulations based on (20). A detailed description of a particular set of simulations can be found in Reggiani *et al.* (1988a); here we summarize the main results. The simulations can be classified according to what physical effects were included in the joint spectral density.

Intracollisional field effect (ICFE)

The spectral density for free electrons in a parabolic band in the presence of a uniform static electric field can be solved analytically (Jauho and Reggiani, 1988; Reggiani *et al.*, 1988b):

$$A(k, \tau) = \exp\left[-i\varepsilon(k)\tau + \frac{E^2}{24m}\tau^3\right], \qquad (21)$$

and the resulting joint-spectral density can be expressed in terms of Fresnel integrals. Here one encounters a conceptual difficulty. The joint-spectral function is not a positive semidefinite quantity, and the probabilistic interpretation required in Monte Carlo simulations breaks down. In the numerical calculations, the suggestion of Barker (1978) was followed: The main peak in K was fit with a Lorentzian thus suppressing the negative oscillations. This procedure has a physical motivation — inclusion of scattering would imply a smearing of all sharp features, and hence the rapid oscillations, which integrate to zero, should be strongly suppressed. The shift and width of the Lorentzian depend on the strength on the electric field, and its orientation with respect to k and q (Reggiani *et al.*, 1988a, for several illustrations). Quite recently, Abdolsalami and Khan (1990) have employed another way of removing the negative parts of the joint-spectral density. Rather than fitting a Lorentzian they argue that a finite box is sufficient to approximate the joint-spectral function. The results of their simulations *differ* from those of Reggiani *et al.* (1988a) underlining once again the need for further investigations.

Collisional Broadening (CB)

In this case the field in (20) is set to zero, and the integral reduces to a convolution in energy space (Kim *et al.*, 1987; Reggiani *et al.*, 1987). Kim *et al.* (1987) solve the spectral density self-consistently for dispersionless-optical-phonon scattering, and use a realistic density of states, while Reggiani *et al.* (1987) work analytically in lowest order perturbation theory, and use a free electron density of states. The resulting joint-spectral densities are similar, and there are no problems with the joint-spectral density going negative. It is tempting to suggest that the difference between ICFE and CB is due to the singular nature of the perturbing electric field — recall that the Hamiltonian includes a scalar potential $U(x) = -x \cdot E$, giving rise to a uniform electric field, which is not bounded, and the spectral function (21) in ω space does not approach uniformly the free spectral density [but does in the distribution sense, Jauho and Wilkins (1984)]. These difficulties suggest that it may be necessary to explicitly account for the finiteness of the sample. A consistent nonlinear theory cannot be formulated without accounting for end effects. This would imply that even the uniform field case should be treated with a spatially inhomogenous theory. The Monte Carlo simulation for the ICFE and CB give similar results. For low fields, $E < 10$ kV/cm, the results differ very little from those obtained with the semiclassical BE while for higher fields there is an increase of carriers both in the low- and high-energy tails of the distribution function. The high-energy electron population enhancement may have some relevance to the onset of impact ionization. In simulations of this kind one should guard against spurious runaway effects that may occur if there is no inherent high-energy cut-off in the problem. In the simulations of Abdolsalami and Khan (1990) the cut-off is introduced by approximating the joint-spectral function by a square function, and therefore they get results largely in agreement with the semiclassical BE.

To our knowledge the only papers where one has attempted to treat ICFE and CB on equal footing are those of Bertoncini *et al.* (1989a, 1989b, 1990), discussed in the next section. Completing the task requires a solution to the nonequilibrium Dyson equation (7). Since ICFE and CB give rise to qualitatively similar results (broadened delta functions) there is no *a priori* justification for treating them separately (Rammer, 1990). More work along these lines would be most welcome.

SPATIALLY INHOMOGENOUS SYSTEMS

All the applications discussed above deal with the uniform field case. However, all microdevices are (almost by definition) extremely inhomogenous, and it is important to ask how much of the above can be generalized to spatially inhomogenous systems. Further, the specification of appropriate boundary conditions represents subtle problems — the contacts,

that allow charge to flow in and out of the device must be included. A proper treatement of a microdevice is a very difficult task and few results have been reported. To get some flavor of the difficulties we write down explicitly the driving term [first commutator in (5)] for a general potential $U(x)$ (for simplicity we consider a one-dimensional system):

$$\left(\frac{\partial}{\partial T}+\frac{p}{m}\frac{\partial}{\partial X}\right) f_W(p,X,T)-\int\frac{dp'}{2\pi\hbar}M(p-p',X)\,f_W(p',X,T)=\left(\frac{\partial f_W}{\partial T}\right)_{coll}, \quad (22)$$

where

$$M(q,X)=\int dx\, e^{-iqx}\left[U(X+x/2)-U(X-x/2)\right] \quad (23)$$

The driving term is nonlocal, which implies considerable difficulties in numerical implementation. A few applications to the resonant tunneling diode have been reported (Frensley, 1987; Kluksdahl *et al.*, 1989). In these calculations the collision term has been treated in the relaxation-time approximation. We would like to issue a warning here — it has been suggested the simple relaxation-time approximation violates particle conservation (Mermin, 1970), and spurious effects may result.

Very recently (Ziep *et al.*, 1986; Jauho and Ziep, 1989; Bertoncini *et al.*, 1989a, 1989b, 1990) an alternative approach has been proposed (however, related procedures have been used earlier, e.g. Herbert and Till, 1982) — rather than working directly with (22) one transforms to a new basis defined by the eigenfunctions of the potential $U(x)$. For the uniform field case, for example, this means that the kinetic equation and the Dyson equation should be 'Airy-transformed' (Bertoncini *et al.*, 1989a, 1989b, 1990). Here we sketch the procedure for a system where the translational invariance is broken in one spatial direction. The eigenfunctions are determined by

$$\left[-\frac{1}{2}\frac{d^2}{dx^2}+U(x)+\frac{1}{2}k_\perp^2\right]\phi_n(x)=\varepsilon_n(k_\perp)\,\phi_n(x) \quad (24)$$

where $U(x)$ is the perturbing one-dimensional potential, for example the position-dependent conduction-band edge found in heterostructures. The index n labelling the eigenfunctions can be either discrete or continuous. The transformed Green's functions are defined by (the self-energies have analogous definitions)

$$G(k_\perp,x,x',\omega)=\sum_{n,n'}\phi_n(x)\,G_{nn'}(k_\perp,\omega)\,\phi_{n'}(x') \quad (25)$$

and the transformed Dyson equation reads

$$G_{mm'}(k_\perp,\omega)=\delta_{mm'}\tilde{G}_m(k_\perp,\omega)+\sum_n\tilde{G}_m(k_\perp,\omega)\,\Sigma_{mn}(k_\perp,\omega)\,G_{nm'}(k_\perp,\omega) \quad (26)$$

with

$$\tilde{G}_n(k_\perp,\omega)=\frac{1}{\omega-\varepsilon_n(k_\perp)+i\eta} \quad (27)$$

The transformation has essentially eliminated the field, or the nonuniform potential, and for many cases of interest the self-energy may well be dominated by the diagonal term [for the uniform field case see Bertoncini *et al.* (1989b, 1990)] in which case (26) is immediately solved. In the diagonal approximation the result is

$$\tilde{G}_{mm'}(k_\perp,\omega)=\frac{\delta_{mm'}}{\omega-\varepsilon_m(k_\perp)-\Sigma_m(k_\perp,\omega)} \quad (28)$$

Equation (28) has one immediate advantage — the resulting spectral density, and joint-spectral density, are positive definite, and satisfy appropriate sum rules. If the diagonality approximation is made in the kinetic equation, one finds that the correlation function can be written in the form

$$G^<_m(k_\perp,\omega)=i\,A_m(k_\perp,\omega)\,f_m(\omega)\ , \quad (29)$$

where the function $f_m(\omega)$ satisfies the following integral equation:

$$f_m(\omega)=\sum_{\eta=\pm1}(N_0+(\eta+1)/2)\sum_{m'}F_{mm'}(\omega+\eta\omega_0)\,f_m(\omega+\eta\omega_0) \quad (30)$$

In the above example, dispersionless optical phonons with energy ω_0 were considered, and the expression for the kernel F is given in Bertoncini *et al.* (1989b, 1990). This type of equation is well suited for numerical evaluation, and numerical results are beginning to emerge. There are a number of ways to continue work along the above lines: (1) other, more complicated structures (for example, the resonant tunneling diode); (2) interpretation of the function $f_m(\omega)$ [it would be tempting to call it a (generalized) distribution function]; (3) development of simulation techniques (Reggiani *et al.*, 1990); and (4) a detailed comparison with results obtained with other techniques.

ACKNOWLEDGEMENT

The author has benefited from many discussions, extended visits, and collaborations on various aspects of quantum-transport theory with collagues too numerous to mention by name. Instead, the author would like to express his sincere gratitude to the following research groups: the Cornell/Ohio State group, the Modena group, the Arizona State group, the Humboldt University group, the Prague group, and the Warwick/Glasgow group.

REFERENCES

Abdolsalami, F., and Khan, F. S., 1990, *Phys. Rev. B*, 41: 3494.
Baranger, H. U., and Wilkins, J. W., 1984, *Phys. Rev. B*, 30:7394.
Barker, J. R., 1973, *J. Phys. C*, 6:2663.
Barker, J. R., 1978, *Solid-State Electron.*, 21:267.
Barker, J. R., 1981, *J. de Physique*, C7: 245.
Barker, J. R., 1987, Talk given during "Nanometer Physics", San Miniato, Italy, March 1987.
Barker, J. R., andFerry, D. K., 1979, *Phys. Rev. Lett.*, 42: 1779
Bertoncini, R., Kriman, A. M., Ferry, D. K., Reggiani, L., Rota, L., Poli, P., and Jauho, A.P., 1989a, *Solid-State Electon.*, 32:1167.
Bertoncini, R., Kriman, A. M., and Ferry, D. K., 1989b, *Phys. Rev. B*, 40:3371.
Bertoncini, R., Kriman, A. M., and Ferry, D. K., 1990, *Phys. Rev. B*, 41:1390.
Chen, L.-Y., and Su, Z.-B., 1989, *Phys. Rev. B*, 40:9309.
Conwell, E. M., 1967, *in* "High Field Transport in Semiconductors," Academic Press., New York.
Craig, R. A., 1968, *J.Math.Phys.*, 9:605.

Danielewicz, P., 1984, *Ann. Phys. (N.Y.)*, 152:239.
Datta, S., 1989, *Phys. Rev. B*, 40:5830.
Frensley, W., 1987, *Phys. Rev. B.*, 36:1570.
Hänsch, W., and Mahan, G. D., 1983, *Phys. Rev. B.*, 28:1902.
Herbert, D. C., and Till, J. J., 1982, *J.Phys. C*, 15:5411.
Hershfield, S., and Ambegaokar, V., 1986, *Phys. Rev. B*, 34:2147.
Hu, G. Y., and O'Connell, R. F., 1987, *Phys. Rev. B*, 36:5798.
Hu, G. Y., and O'Connell, R. F., 1988, *Physica*, 149A:1.
Hu, G. Y., and O'Connell, R. F., 1989, *Phys. Rev. B*, 39:12717.
Jacoboni, C., and Reggiani, L., 1983, *Rev. Mod. Phys.*, 55:645.
Jauho, A. P., 1990, *in* "Nanometer Physics", Jacoboni, C. (Ed.),to be published by Plenum.
Jauho, A. P., and Wilkins, J. W., 1982, *Phys. Rev. Lett.*, 49:762.
Jauho, A. P., and Wilkins, J. W., 1984, *Phys. Rev. B*, 29:1919.
Jauho, A. P., and Reggiani, L., 1988, *Solid-State Electron.*, 31:535.
Jauho, A. P., and Ziep, O., 1989, *Physica Scripta*, T25:329.
Kadanoff, L. P., and Baym, G., 1962, "Quantum Statistical Mechanics", Benjamin, Reading, Mass.
Keldysh, L. V., 1965, *Sov. Phys. JETP*, 20:1018.
Khan, F. S., Davies, J. H., and Wilkins, J. W., 1987, *Phys. Rev. B*, 36:2578.
Kim, K., Mason, B. A., and Hess, K., 1987, *Phys. Rev. B*, 36:6547.
Kluksdahl, N. C., Kriman, A. M., Ferry, D. K, and Ringhofer, C., 1989, *Phys. Rev. B*, 39:7720.
Landauer, R., 1970, *Phil. Mag.*, 21:863.
Langreth, D. C., 1966, *Phys. Rev.*, 148:707.
Langreth, D. C., 1976, *in* "Linear and Nonlinear Electron Transport in Solids," Devreese, J. T., and Van Doren, E., (Eds.), Plenum, New York, p.2.
Langreth, D. C., and Wilkins, J. W., 1972, *Phys. Rev. B*, 6:3189.
Levinson, I. B., 1970, *Sov. Phys. JETP*, 30:362.
Lipavsky, P., Spicka, V., and Velicky, B., 1986, *Phys. Rev. B*, 34:6933.
Mahan, G. D., 1984, *Phys. Rep.*, 110:321.
Mahan, G. D., 1987, *Phys. Rep.*, 145:253.
Mason, B. A., and Hess, K., 1989, *Phys. Rev. B*, 39:5051.
Mermin, N. D., 1970, *Phys. Rev. B*, 1:2362.
Rammer, J., 1990, Nordita Preprint 90/7 S, unpublished.
Rammer, J., and Smith, H., 1986, *Rev. Mod. Phys.*, 58:323.
Reggiani, L., 1985, *Physica*, 134B:123.
Reggiani, L., Lugli, P., and Jauho, A. P., 1987, *Phys. Rev. B*, 36:6602.
Reggiani, L., Lugli, P., and Jauho, A. P., 1988s, *J.Appl. Phys.*, 64:3072.
Reggiani, L., Lugli, P., and Jauho, A. P., 1988b, *Physica Scripta*, 38:117.
Reggiani, L., 1990, private communication.
Sarker, S., 1985, *Phys. Rev. B*, 32:743.
Seminozhenko, V. P., 1982, *Physics Rep.*, 91:103.
Strinati, G., Castellani, C., and Di Castro, C., 1989, *Phys. Rev. B*, 40:12237.
Suhrke, M., 1989, private communication.
Thornber, K, 1978, *Solid-State Electron.*, 21:259.
Vinogradov, A. V., 1986, *Sov. J. Quantum Electron.*, 16:195.
Van Wees, B.J., van Houten, H., Beenakker, C. W. J., Williamson, J. G., Kouwenhoven, L., Van der Marel, D., and Foxon, C. T., 1988, *Phys. Rev. Lett.*, 60:848.
Wharam, D.A., Thronton, T. J., Newbury, R., Pepper, M., Ahmed, H., Frost, J. E. F., Hasko, D. G., Peacock, D. C., Ritchie, D. A., and Jones, G. A. C. J, 1988, *J. Phys. C*, 21:L209.
Ziep, O., Suhrke, M., and Keiper, R., 1986, *Phys. Stat. Sol. (b)*, 134:789.

(17)

NEURAL AND CONSTRAINED INTERCONNECT AUTOMATA

Lex A. Akers

Center for Solid State Electronics Research
Arizona State University
Tempe, Az 85287-6206

INTRODUCTION

Granular nanoelectronics offers the possibilities of spectacular new electronic systems with intelligent computational capabilities. Current computational engines contain over 1 million active devices, and are excellent at storing and retrieving bits of data in precise locations, and at performing complex calculations involving thousands of numbers and millions of operations per second (megaflops). Granular nanoelectronics could allow these systems to scale to billions of active devices operating in the trillions of operations per second (teraflops) range. Such mega-computers would accelerate simulations of multi-dimensional models of transport in semiconductors, complex control functions, data base manipulations, and graphics. However, even these computers will be very limited in their abilities to solve certain problems. For example, computers are not good at extracting information from the data that they store. Hence, they are not good at generalization from information stored in memory, self-organization, autonomous acquisition of knowledge, and learning associations. These are skills a biological system acquires within a few weeks after birth. Real-time object recognition and sensory fusion are additional examples of early acquired skills in biological systems that require pattern recognition speeds far beyond the current or near-term predicted capabilities of computers. Can such computations be implemented with non-organic material, and if so can granular nanoelectronics assist in such an implementation? We will demonstrate that cooperative information processing systems can solve such problems. We will also discuss why granular nanoelectronics has particular characteristics that actually encourages the implementation of these systems.

Conventional systems made with very large scale integrated circuits(VLSI) are designed for either serial or limited parallel data flow and data manipulation. The synthesis and implementation of conventional systems requires that each switching device behave in the same manner within the total system as it does when it is isolated. The full function of the system or integrated circuit (IC) is determined solely by the interconnections used to join the individual devices. A different function can only be implemented in the system by redesigning the interconnections. This requires in most cases both new polysilicon and metal masks, and refabrication. The conventional clear separation of device design from system design thus depends upon being able to isolate each individual device from the environment of the other devices except for the planned effects occurring through the interconnection network. This gross over-simplification is likely to be seriously in error for ultra-submicron dimensioned ultra-large-scale integrated (ULSI) systems, where the isolation of one device from another will be difficult if not impossible to achieve. Instead, one is driven to begin to think about methods of actually using the interactions between devices as a method to accomplish massively paralleled and distributed information processing. This approach is exactly the approach used in *cooperative information processing systems*. It is also the approach currently used to model neural networks(Akers *et al.*, 1988).

Device-device coupling mechanisms are numerous and include such effects as capacitive coupling, of which line-to-line parasitic capacitance is just one example (Ferry, 1982), and wave-function penetration (tunneling and charge spillover). The former is already significant in VLSI, and one cannot limit device-device interactions to just those size scales of the order of the wave function or the inelastic mean free path. The parasitic interaction arising from this line-to-line coupling leads to a direct device-device interaction outside the normal circuit or architectural design (Ferry, 1988). The constraints imposed by device-device interactions will have to be included in future architectural designs of compact ULSI systems, and this will most easily be accomplished if these constraints are reflected in the system theory description of the architecture itself. It is found that the pair-wise nearest neighbor interaction dominates the device-device effects, which suggests two important points. First, the architecture begins to resemble neural networks or cellular automata in format, and these system descriptions hold some usefulness in architectural design (Ferry *et al.*, 1989). Secondly, the nearest neighbor interactions suggest that arrays of devices will begin to look like superlattices of devices (Ferry, 1982), and such lateral surface superlattices can be used to generically look at cooperative effects.

In the next section, we will develop the mathematical description of cooperative, distributed information processing systems. An example will illustrate how such a system can be trained to perform associations and generalizations of information. Then we will describe some VLSI systems which are being used as test vehicles for experimentation in cooperative, distributed information processing systems.

COOPERATIVE INFORMATION PROCESSING SYSTEMS

A cooperative information processing system consists of elementary units interconnected to form a network. A variety of interconnection styles have been investigated, as shown in Fig. 1, with each interconnection pattern referred to as an architecture. The computations performed by the network are highly dependent on the architecture. Each connection has a transmittance value associated with it, referred to as a weight, or in neural systems as a synapse. Therefore, each connection specifies how strong the interactions are between the variables it connects. In general, the magnitude of the weights are independent variables, and hence the architecture is considered to be plastic.

The dynamics of a *conventional system* is usually described by a family of curves. The overall structure of the conventional system is considered fixed, with the variability concentrated into a few parameters. The variations of these parameters leads to another curve in the family of responses. The family of curves shown in Fig. 2 illustrates this point. The system is a distributed interconnection line modeled by a partial differential equation. The various curves result from changing a small set of parameters. The dynamics of *cooperative systems* are much richer and more complex (Scofield *et al.*, 1985). The plasticity of the interconnections, along with the variability of each weight associated with each interconnection, implies entire families of system dynamics are possible. One can view the dynamics as a hierarchy of three levels. The first level is the states of the network, the second level is the magnitude of the weights, and the third is the architecture. This degree of variability provides exciting new types of dynamics, and powerful computational capabilities.

Cooperative Dynamics

The computational capabilities and the mathematical description of cooperative systems has been rediscovered many times (Farmer, 1990). Lately, this paradigm is being used by the neural network community to simulate biological neural networks (Akers *et al.*, 1988). However, these basic dynamics are by no means unique to neural networks. Several other systems such as classifier systems (Compiani *et al.*, 1988), Boolean networks (Kauffman,

1969), systolic and cellular arrays (Akers *et al.*, 1989), the immune network (Farmer *et al.*, 1986), and psychology behavior (Rummelhart *et al.*, 1986), have been described with the same dynamics. This model also has potential application in areas such as economics (Friedman, 1989), game theoretic models (Smith, 1986), and in the modeling of granular nanoelectronics systems.

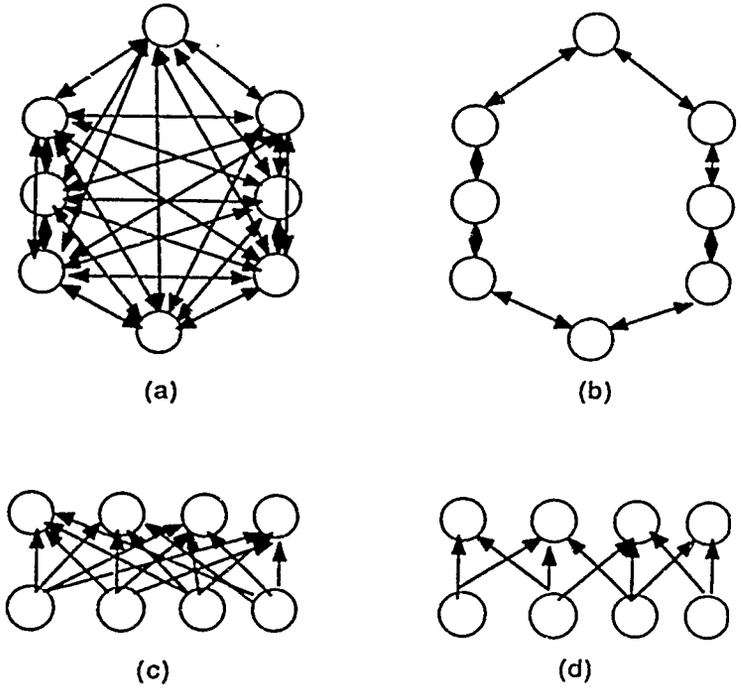


Fig. 1 Neural architectures: (a) Single layer fully interconnected, (b) single layer limited interconnected, (c) multi-layered, fully interconnected, (d) multi-layered, limited interconnected.

These models of course still represent a particular level of abstraction. For example, by reducing the state of a neuron in a neural system to a single number, we are collapsing its properties relative to a real biological neuron. Other much more comprehensive mathematical formalisms exist, for example the Hodgkin-Huxley (1959) model of the membrane currents in terms of the membrane potential. At the level of the Hodgkin-Huxley model, the state of a neuron is described as a function whose evolution is governed by a partial differential equation. Yet at the level of the cooperative description, the neuron can be described by either a set of ordinary differential equations, or rules describing the time evolution of a node, and the weights connecting the nodes. The partial differential equation is not a cooperative model since there are no identifiable interconnection weights. The system just evolves in time. But, describing the system as a collection of nodes allows the nature of the solution to depend critically on the coupling parameters. System behaviors can be observed to emerge which are only loosely coupled to the details dynamics of a node. Therefore, suppressing this level of detail allows us to focus on macroscopic or system level dynamics without completely specifying the nodes.

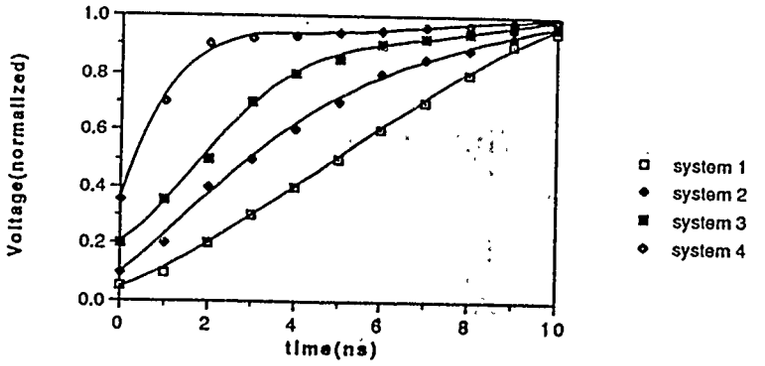


Fig. 2 Voltage vs time for a distributed interconnect line. A different set of parameters were used for each system.

System Description

The structure of any cooperative model is a graph, consisting of vertices, commonly called nodes. The nodes communicate with other nodes through connections. The graph describes the architecture of the system and provides the framework for the dynamics to occur. The connections between the nodes can either be unidirectional, or bidirectional. While in most cases the graphs are best described by a picture, a more formal description is necessary. One common graph representation is a connection matrix. The nodes are assigned an order, which corresponds to the rows and columns of a matrix. The order can be important in that efficient matrix manipulations are available for certain banded matrixes. Non-zero entries identify nodes that are connected. A matrix describing the weights can be combined with the connection matrix if desired. A dense graph has almost every node connected to almost every other node. A sparse graph has most nodes only connected to a small fraction of the other nodes. As will be described later, dense graphs are difficult if not impossible to implement in hardware, and mapping dense graphs into sparse graphs are necessary.

In conventional models the graph of the system is fixed. Only the state of the system is changing, and the state contains sufficient information to determine the future evolution of the system. The only methods to change the dynamics reside in a few system parameters. These parameters are normally fixed during any given experiment, but can change between different experiments. These parameters can be thought of as quantities that change on a time scale much slower than those we are modeling with the system. Cooperative systems can be thought of as giving the parameters explicit dynamics of their own. Normally though, this involves a number of time scales; however, this is not a requirement. The fast scale dynamics, which results in changes of state is usually associated with short-term information processing. This is referred to as the state transition rule of the system. The intermediate scale dynamics occurs from changes in the parameters, which we refer to as weights, and is associated with learning. This is referred to as the learning rule. And lastly, the longest time scale is changes in the graph. This is referred to as architectural dynamics. Architectural dynamics can also be used for learning.

Defining the state, learning rule, and architectural dynamics separately is just a convenience to relate the system dynamics to conventional systems. One could think of the above as just the states of a larger dynamical system with multiple time scales. Exciting possibilities exist when one relaxes these separations to include weights and states changing on similar time scales.

Mathematical Description

The states reside at the nodes in most descriptions of cooperative systems, but this is not a requirement. A state is denoted x_i, where the integer i varies from 1 to the total number of nodes in the network. Also at the node there is a parameter Φ_i, which refers to the threshold of the node. The weights, w_{ij}, refer to the connection strength between nodes i and j. The connection strength determines the interactions between nodes, and hence are an important parameter in cooperative systems. Of course the threshold parameter, Φ_i, also has an influence on this coupling. Thus, it can be misleading to assume the weight is equivalent to the coupling strength. The coupling strength may change at any given time depending on the state of the nodes, and the threshold parameters. Therefore a network independent representation is useful to describe the weights. For continuous systems, the coupling strengths can be expressed in terms of the Jacobian. When the transition rule is an ordinary differential equation of the form

$$\frac{\partial y}{\partial x} = F(x_1,\ldots,x_n) \ , \tag{1}$$

the instantaneous coupling strengths of the weight from node i to node j is the corresponding term in the Jacobian matrix,

$$w_{ij} = \frac{\partial x_i}{\partial x_j} \ . \tag{2}$$

If the weight is positive, the weight is referred to as being excitatory, and if the weight is negative, the weight is referred to as being inhibitory. The average coupling strength is $[w_{ij}]$, where [] denotes some appropriate average.

We are going to use the neural systems paradigm description of cooperative systems. While other systems descriptions are possible, the neural systems description has dynamics that are particularly of interest to granular nanoelectronics. This description requires each node, hereafter referred to as a neuron, to sum the product of the weight and the input signal, subtract a threshold, and apply an activation function to the result.

The response of the neuron can be described as,

$$x_j(t+1) = F\left\{ \sum_{i=1}^{n} w_{ij}x_i(t) - \Phi_j \right\} \ . \tag{3}$$

The function F is called the activation function. F is nonlinear, and used to bound the output. Sigmoidal functions such as tanh are typically used. The instantaneous coupling strength is

$$\frac{\partial x_j(t+1)}{\partial x_i(t)} = w_{ij}\frac{\partial F(u)}{\partial u} \ , \tag{4}$$

where the last partial derivative is of the activation function with respect to its total argument.

Architectures

Of the many types of possible architectures, two have been extensively simulated, and implemented in hardware (Hopfield, 1982, Akers *et al.*, 1989). The first type of architecture is the single layered, fully interconnected structure, as shown in Fig. 1a. A Liapunov or energy function for the network can be defined. A learning rule (Hopfield, 1982) will force the learned vectors to be stable attractor points of the energy terrain. An input is projected onto the energy terrain, and is attracted in terms of the Hamming distance to the nearest energy minimum. The system response is then just the closest learned vector for the input. This architecture functionally performs as an associative memory, but has very limited storage capacity.

The most popular neural architecture is the multiple-layered, fully interconnected between layers architecture, as shown in Fig. 3. The layers of neurons between the input and the output layer are termed referred hidden layers. All of the interconnections in Fig. 3 are feed-fordward. This restriction allows the network to be viewed as simply a network that implements a particular family of nonlinear functions, parameterized by the weights and the thresholds. If the restriction to the connections of being just feed-fordward is removed, much more interesting dynamics can occur. This recurrent network allows any given neuron to change state more than once during a computation. This more interesting dynamics effectively gives the network memory, increasing the number of functions the network is able to implement. However, the feedback adds new problems. For example, a decision must be made on when to stop the computation, which complicates the learning problem (Farmer, 1990). An example of the ability of the multi-layered network to learn to replicate input-output mappings and generalize for undefined input-output pairs is shown below.

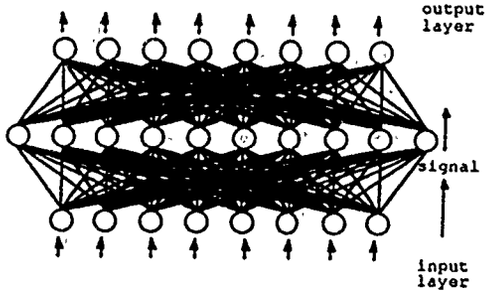


Fig. 3 Layered fully connected architecture.

Local Learning Rules

A given problem is solved by selecting the set of parameters, w_{ij} and Φ_i, to implement a particular mapping. This selection is accomplished with a learning rule. The Hebbian learning rule (Hebb, 1949) is one of the simplest. The learning is non-supervised, in that no information about the output is provided, and only local information is needed to calculate the weights. In theory, the weights are strengthened between neurons with coincident activity. This can be expressed as

$$\Delta w_{ij} = \alpha x_i x_j \ , \tag{5}$$

where α is an amplification factor.

A more advanced unsupervised adaptation (training) rule has been developed by Haghighi (1990). This learning rule was developed specifically for architectures with only local connections between layers. This is also called having a partial receptive field. This is in contrast to the fully interconnected architectures previously described, and the only practical architecture to electronically implement in hardware for large numbers of processing nodes. The learning rule can though be generalized to fully interconnected architectures. This learning rule determines the correlations between the inputs of a node. The nodes at the upper layers of a multi-layer hierarchical network combine the partial receptive fields from the previous layers,

thereby achieving larger receptive fields. One of the methods of accomplishing this objective is that each node should maximize the fraction of its output average power which conveys information about its inputs. In so doing, information loss across many processing layers will be kept to a minimum in a neural network. This function effectively performs feature selection, and allows the output of a node to optimally preserve joint information among inputs.

The correlation coefficient ρ among two random variables indicates their linear dependence and is sufficient to completely characterize the common structure between two jointly Gaussian random variables. It can be easily verified that this measure is invariant to translation, rotation, and scaling of the two random variables. A node consists of several inputs and one output. The output y is the weighted spatio-temporal summation of the inputs. Let x_i (i=1...m) denote the inputs. The weights are labeled w_i). The variance of random variable y is denoted by σ_y^2. The covariance between x_i and x_j is c_{ij}, and their correlation coefficient is ρ_{ij}. The fractional variance index J is defined as the ratio of the signal to the sum of signal and noise variance, where the signal is the correlation among its inputs.

The objective of a node is to adjust the weights $W=[\, w_1 \ldots w_m]^T$ such that the output optimally indicates the presence or absence of linear correlation among inputs. For convenience, we have converted the weight matrix into a weight vector. Without loss of generality, inputs are assumed to be zero mean random variables. Large values of output variance indicate large fluctuations around the mean. Thus, the output variance of a random variable is a measure of its dynamic range requirements. VLSI implementation of a computing node places constraints on the maximum allowable accuracy and resolution of voltages or currents. Taking into account such undesirable properties as arithmetic round-off and circuit parasitics, an upper limit to the distinct possible states of a circuit parameter such as voltage can be specified. Obviously it is always desirable that parameters of interest occupy this maximum allowable range of values. This is precisely the objective of our analytical derivations. In brief, our objective is to select values for a computing node input weights which allow maximum number of possible representations for the desired parameters (input correlation coefficient) at the output.

For simplicity, initially nodes with 2 inputs are considered. The results are then extended to multiple inputs.

Case 1: m=2

A computing node's output variance for the two hypotheses H_1 (uncorrelated inputs) and H_2 (correlated inputs) is

$$\sigma_y^2(H_1) = w_1^2\sigma_1^2 + w_2^2\sigma_2^2 \,, \tag{6}$$

$$\sigma_y^2(H_2) = w_1^2\sigma_1^2 + w_2^2\sigma_2^2 + 2w_1w_2c_{12} \,. \tag{7}$$

The difference between the two output variances is due to the correlation term $2w_1w_2c_{12}$. Our objective is to choose w_1 and w_2 such that the output contribution from the linear correlation among inputs is optimized. This is accomplished by optimizing J where

$$J = \frac{2w_1w_2c_{12}}{w_1^2\sigma_1^2 + w_2^2\sigma_2^2 + 2w_1w_2c_{12}} \,. \tag{8}$$

For $\rho_{12} = \rho_{21} = \rho$, the values

$$w_1 = \frac{K}{\sigma_1} \,, \qquad w_2 = \frac{K}{\sigma_2} \tag{9}$$

will maximize J for $\rho > 0$, and minimize it for $\rho < 0$. On the other hand,

$$w_1 = \frac{K}{\sigma_1} \,, \qquad w_2 = -\frac{K}{\sigma_2} \tag{10}$$

will minimize J for $\rho > 0$ and maximize it for $\rho < 0$. The output variance for (9) and (10) is

$$\sigma_y^2 = 2K^2(1 \pm \rho) \,. \tag{11}$$

Thus, fluctuations in the output variance σ_y^2 are caused by the correlation among inputs. If the weights are selected according to (9), the output of a node will optimally indicate positively correlated input patterns. Such an adaptation mechanism will simultaneously minimize the effect of negatively correlated inputs. The adaptation rule in (10) has the opposite effect.

Case 2: $m \geq 2$

Adaptation rule (9) can be extended to nodes with more than two inputs. An additional requirement is that the inputs must be equally correlated and

$$\rho = -\frac{1}{m+1} \,. \tag{12}$$

Equation (12) guarantees that the output variance is positive.

This learning rule has been used to develop higher levels of representations or symbols of input images. The input image was dynamically blurred by uncorrelated Gaussian horizontal and vertical random movements with a mean of 0, and a variance of 1 pixel. Since computation of the linear correlation coefficient between two inputs requires multiple samples, it is appropriate to specify a time window length for each layer of the pyramid. The generated symbols are compact and can be manipulated as standard symbols. The leaning rule also has been used to detect objects in motion(Akers *et al.*, 1990).

While the Hebbian and the Correlation rule is simple and easy to implement in hardware, the learning rule does not use information about the input-output pairs. The learning rule just clusters the inputs. To exploit the information about the input-output pairs, a learning algorithm called back propagation is used.

Global Learning Rule

Back propagation is a powerful algorithm for adapting layered networks, but this method typically requires a continuous, differentiable activation function such as a sigmoid in order to determine the error gradient at every node in the network. This problem may be circumvented by determining a "desired" state for every node in the network. Le Cun (1985) has described a simple procedure for determining the desired states working backwards beginning with the desired responses in the output layer.

In back propagation(BP), the output error is defined as,

$$E = \frac{1}{2}\sum_{C}\sum_{J}(y_j - d_j)^2 \ , \tag{13}$$

where C is an index over the number of input-output pairs, J is the index of the output units, y_j is the actual output value and d_j is the desired output value at each output unit. The output of each processing element is a nonlinear function of its input, x_j,

$$y_j = \frac{1}{1+\exp(-x_j)} \ . \tag{14}$$

where the input x_j is a weighted sum of the outputs of units in the previous layer. The weighted transmittance value between the previous output y_i and the next input x_j is w_{ij}. x_j is expressed as

$$x_j = \sum_i y_i w_{ij} \ . \tag{15}$$

The use of a differentiable node transfer function allows the use of the chain rule to determine the change in the output error resulting from each transmittance value change. The procedure for adaptation involves applying input vectors at the input layer and accumulating the difference between the actual and desired output vectors for the entire example set. The change for each transmittance value as a function of total output error is

$$\Delta w(t) = -c\frac{\partial E}{\partial w(t)} + \alpha\,\Delta w(t-1) \ , \tag{16}$$

where t is incremented by 1 for each sweep through the whole set of input-output cases, c is a proportionality constant which determines the magnitude of transmittance adjustment at each step, and α is a momentum factor which determines the contribution of the previous gradients to the transmittance change. This procedure is repeated until the total accumulated error falls below some minimum value.

For networks using binary threshold elements, additional considerations are necessary. The desired values for the states of the internal processing elements as well as the output nodes must be defined. A procedure has been described (Plaut *et al.*, 1986) in which internal desired states are calculated as

$$d_i = 1 \ , \quad \text{if } \sum_j w_{ij}(2d_j - 1) > \mu_i \ ,$$

$$d_i = 0 \ , \quad \text{otherwise,} \tag{17}$$

where j is the index over the succeeding layer of processing elements and i is the index of the layer under consideration. μ_i is an arbitrary threshold for the generation of the backward-propagated desired state. A "criticality" term for each desired state associated with each hidden unit is generated. This ratio is a numerical representation of the "confidence" associated with the desired state selected for each hidden unit and is used to modify the weight update. The transmittance change for any connection then becomes,

$$\Delta w_{ij}(t) = \varepsilon(d_j - y_j)c_j y_i + \alpha\Delta w_{ij}(t-1) \ , \tag{18}$$

where c is the criticality of any node in a given layer, defined by

$$c_i = \frac{\left|\sum_j w_{ij}(2d_j - 1)c_j\right|}{\sum_j \left|(2d_j - 1)c_j\right|} \ . \tag{18}$$

In this way, desired states which have a high criticality will exert the greatest influence on the modification of a given connection.

Each weight in this type of network develops in a way that not only converges on the training set, but also constrains all units to the saturated endpoints of their activation functions. In addition, since all units are forced to the endpoints, each unit ends up acting as a simple threshold element implementing a boolean function. For this reason, binary threshold elements may be used to replace the continuous sigmoids in feedforward operation after training has completed.

The following example shows how this algorithm finds a local minimum solution to the XOR problem. Figure 4 shows the "prewired" routing through the sparse array and the necessary boolean logic functions formed by the unique weight patterns at each node. Figure 5 shows the weight pattern formed using back propagation of desired states. Of interest is the similarity and location of the boolean functions formed within the net. The network has learned the input-output mappings, and will give correct generalization for inputs not in the training set.

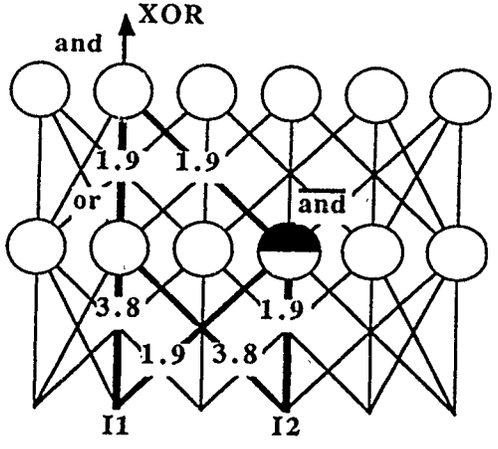


Fig. 4 Prewired XOR algorithm.

Additional considerations are required for sparsely connected networks adapted with BP (Walker *et al.*, 1988). Efficient internal network representations are necessary for generalization of a given problem in a neural network. A general rule for accurate generalization with full interconnection between layers is that they must contain a minimum number of hidden units necessary to encode the invariant properties of the input training set. In a similar fashion, sparsely connected networks which generalize have a minimum number of active internal units. Learning algorithms for these networks must not only develop internal abstractions with a minimum number of processing elements, but must also route required signals to the proper locations within the network. Low fan-in and the use of random initial

weight values serves to reduce the probability that a signal necessary for a given hidden unit to generate a required abstraction will be routed in the interconnections. For this reason, the generation of a "high" desired state within a hidden unit must be made artificially difficult. In this way, the majority of connections will be eliminated from the forward signal path unless they are needed. Fortunately, the presence of a positive threshold value in each processing element means that each node is hard to turn on and easy to turn off. Another necessary *a priori* consideration required is that the inputs must be arranged in a way that ensures that they can be connected to all of the correct output nodes with the number of layers provided.

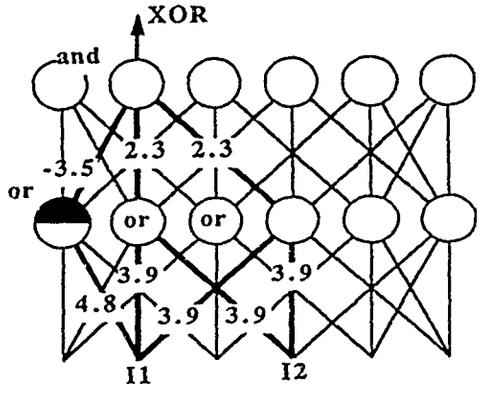


Fig. 5 Adapted XOR network.

Sparsely connected neural networks are generally more prone to local minima during training. Fully-connected neurons providing more signal paths and a richer, more redundant set of internally formed abstractions more easily find a near-optimal solution (in the least mean square sense) in weight space. Since gradient descent is basically a shortcut through an NP-complete search in weight space, greater redundancy and overlapping of internal representations of information improve the probability of convergence to a near-optimal solution of the training set. Also sparse networks provide less redundancy and have a more difficult time forming the necessary internal abstractions since now routing of signals to the correct hidden unit forming a needed abstraction becomes important. An advantage however, is that since fewer points in weight space are available to converge on the given training set, the probability is higher that the given weight set has captured the invariant properties of the desired network function.

VLSI IMPLEMENTATIONS OF NEURAL NETWORKS

Neural networks is a rapidly growing field primarily because of the promise of solutions to problems that have continued to confound computer science and artificial intelligence. In the next few years, neuromorphic systems will provide solutions to problems requiring parallel searches though spatial and spatial-temporal information, allow implementation of self-organized associative memories, and permit systems to autonomously collect knowledge from observations of it's environment. Applications of these systems to problems of sensor processing, knowledge processing, natural language, vision, and real-time control offer new markets for electronic systems.

Neural Architectures for VLSI

Hardware implementation of the equations that describe a cooperative information processing system span a wide range of approaches. At one end of the spectrum are hardware accelerators which enhance the processing speed of the host computer. The hardware for this method is easy to install and use, relatively inexpensive, and offers 2 to 3 orders of magnitude of enhanced processing speed. The other end of the spectrum is the design and use of custom Very Large Scale Integrated (VLSI) chips. This approach contributes an additional 3 to 4 orders of magnitude of enhanced processing speed over accelerator boards. These VLSI systems have low power consumption, are small in size, and are inexpensive in large volumes. Only the VLSI chips offer the potential of matching the processing speed, high density, and low power consumption of complex biological systems.

Two disparate groups of workers are presently engaged in VLSI chip implementations of neural networks. The first is committed to electronic-based implementations of neural networks, and use standard or custom VLSI chips. The driving force for this group is to implement the neural system equations, not to design novel integrated circuits. The electronic systems are only a vehicle to obtain the desired system behavior and through-put. The second group desires to build fault-tolerant, adaptive VLSI chips, and is much less concerned with whether the design rigorously duplicates the neural models. This group is composed of VLSI designers who realize the opportunities of obtaining biological behavior in hardware. They see numerous problems for VLSI designers on the horizon, and recognize that nature has found solutions to many of these problems. VLSI technology currently has the capability of fabricating systems which border on biological complexity. The central problem in the construction of a VLSI neural network is that the design constraints of VLSI differ from those of biology (Walker *et al* ., 1989). In particular, the high fan-in/fan-out of biology imposes connectivity requirements such that the VLSI chip implementation of a highly interconnected neural network of just a few thousand neurons would require a level of connectivity which exceeds the current or even projected interconnection density of Ultra Large Scale Integrated (ULSI) systems (Ferry *et al.*, 1988a). Fortunately, highly-layered, limited-interconnected networks can be designed that are functionally equivalent to highly-connected systems (Akers *et al.*, 1989). Figure 6 illustrates the limited interconnected, multi-layered architecture used to implement large numbers of neurons in hardware.

Electronic Implementations

One classical problem solved by neural networks, and described earlier, is the retrieval of a complete vector from partial knowledge of the vector, hence an associative or content addressable memory. The initial condition represents part of the complete set of information to be retrieved, and the system will relax to the closest complete set. Other problems which have been implemented with this type of network are optimization problems. Examples are the traveling salesman problem (Hopfield *et al.*, 1985), and an analog-to-digital converter (Tank *et al.*, 1986).

A team at Bell Laboratories (Graf *et al.*, 1986) has designed, fabricated, and tested a number of neural network electronic circuit implementations. These circuits include a thin-film array of read-only resistive synapses (Hubbart *et al.*, 1986), an associative memory with 256 neurons on a single chip using a combination of analog and digital VLSI technology plus custom microfabrication process (Jackel *et al.*, 1987), and an array of programmable weights and amplifiers serving as electronic neurons (Graf *et al.*, 1987).

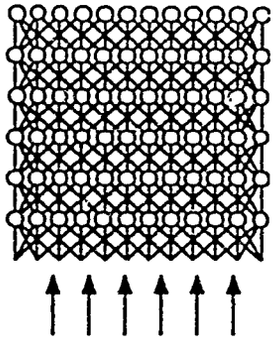


Fig. 6 Multi-layered, limited interconnect architecture.

Connections between neurons, or the weights, form a major component of an electronic neural network. A virtue of resistive weights is that they can be packed very densely. Using layered thin films, Bell-Lab have made resistors which are considerably smaller than the smallest state-of-the-art transistor and which are VLSI process compatible. A 22 x 22 array of microfabricated resistors was fabricated and programmed to serve in a Content Addressable Memory (CAM) mode. The resistive array was made by sandwiching amorphous silicon between tungsten as shown in Fig. 7.

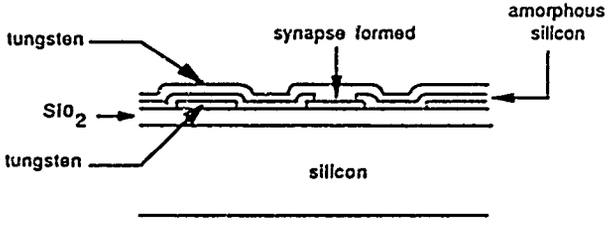


Fig. 7 Bell Labs microfabricated resistor.

The upper tungsten layer serves as the input; the amorphous silicon serves as the resistive weight material since current from axons (the lower tungsten layer), to inputs, had to pass through this silicon. The array has only two weight strengths, no connection or connection through the amorphous silicon of about 100 kΩ . The array was coupled to an array of twenty-two CMOS inverters to form the complete network. The network was coded to store four 22-bit words in a distributed fashion. The network was able to find the best match in hamming distance for a test word to memory word that differed from the test word by 4 or less bits.

Another fabricated chip, which uses the microfabricated resistors, is an Electronic Neural Network (ENN) memory with 256 neurons. Amplifiers with inverting and noninverting outputs are used for the neurons to make inhibitory and excitatory connections. The ENN consists of fully connected array of amplifiers. The reduction in the connection area by using microfabricated resistor is thus a major contribution.

The disadvantage of the resistive array is that once it is fabricated, the synaptic strength cannot be changed. Graf, from Bell-Labs, designed a self contained VLSI chip with fifty-four neurons and about 3000 programmable synapses. The synapses can either be excitatory, inhibitory, or open (no connection). A novel aspect of the design is that each synapse actively drives the axon, reducing the drive requirement of the neuron amplifiers as compared to a circuit with simple resistive synapses. Figure 8 shows a programmable synapse connecting the output of neuron j to neuron i.

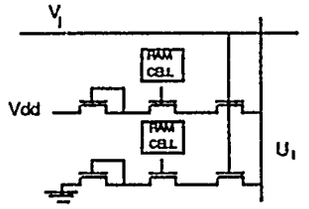


Fig. 8 Bell-Lab's Programmable Weight.

The two RAM cells determine whether the connection is inhibitory (gate towards ground enabled), excitatory (gate towards V_{DD} enabled), or open (both gates disabled). When the out of neuron j is high, current flows into or out of the input line of neuron i, depending on how the two RAMs are loaded. The best configuration of this chip for performing associative recalls (Graf *et al.* 1987), is the one shown in Fig. 9. The neurons are partitioned into vector neurons, that accept the input keys, and label neurons, that identify the best-match memory. The input key passes through the vector neurons and then is tested by the label neurons whose synapses are templates for the memories they represent. The label neurons are excited in proportion to their template match to the input vector. The label neurons are connected through a matrix to the input vector. The label neurons are connected through a matrix of mutually inhibitory connections that act as a winner-take-all circuit; that is, the label whose template is the best match to the input vector is turned on, all the others are turned off. The label neuron, in turn, feeds back to the vector neurons through a template that impose that cell's memory word on the vector neurons, accomplishing associative recall. The chip uses standard 2.5 µm CMOS fabrication technology, but the calculation is a mixture of analog and digital processing.

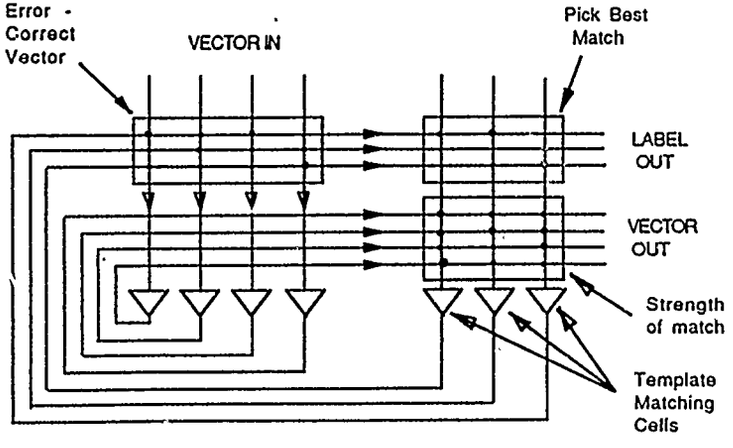


Fig. 9 Configuration for Associative Recall.

In an investigation at Arizona State University, on the usefulness of neural networks in performing computation, a digital SNS capable of varying its interconnection neighborhood and interconnection weights has been designed and fabricated at MOSIS (Akers *et al.*, 1988a).

The custom designed VLSI chip was among the first implementations of a digital neural networks based on a programmable interconnection pattern. The digital neural network consists of 12 neurons in a systolic array architecture. Each neuron performs the evaluation function, as in the Hopfield model, with a programmable threshold. All weights are limited to the range ±1, with a resolution varying between four and eight bits of representation. Each processing element contains four sections: the router, the memory, the accumulator, and the control unit. Figure 10 is a block diagram of one cell. All actions are synchronized by an external clock and control signals. The router is responsible for directing each neuron's output state to its neighbors. A neuron has four connections to each of its four physical neighbors. The router is composed of flip-flops, which implement a two-dimensional shift matrix. The input to the flip-flop is controlled by a four-input multiplexer. This allows each cell to direct its own output to any one of its four nearest neighbors. The selection of the neighbor receiving the signal is controlled by direction control signals. All flip-flops of the shift matrix are controlled by the same signals. On each cycle of the shift clock, information moves in any one of four directions. This signal routing scheme is similar to that found in the connection machine (Hillis, 1985). Each cell is designed to be autonomous, except for the external clock and control signals. Many cells can be connected together to form a matrix of neurons. The edges of the matrix may be connected to latches that can be read or written by a microprocessor. This test vehicle is being used to study the effects of interconnectivity on the transition from local to distributed data storage, digital learning algorithms, and system robustness against individual device failure. The chip was fabricated using 3 μm CMOS double metal technology and packaged in a 84 Pin Grid Array.

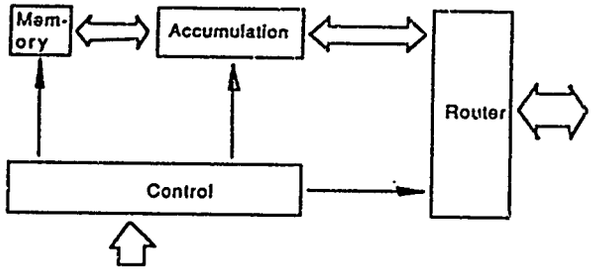


Fig. 10 Simplified Functional Block of a Neural Cell.

Electronic neural networks implemented in silicon are limited by VLSI constraints, and not neural constraints. A principal VLSI constraint is cost, and the cost of a chip is directly related to it's die area. By exploiting the natural functions available with analog circuits, like summation, less die area is consumed than with a digital implementation of the same function. Of course, analog VLSI designs have additional problems not found in digital systems. For example, power dissipation can be a serious problem for large chips. While digital circuits are relatively insensitive to varying device parameters, in most cases analog circuits are very sensitive to these variations. Analog neural-like circuits must be designed not to rely on the absolute behavior of each individual transistor, but on the cooperative behavior of large numbers of devices.

To take advantage of the compactness of analog circuits, and the device process tolerance associated with digital circuits, a hybrid analog-digital circuit has been developed (Akers *et al.*, 1988b). This cell is designed for use in limited-interconnect architectures, so tens of thousands of neurons can be fabricated on a die. Figure 11 is the circuit diagram of the limited-interconnect analog neural cell and Fig. 12 is the timing diagram. The operation of the cell is as follows. Weights, W1, W2, and W3, are stored dynamically on the gates of transistors T1, T2 and T3. Notice only PMOS transistors (T18, T19, and T20) are used to pass and isolate the weights instead of transmission gates. This is allowable since only weights above the device threshold are important, and hence a degraded low state voltage has no effect on circuit performance. For inputs of 5 volts, the drain parasitic capacitors of T7, T8, and T9 are charged by current flowing through the pass transistors T4, T5, and T6, to a voltage equal to the weight minus the device threshold voltage. For inputs of 0 volts, the pass transistors will allow the capacitors to discharge. While exact multiplication of the input and the weight is not done, shifting of the circuits' logical threshold voltage and modifying the training algorithm compensates for this behavior. Once the storage capacitor in each branch is charged, clock Φ1 is turned off to isolate the signal from the input. Turning on Φ2 allows the signals to be analog summed and compared to the logical threshold of the first inverter. The circuit on top of the PMOS pull-up device allows the logical threshold and hence the neural threshold to be set at a voltage lower than $0.5V_{dd}$. The output inverter restores the output voltage level and drives the next stage. Since this circuit uses only positive weights, a shunting transistor T11 is used to provide inhibition. Arrays made with this cell can perform as a complete logic family. Figure 13 shows a circuit simulation of the cell.

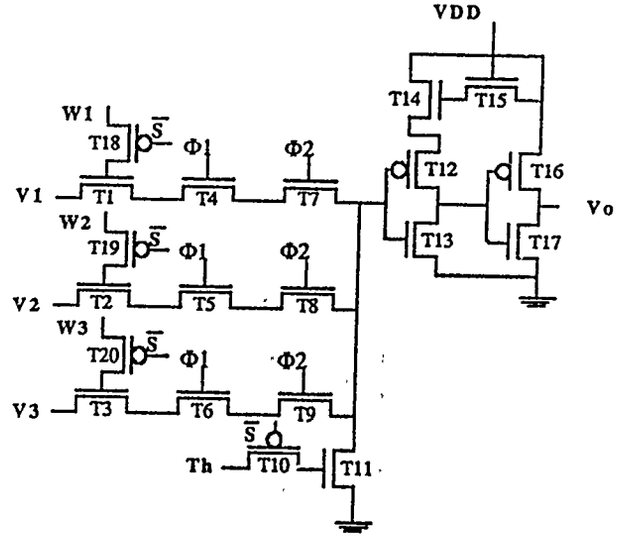


Fig 11 Analog synthetic neural cell.

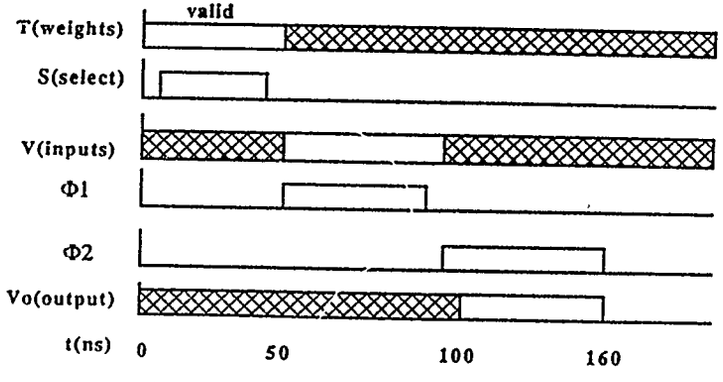


Fig 12 Timing Diagram for analog neural cell.

* A system formed by replicating the analog cell operates in the following manner. To keep the number of I/O pins at reasonable numbers as the system is expanded, we share the input, weights, and output pins. For the implementation of 512 neurons in an array 32 wide by 16 long, 32 lines are used for the weights, and I/O. The four unique weights are multiplexed to the single line running to each cell. The whole array can be loaded in approximately 10 µs, an order of magnitude faster than the discharge time constant for an individual gate. Figure 14 shows the block diagram for the chip. For efficient data flow through the layers, the clock lines are interchanged in every other row. After the weights have been loaded, the inputs are loaded, and the output vector ripples through the layers. Using this architecture, a 512-element, feedforward neural IC has been designed (Akers *et al.*, 1988b).

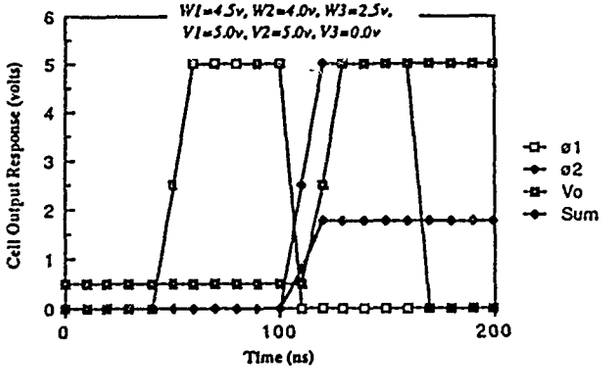
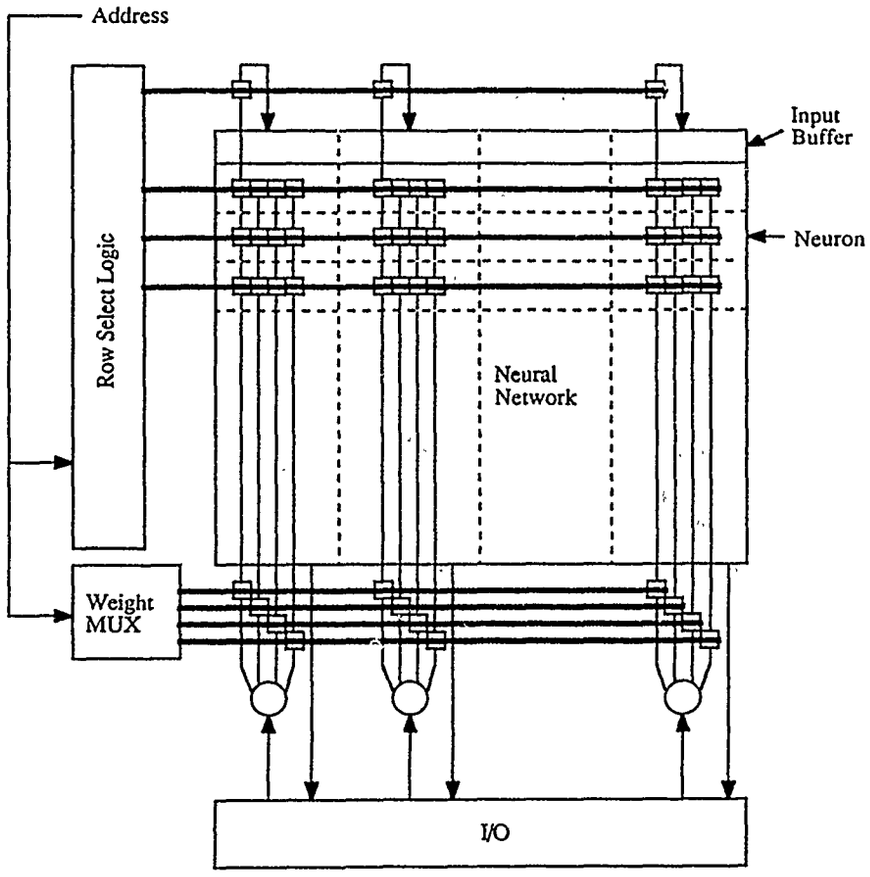


Fig 13 Circuit simulation of cell.

A new neural cell has recently been developed by Hasler (1989). The cell, called the Pulsed Neural Cell also uses pulses to achieve the requirements of an electronic neural system. A block diagram of the pulse neural cell is shown in Fig. 15. This circuit performs quasi-linear, four-quadrant multiplication without large penalties in area or speed. The neuron and synapse are still compact, with the pulsed neural cell consuming only 2.5 times the area of the Analog Neural Cell. The neuron is capable of either linear or sigmoidial outputs which can be programmed by the training algorithm.

The weight strengths are stored dynamically on a parasitic gate capacitor as analog values. These analog values will decay with time, and must be refreshed The basis for the analog refresh circuit comes from DRAM refresh circuits, where an analog value is read, converted to a quantitized value, and then written back to the storage capacitor. However, multivalued analog signals require more complex circuits. A unity gain buffer after the storage capacitor is needed to send the analog signal back to the refresh circuitry. Since the buffer will be part of another subcircuit and therefore not designed to drive long transmission lines, considerable time (100 ns) will be required to accurately sense the weight value, but this will consume only a small amount of the 100 to 500µs computation time. The element used to multiply the weight and the input signal is analogous to a stage in a pulse multiplier, and is shown in Fig. 16.

The outputs of the pulse multipliers are summed in a capacitor. The analog voltage representing the summed products of the inputs and the weights is converter to a pulse by the circuit in Fig. 17. The output of the neuron can be trained to be either a linear or nonlinear function of the input. This capability allows the circuit element to act either as a linear summer to increase the correlations in the input vector and effectively increase the fan-in, or as a standard neuron. Both the analog voltage and the resulting pulse width is shown in Fig. 18.



Common connection pad.

Weight, input, output lines, METAL2

Control lines (MUX control and Select), METAL1

Signal flow control element (Pass Transistor or Transmission Gate)

Fig. 14 Chip Architecture.

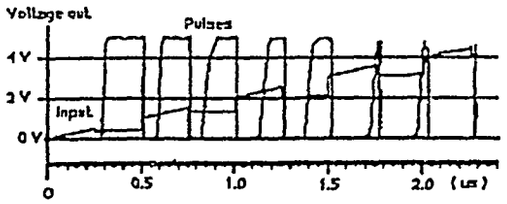


Fig. 18. Analog input to pulse width output circuit.

THE FUTURE: THE ULTIMATE INTEGRATED CIRCUIT

The IC of the future must be based on a regular array of regularly connected elements, or it is unlikely that we will ever be able to complete the initial design. The connections however will be plastic. While conventional computational engines will still be needed, the control of such chips will have to be adaptive. These adaptive controllers will have to learn by example, rather than be programmed, and be highly fault tolerant. Highly paralleled, distributed, cooperative information processing systems have all of the previously mentioned characteristics. The challenge is to use granular nanoelectronics systems to implement such systems.

ACKNOWLEDGEMENTS

Paul Hasler, Mark Walker, and Siamack Haghighi are thanked for contributing part of the pulse neural cell section, the back propagation, and the local training rule section, respectively.

REFERENCES

Akers, L.A., Highighi, S., and Rao, A., 1990, in "Adavanced Neural Computers", Elsevier Science Publishers, New York (in Press).

Akers, L.A., Walker, M, Ferry, D.K., and Grondin, R.O., 1988, Proc. of the 22 Annual Aislomar Conf. on Signals, Systems, and Computers.

Akers, L.A., Walker, M, Ferry, D.K., and Grondin, R.O., 1988a, in "Neural Computers", Springer-Verlag, Berlin, pp. 407-416.

Akers, L.A., Walker, M, Ferry, D.K., and Grondin, R.O., 1989,. in "VLSI for Artificial Intelligence", Kluwer Academic, Norwell, MA, pp. 407-416.

Akers, L.A. and Walker, M., 1988b, Proc. IEEE Neural Net. Conf., San Diego III-151.

Compiani, M., Montanari, D., Serra, R., and Valastro, G., 1988, in "Parallel Architectures and Neural Networks".

Farmer, J., 1990, Los Alamos Technical Report LA-UR-90-228.

Farmer, J., Packard, N., and Perelson, A., 1986, Physica, 22D:187.

Ferry, D., 1982, in "Adv. in Electronics and Electron Phys.", 58:311.

Ferry, D., 1988, in "The Physics of Submicron Semiconductor Devices", Plenum, New York, pp. 503-520.

Ferry, D.K., Akers, L.A., and Greeneich, E., 1988a, "Ultra-Large Scale Integrated Microelectronics", Prenctice-Hall, Englewood Cliffs, NJ.

Ferry, D.K., Akers, L.A., and Grondin, R.O., 1989, in "Physics of VLSI at the 0.5 to 0.05 Micron Dimension", Academic Press, New York, pp. 377-412.

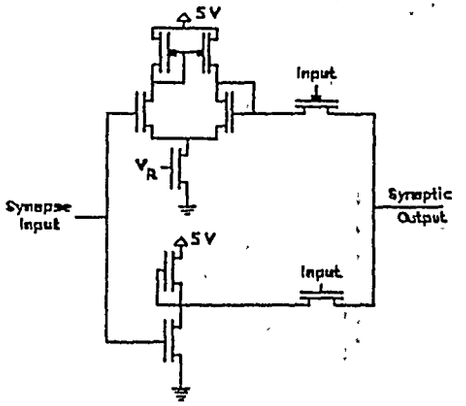


Fig. 15 Block diagram of pulse neural cell.

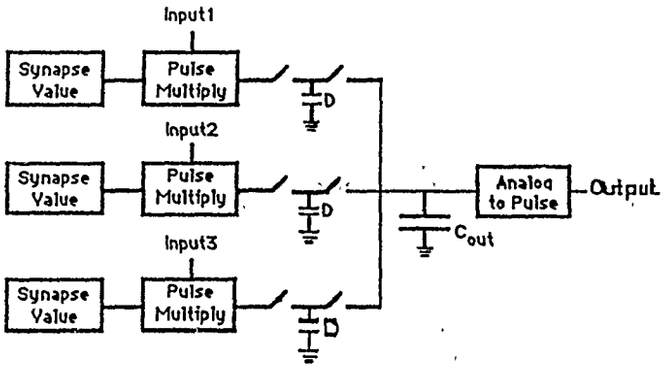


Fig. 16 Pulse multiplier circuit.

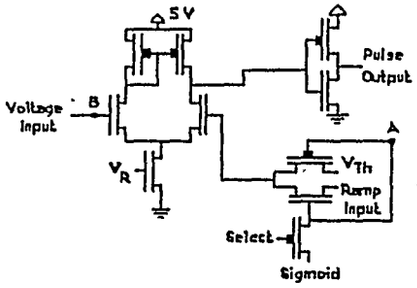


Fig. 17 Analog to pulse converter.

Friedman, D, 1989, Evolutionary Games in Economics, Technical Report, Stanford.
Graf, H. P. and DeVgar, P., 1987, in "Advanced Research in VLSI : Proceedings of 1987 Stanford Conference", Cambrigde, MA, MIT Press.
Graf, H. P., Jackel, L., Howard, R., Howard, B., Stranghn, B. Denker, J., Hubbard, W., Tennant, D., and Schwartz, D., 1986, AIP Conf. Proceeding, Snowbird 151:182.
Haghighi, S., and Akers, L.A., 1990, Proc. IEEE Neural Net. Conf. San Diego (in Press).
Hasler, P., 1989, Proceedings of the Wescon Conference 1989.
Hebb, D.O., 1949, "The Organization of Behavior", New York, Wiley Interscience.
Hillis, D., 1985, "The Connection Machine". MIT Press, Boston.
Hodgin, A.L., and Huxley, A.F., 1959, Proc. Roy. Soc.(London) B148:1.
Hopfield, J., 1985, Biol. Cybern., 1:55.
Hopfield, J.,1982, Proc. Nat. Acad. Sci. 79:2554.
Hubbard W., Schwartz, D., Denker, J., Graf, H., Howard, R., Jackel, L., Straughn, G., and Tennant, D.,1986, Neural Networks for Computing, AIP Conf. Proc. 151.
Jackel L. D., Graf, H. P. and Howard, R. E., 1987, Applied Optics, 26:23.
Kauffman, S., 1984, Physica, 10D.
Le Cun, Y.,1985, In "Disordered Systems and Biological Organization", Springer-Verlag, New York.
Plaut, D., Nowlan, S., and Hinton, G., 1986, Computer Science Tech. Rep. Carnegie-Millon.
Rummelhart, D., and McClelland, J., 1986, "Parallel Distributed Processing", Vol. 1, MIT Press, Cambridge.
Scofield, C., and Cooper, L.N., 1985, Contemp. Phys. 26(2):125.
Smith, J., 1986, Physica, 22D.
Tank, D., and Hopfield, J , 1986, IEEE Trans. Cir. Syst. 33:533.
Walker, M. and Akers, L.A., 1988, Proc. IEEE Conf. Computers and Comm. Phoenix, p.19.
Walker, M. Hasler, P., and Akers, L.A., 1989, Phoenix Conference on Computers and Communications.

(19)

OPTICAL PROPERTIES OF SHORT PERIOD SUPERLATTICES

C. Hamaguchi, T. Matsuoka and K. Taniguchi

Department of electronic Engineering
Osaka University, Suita City, Osaka 565, Japan

INTRODUCTION

The purpose of this lecture is to describe the optical properties of short period superlattices. Optical properties have been known to represent the electronic properties of the semiconductors and thus the energy band structures have been probed by measuring the optical properties. This procedure, especially modulated reflectance spectroscopy, has been successfully used to determine the energy band structures of bulk semiconductors, such as Ge, Si, GaAs, GaP, and so on (Cardona, 1969; Seraphin, 1972; Aspnes, 1980). As is well known, molecular beam epitaxy (MBE) enables us to produce various kinds of semiconductor structures such as quantum wells (QWs), superlattices (SLs), and heterojunction devices. These new structures modify the electronic structures by using energy band discontinuity at the interfaces. Therefore their optical and electrical properties are modified in the heterostructures (Ando *et al.*, 1982). For example, electronic states are quantized to form subband structures in QW structures and minibands in SLs, resulting in uniqueness of their density of states, and thus optical properties are quite different from those of bulk materials.

As is well known, the reflectivity of a semiconductor as a function of incident photon energy consists of several peaks arising from various critical points, such as E_0, E_0', E_1, E_2 and their spin orbit splittings. These peaks are sometimes smeared out by a structureless background (Cardona 1969). The discovery of modulation spectroscopy elucidated fine structures of the optical properties and gave very accurate determination of the critical points in the bulk semiconductors (Cardona, 1969; Seraphin, 1972; Aspnes, 1980). The basic idea of modulation spectroscopy is to observe derivative spectra with respect to incident photon energy, which is achieved by modulating the physical parameters of the material by an external perturbation such as stress, electric field, temperature and so on, and by detecting a change in the optical properties by using phase sensitive detection.

In this lecture, we place our main emphasis on the optical properties of heterostructures, QWs and SLs, and show modulation spectroscopy is very powerful to investigate the optical properties of these structures with high accuracy. First, we describe electronic structures of several heterostructures and optical properties of solids. Second, we review modulation spectroscopy. Finally, we present experimental results. This overview will give us well refined understanding of optical properties of small structures. In particular, a difference in the optical properties between three, two and one dimensional structures will be discussed. In the next section, energy band calculations by the simple Kronig-Penney model and the empirical tight-binding theory will be given. The former model is very useful for understanding the basic idea of superlattices and the latter is more sophisticated and helpful for the discussion of optical transitions in SLs. In the third section, we deal with optical properties of solids and give a review of modulated reflectance spectroscopy. Then we present experimental procedures and experimental results of photoreflectance spectra in QWs and SLs. We will show how to understand the photoreflectance data and a comparison with photoluminescence data. The results give an information of the cross over of the direct and indirect transitions in short period superlattices.

ENERGY BAND STRUCTURES

Energy band structures of SLs have been calculated by using various methods, such as a simplified Kronig-Penney method (Esaki *et al.*, 1970), envelope-function approximation (Bastard, 1981, 1982), and more elaborated methods (Schulman *et al.*, 1979; Harrison, 1981, Nakayama *et al.*, 1985; Drummond *et al.*, 1987; Yamaguchi,1987; Jian-Bai-Xia, 1988; Wei *et al.*, 1988; Eppenga *et al.*, 1988; Gopalan *et al.*, 1989; Fujimoto *et al.*, 1989). In this section we will present two methods, Kronig-Penney method and empirical tight-binding method. The former method is very useful for understanding the basic idea of the superlattices, while the latter will give more detailed discussion on the energy bands and optical properties although the method has some restriction when we discuss higher energy band transitions. Since we are interested in the optical properties of SLs, we need detailed information on the momentum matrix elements or oscillator strength for the interband transitions, which will be obtained by an analysis based on the tight-binding energy band calculations.

Kronig-Penney Model

First, we assume that the energy bands structures are known and the energy band minima in which we are interested in are well expressed in terms of parabolic dispersion with a constant effective mass m*. This means that the energy band is given by

$$\varepsilon = \frac{h^2k^2}{2m^*} , \qquad (1)$$

where ε is the energy of the electron with wave vector k measured from the band edge, h is Planck's constant divided by 2π, and m^* is the effective mass, where we assume an isotropic (spherical mass).

Superlattices are formed by growing alternative layers of different semiconductors. In this lecture we deal with $(GaAs)_m/(AlAs)_n$ SLs consisting m layers of GaAs (thickness a) and n layers of AlAs (thickness b) as a typical example. As is well known, the difference in the electron affinity between the two materials results in a conduction band discontinuity as shown in Fig. 1.

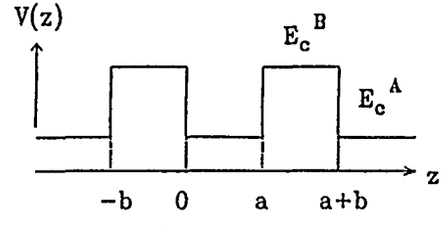


Fig. 1 Superlattice of A (GaAs) with thickness a and B (AlAs) with thickness b.

The energy band structure of the superlattice shown in Fig. 1 is obtained by solving Schrödinger equation within the frame of the effective mass approximation

$$\left[-\frac{h^2}{2m^*}\nabla^2 + V_c(r)\right] F(r) = \varepsilon F(r) \ , \tag{2}$$

where $V_c(r)$ is the periodic potential of the superlattice, ε is the energy and $F(r)$ is the wave function. The superlattice potential has a periodicity of (a+b) and given by

$$V_c(r) = \begin{cases} E_c^B - E_c^A \equiv V_c \ , & \text{in barrier region,} \\ 0 \ , & \text{in well region,} \end{cases} \tag{3}$$

where E_c^A and E_c^B are the conduction band edges of the two semiconductors A (GaAs) and B (AlAs). The wave function $F(r)$ is understood as the envelop function and connected with the Bloch function $\psi_c(r)$ and actual wave function $\Psi(r)$ through the following equation

$$\Psi(r) = F(r)\,\psi_c(r) \ . \tag{4}$$

As seen in Fig. 1, the potential is a function of z, the growth direction of the SL and usually taken to be the (001) direction. This means that the wave function is written as $F(r) = X(x)Y(y)Z(z)$, and that the electron energy is given by $\varepsilon = \varepsilon_x + \varepsilon_y + \varepsilon_z$, which satisfy

$$\begin{aligned} &-\frac{h^2}{2m^*}\frac{d^2}{dx^2}X(x) = \varepsilon_x X(x) \ , \\ &-\frac{h^2}{2m^*}\frac{d^2}{dy^2}Y(y) = \varepsilon_y Y(y) \ , \\ &\left[-\frac{h^2}{2m^*}\frac{d^2}{dz^2}Z(z) + V_c(z)\,Z(z)\right] = \varepsilon_z Z(z) \ . \end{aligned} \tag{5}$$

The envelope functions X(x) and Y(y) are given by

$$\begin{aligned} X(x) &= \exp(ik_x x) \ , \quad \varepsilon_x = \frac{h^2k_x^2}{2m^*} \ , \\ Y(y) &= \exp(ik_y y) \ , \quad \varepsilon_y = \frac{h^2k_y^2}{2m^*} \ . \end{aligned} \tag{6}$$

From (5), we find that the envelope function Z(z) is expressed as

$$Z(z) = U(z)\exp(ik_z z) \ , \quad U(z+L) = U(z) \ , \tag{7}$$

from the analogy of Bloch function, where k_z is wave vector of an electron in the direction z, and U(z) has the periodicity of the superlattice $L = a + b$. When we define

$$U(z) = \begin{cases} U_A(z) \ , & \text{in well region,} \\ U_B(z) \ , & \text{in barrier region,} \end{cases} \tag{8}$$

we obtain

$$\begin{aligned} &\left[-\frac{d^2}{dz^2} + 2ik_z\frac{d}{dz} + k_A^2 - k_z^2\right] U_A = 0 \ , \\ &\left[-\frac{d^2}{dz^2} + 2ik_z\frac{d}{dz} + k_B^2 - k_z^2\right] U_B = 0 \ , \end{aligned} \tag{9}$$

where

$$k_A^2 = \frac{2m^*\varepsilon_z}{h^2} \ , \quad k_B^2 = \frac{2m^*(\varepsilon_z - V_c)}{h^2} \ . \tag{10}$$

Equation (9) can be solved by choosing the solutions

$$\begin{aligned} U_A &= C_A^+\exp[i(k_A - k_z)z] + C_A^-\exp[i(k_A + k_z)z] \ , \\ U_B &= C_B^+\exp[i(k_B - k_z)z] + C_B^-\exp[i(k_B + k_z)z] \ . \end{aligned} \tag{11}$$

We assume boundary conditions:

$$\begin{aligned} &U_B(0) = U_A(0) \ , \quad U_B(-b) = U_A(a) \ , \\ &\frac{dU_B(0)}{dz} = \frac{dU_A(0)}{dz} \ , \quad \frac{dU_B(-b)}{dz} = \frac{dU_A(a)}{dz} \ . \end{aligned} \tag{12}$$

From the boundary conditions we obtain

$$\begin{vmatrix} 1 & 1 & 1 & 1 \\ \alpha_- & -\alpha_+ & -\beta_- & \beta_+ \\ u_- & u_+^{-1} & -v_-^{-1} & -v_+ \\ \alpha_- u_- & -\alpha_+ u_+^{-1} & -\beta_- v_-^{-1} & \beta_+ v_+ \end{vmatrix} = 0 \ , \tag{13}$$

where $u_\pm = \exp(i\alpha_\pm a)$, $v_\pm = \exp(i\beta_\pm b)$, $\alpha_\pm = k_A \pm k_z$, $\beta_\pm = k_B \pm k_z$. Solving (13), we obtain

$$\cos(k_z L) = \cos(k_A a)\cos(k_B b) - \frac{k_A^2 + k_B^2}{2k_A k_B}\sin(k_A a)\sin(k_B b) \ . \tag{14}$$

This equation is easily solved numerically and gives a dispersion relation between electron energy ε and wave vector k_z.

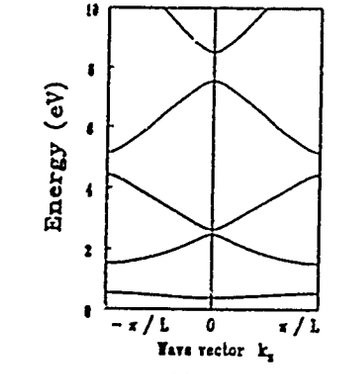
In superlattices the electron effective masses are generally different and thus the boundary conditions (12) should be modified as pointed out by Bastard (1981, 1982)

$$\frac{1}{m_B^*}\frac{dU_B(0)}{dz} = \frac{1}{m_A^*}\frac{dU_A(0)}{dz} \ , \quad \frac{1}{m_B^*}\frac{dU_B(-b)}{dz} = \frac{1}{m_A^*}\frac{dU_A(a)}{dz} \ , \tag{15}$$

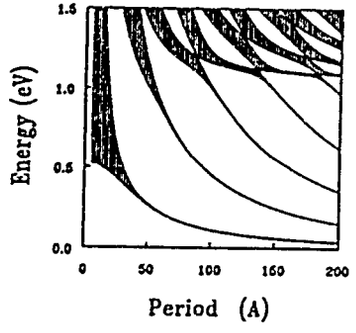
where the boundary condition takes into account the continuity of the probability current. m_A^* and m_B^* are the effective masses in region A and B, respectively. This boundary condition leads to the result of (14) with new variables

$$k_A^2 = \frac{2m_A^* e_z}{h^2} \quad , \quad k_B^2 = \frac{2m_B^*(e_z - V_c)}{h^2} \quad . \tag{16}$$

A typical example of the energy band structure of a superlattice calculated by the Kronig-Penney model is shown in Fig. 2 for the case of (GaAs)5/(AlAs)5 superlattice (L = 28Å).



(a)



(b)

Fig. 2 (a) Energy band structure calculated by the Knonig-Penney model for a = 14 Å and b = 14 Å ((GaAs)5/(AlAs)5). (b) Energy band edge as a function of superlattice period.

We have to note here that Knonig-Penney model is extremely simplified. As seen in Fig. 2, the Brillouin zone edge of the host material is $2\pi/a$ and $2\pi/b$, while the edge of the superlattice is given by $2\pi/(a+b)$, where the lattice constants of the host materials are a and b. This means that the Brillouin zone is folded in the growth direction of the superlattice, resulting in mixing of the wave functions of different bands. The energy band structures are very complicated and thus the energy bands of superlattices become very complicated. Therefore we need more sophisticated calculations in order to get details of the energy bands of superlattices, which will be made by pseudopotential method, APW and tight-binding methods. In this lecture we will concern with the empirical tight-binding method because the method is very helpful for understanding energy band structures of semiconductors and their superlattices.

Tight-Binding Method

The tight-binding method is based on linear combination of atomic orbitals (LCAO). In this paper we use Harrison's model of semiconductors by constructing a nearest neighbor tight-binding theory (Harrison, 1981) and extend the model by including the second-nearest neighbor interactions. The tight-binding theory utilizes chemical bondings of minimum number of atomic orbitals. A simple calculation is carried out by taking into account bonding of sp^3 basis, but this was found to fail in producing an indirect gap for Si (Vogl *et al.*, 1983). This disadvantage can be removed by introducing an excited s state, s^*, on each atom, in addition to sp^3 basis. The sp^3s^* model of the empirical tight-binding has been successfully used to calculate energy band structures of various semiconductors (Vogl *et al.*, 1983). We present the outline of the theory at the beginning and later show how to extend the theory to calculate energy bands of superlattices.

We start from the zincblende-structure tight-binding Hamiltonian in a basis of quasi-atomic functions

$$|nbk\rangle = \frac{1}{\sqrt{N}} \sum_{i,b} |nbR_i\rangle \exp[i(k \cdot R_i + k \cdot V_b)] \quad . \tag{17}$$

The quantum numbers n run over the s, p_x, p_y, p_z, and s^* orbitals, the N wave vectors k lie in the first Brillouin zone, and the site index *b* is either *a* (for anion) or *c* (for cation). The anion positions are R_i and the cation positions are R_i+V_b with $V_b = \delta_{c,b}(a_L/4)$ (1,1,1) and a_L being the lattice constant, in terms of Kronecker δ. The quasi-atomic functions are Löwdin orbitals, which are symmetrically orthogonalized atomic orbitals. The Schrödinger equation for the Bloch function $|k\lambda\rangle$ is

$$[\, H - \varepsilon(k,\lambda)\,]\, |k\lambda\rangle = 0 \quad , \tag{18}$$

or, in this basis,

$$\left[\, \langle nbk|\, H\, |mb'k\rangle - \varepsilon(k,\lambda)\delta_{nm}\delta_{bb'} \,\right]\, \langle mb'k|k\lambda\rangle = 0 \quad . \tag{19}$$

The solutions are

$$|k\lambda\rangle = \sum_{n,b} |nbk\rangle \langle nbk|k\lambda\rangle \quad . \tag{20}$$

The band index λ has ten values for semiconductor single crystals such as Ge, Si, GaAs, AlAs, and so on. The Hamiltonian matrix in the $|nbk\rangle$ basis is given by 10x10 matrix (Vogl *et al.*, 1983) and we will not repeat here. The tight-binding matrix elements are also listed in their paper.

In the above treatment we took into account only the nearest neighbor interactions. In addition the above calculations were done by fitting the energies at the Γ and X points, but not at the L point. In order to fit the energy gap at the L point we use the method proposed by Newman *et al.* (1984) and Yamaguchi (1987) by taking into account the second-nearest neighbor interactions. Here we will show a typical example of the energy band structure calculated by this method. Figure 3 shows the energy band structure of GaAs calculated by the tight-binding method.

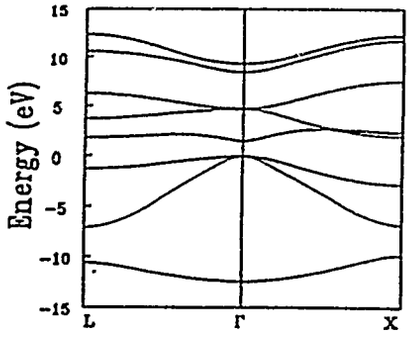
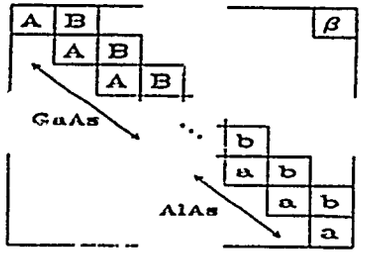


Fig. 3 Energy band structure of GaAs calculated by the tight binding method.

We extend this method to the calculations of energy band structures of superlattices. Once the bulk tight-binding parameters are determined, the Hamiltonian matrix for the superlattice is easily obtained, which is schematically shown in (21), where each block has 10x10 submatrix (Schulman *et al.*, 1979). The blocks A and a contain 10x10 matrix elements between orbitals within a layer for GaAs and AlAs, respectively, and the blocks B and b contain matrix elements between adjacent layers. The corner blocks β contain matrix elements between adjacent layers where each layer is in an adjacent slab.



(21)

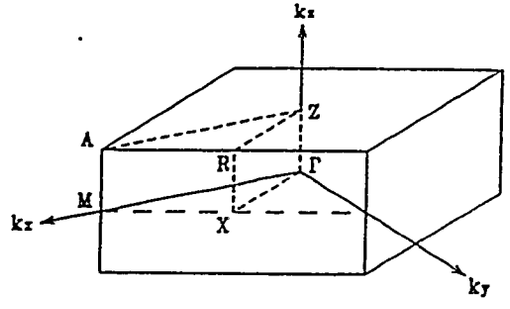
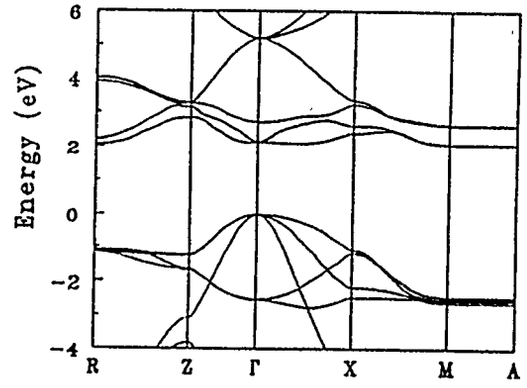
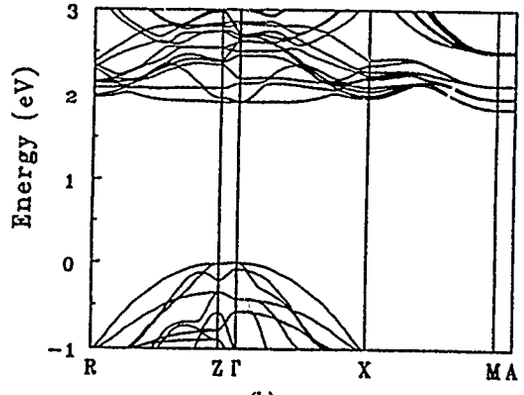


Fig. 4 Brillouin zone of $(GaAs)_n/(AlAs)_n$ and the notation.

The eigen value equation for a superlattice is the same as (19) and thus the solution is given by (20), where the band index λ has 10x(m+n) values for $(GaAs)_m/(AlAs)_n$ superlattice. Calculations of energy band structures are straight forward.



(a)



(b)

Fig. 5 Energy band structures of (a) $(GaAs)_1/(AlAs)_1$ and (b) $(GaAs)_5/(AlAs)_5$ calculated by the tight-binding method based on sp^3s^* basis.

In Fig. 4, we present the Brillouin zone of the SL with the notation of the critical points, where the growth direction of the SL is (001) (the z-direction). A typical example of the energy band structure of a SL is shown in Fig. 5(a) for $(GaAs)_1/(AlAs)_1$, where we used the parameters reported elsewhere (Nakazawa *et al.*, 1989; Fujimoto *et al.*, 1989, 1990) and the valence band discontinuity is 0.54 eV. We can see very clearly the zone-folding in the z-direction. In addition, an indirect gap appears with the lowest conduction band at the R point but the energy of the minimum is very close to the lowest conduction band edge at the Γ point. The effective mass of the lowest conduction band is heavier than that of the second lowest conduction band. This seems to be due to the fact that the lowest conduction band reflects the property of the conduction band of AlAs. The results for $(GaAs)_5/(AlAs)_5$ are shown in Fig. 5(b), where we find that the energy bands consist of many zone-folded bands and that the conduction band edges at various points lie in a narrow range of energy. These results and the accuracy of the energy band calculation indicate that the assignment of the critical points is a

very difficult task. We will discuss the energy band structures in connection with photoreflectance spectroscopy in a later section.

OPTICAL PROPERTIES

Optical Constants

Electromagnetic waves in materials can be derived from Maxwell equations, which are expressed by electric field E, electric displacement D, magnetic field H, and magnetic flux B in the following forms

$$\nabla \times E = -\frac{\partial B}{\partial t}, \qquad \nabla \cdot D = 0,$$

$$\nabla \times H = \frac{\partial D}{\partial t}, \qquad \nabla \cdot B = 0, \tag{22}$$

where

$$D = \varepsilon_0 E + P = \kappa \varepsilon_0 E \equiv \varepsilon E,$$

and

$$B = \mu_r \mu_0 H \equiv \mu H,$$

and where P is polarization, ε_0 and μ_0 are dielectric constant and permittivity in free space, κ and μ_r are relative dielectric constant and permittivity of the material. From (22), we obtain

$$\nabla^2 E = \varepsilon \mu \frac{\partial^2 E}{\partial^2 t} = \frac{1}{c^2} \frac{\partial^2 E}{\partial^2 t}, \tag{23}$$

where $c = 1/\sqrt{\varepsilon \mu}$ is the speed of light in the material and the speed of light in free space is given by $c_0 = 1/\sqrt{\varepsilon_0 \mu_0}$. When we define a refractive index $n=\kappa^{1/2}$ (asuuming $\mu_r = 1$). The solution of (23) is given by, for a plane wave with wave vector k_z and angular frequency ω,

$$E = E_0 \exp[i (k_z z - \omega t)] = E_0 \exp[-i \omega (t - \frac{n}{c_0} z)], \tag{24}$$

where E_0 stands for the amplitude of the electric field. In general the dielectric constant in materials is expressed by a complex constant $\kappa=\kappa_1+i\kappa_2$. We introduce complex refractive index n^* as

$$n^{*2} = (n_0 + i k_0)^2 = \kappa_1 + i \kappa_2, \tag{25}$$

which leads to the following relations between the dielectric constant and refractive index

$$\kappa_1 = n_0^2 - k_0^2, \qquad \kappa_2 = 2 n_0 k_0. \tag{26}$$

The electric field is then written as

$$E = E_0 \exp[-i \omega (t - \frac{n^*}{c_0} z)] = E_0 \exp[-i \omega (t - \frac{n_0 + i k_0}{c_0} z)]. \tag{27}$$

The power density of electromagnetic wave is proportional to $|E|^2$ and hence

$$|E|^2 = E_0^2 e^{-\frac{2 k_0 \omega}{c_0} z} = E_0^2 e^{-\alpha z}. \tag{28}$$

The electromagnetic radiation in the material is damped with a characteristic constant α which is called absorption coefficient. From (26) and (28), we obtain important relations

$$\alpha = \frac{2 k_0 \omega}{c_0} = \frac{\omega \kappa_2}{c_0 n_0}. \tag{29}$$

Using the refractive index defined by (25), the near-normal reflectivity $R(\omega)$ is written as

$$R(\omega) = \frac{(n_0 - 1)^2 + k_0^2}{(n_0 + 1)^2 + k_0^2}, \tag{30}$$

where we find the reflectivity is also a function of n_0, k_0 or κ_1, κ_2.

The general expression for κ_2 in the one-electron approximation for a semiconductor is easily calculated by using perturbation calculations and the imaginary part of the dielectric constant is given by

$$\kappa_2 = \frac{\pi e^2}{\varepsilon_0 m^{*2} \omega^2} \sum_{k,k'} |e \cdot p_{cv}|^2 \delta[\varepsilon_c(k) - \varepsilon_v(k') - h\omega] \delta_{kk'}, \tag{31}$$

where c and v denote conduction (empty) and valence (filled) bands, respectively, e is the unit polarization vector of the photon field with photon energy $h\omega$, and

$$e \cdot p_{cv} = \langle ck | e \cdot p | vk' \rangle, \tag{32}$$

is the momentum matrix element between the conduction band $|ck\rangle$ and the valence band $|vk'\rangle$. We can assume that the momentum matrix element is almost independent of the wave vector k, and thus (31) may be rewritten as

$$\kappa_2 = \frac{\pi e^2}{\varepsilon_0 m^{*2} \omega^2} |e \cdot p_{cv}|^2 \sum_k \delta(\varepsilon_{cv}(k) - h\omega), \tag{33}$$

where m^* is an appropriate reduced mass characterizing the joint conduction-valence band density of states, and

$$\varepsilon_{cv}(k) = \varepsilon_c(k) - \varepsilon_v(k), \tag{34}$$

is the interband energy. The summation of the delta function is just the joint density of states which is given by

$$J_{cv}(\hbar\omega) = \sum_k \delta(\varepsilon_{cv}(k) - \hbar\omega) = \frac{2}{(2\pi)^d} \int d^d k\, \delta(\varepsilon_{cv}(k) - \hbar\omega) \ , \qquad (35)$$

where d is the dimensionality of the electron-hole bands in the size-constrained system. When we assume parabolic dispersion for the electronic states, we obtain the joint density of states for three, two, and one dimensional cases as shown in Table 1.

Table 1 Joint density of states at the critical points for three, two, and one dimensions, where ε_G is the energy gap. The index j of M_j represents the number of negative reduced masses, and M_j critical points are called van Hove critical points.

critical point	$J_{CV}(\hbar\omega)$	
	$\hbar\omega \le \varepsilon_G$	$\hbar\omega \ge \varepsilon_G$
3D M_0	0	$C_1 (\hbar\omega - \varepsilon_G)^{1/2}$
3D M_1	$C_2 - C_1 (\varepsilon_G - \hbar\omega)^{1/2}$	C_2
3D M_2	C_2	$C_2 - C_1 (\hbar\omega - \varepsilon_G)^{1/2}$
3D M_3	$C_1 (\varepsilon_G - \hbar\omega)^{1/2}$	0
2D M_0	0	B_1
2D M_1	$(B_1/\pi)(B_2 - \ln\lvert\varepsilon_G - \hbar\omega\rvert)$	$(B_1/\pi)(B_2 - \ln\lvert\varepsilon_G - \hbar\omega\rvert)$
2D M_2	B_1	0
1D M_0	0	$A (\hbar\omega - \varepsilon_G)^{-1/2}$
1D M_1	$A (\varepsilon_G - \hbar\omega)^{-1/2}$	0

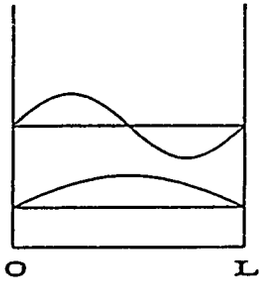


Fig. 6 Quantum well and two dimensional electronic states.

Other important relations are the Kramers-Kronig relations, which result from the requirement of causality on the response function:

$$\kappa_1 - 1 = \frac{2}{\pi} \mathcal{P} \int_0^\infty \frac{\omega' \kappa_2(\omega')}{\omega'^2 - \omega^2} d\omega' \ , \qquad (36a)$$

$$\kappa_2 = \frac{2\omega}{\pi} \mathcal{P} \int_0^\infty \frac{\kappa_1(\omega')}{\omega'^2 - \omega^2} d\omega' \ , \qquad (36b)$$

where $\mathcal{P}$ denotes the Cauchy principal part of the integral.

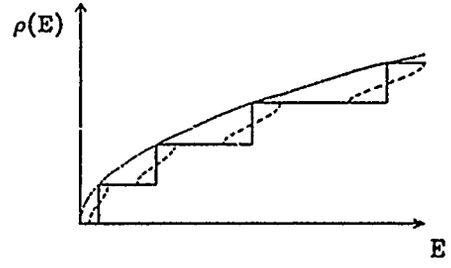


Fig. 7 Joint density of states for two- (solid curve) and three-dimensional cases (dotted curve). Joint density of states for superlattice is also shown by dashed curve for comparison.

As a comparison of the joint density of states between three- and two-dimensional critical points, we deal with the case of optical transitions in a quantum well with infinite barrier height shown in Fig. 6. The electronic states in a infinite quantum well are obtained by solving Schrödinger equation in one dimension

$$-\frac{h^2}{2m^*} \frac{d^2 F_n(z)}{dz^2} = \varepsilon_n^c F_n(z) \quad , \ F_n(0) = F_n(L) = 0 \ , \ 0 \le z \le L \ , \qquad (37)$$

and the energy is given by

$$\varepsilon^c = \varepsilon_n^c + \frac{h^2}{2m^*} (k_x^2 + k_y^2).$$

The same results are obtained for holes and thus the interband energy is given by

$$\varepsilon_{cv}(k_x, k_y) = \varepsilon_c - \varepsilon_v \frac{\pi^2 h^2}{2L^2} \left(\frac{n^2}{m_e^*} + \frac{m^2}{m_h^*} \right) + \frac{h^2 (k_x^2 + k_y^2)}{2\mu}$$

$$= \varepsilon_{nm} + \frac{h^2 (k_x^2 + k_y^2)}{2\mu} \ , \qquad (38)$$

where μ ($1/\mu = 1/m^*_e + 1/m^*_h$) is the reduced mass (we assumed isotropic masses for electron and hole). The joint density of states is therefore given by

$$\rho_{2D} = \frac{\mu}{\pi\hbar^2} \sum_{m,n} \theta(\hbar\omega - \varepsilon_{mn}) \ , \qquad (39)$$

where we find that the density of states is the step function as shown in Fig. 7. In Fig. 7 we also plotted the density of states for a superlattice, which will not discussed in detail but the results are easily understood when we take into account the minibands in conduction and valence bands.

Modulation spectroscopy

Let us consider the perturbation on the reflectivity R. We are interested in the electric field induced change in the reflectivity. When we apply an electric field to the sample, the reflectivity is changed through the change in the dielectric constants, κ_1 and κ_2. The reflectivity change is therefore written as (Cardona, 1969; Seraphin, 1972; Aspnes, 1980)

$$\frac{\Delta R}{R} = \alpha(\kappa_1,\kappa_2)\Delta\kappa_1 + \beta(\kappa_1,\kappa_2)\Delta\kappa_2 \ , \qquad (40)$$

where α and β are called the Seraphin coefficients. The electric field destroys the translational symmetry of the crystal, resulting in a complexity of the optical properties. Electrons and holes are accelerated in an electric field, and then the wavevector k is no longer a good quantum number in the direction of the field. The effect of the field is divided into two cases. One is "low field" modulation, called third derivative modulation spectroscopy, and the other is "high field" modulation, called the Franz-Keldysh effect. Since Franz-Keldysh effect is known to be very complicated because of the difficulty in achieving uniform distribution of electric field, we describe here the low field modulation (Aspnes *et al.*, 1972; Aspnes, 1973).

The third derivative functional form was derived by Aspnes *et al.* (1972) by means of a perturbation treatment. Following Aspnes *et al.* (1972) we expand the energy difference, ε_{cv} and the optical matrix elements in a Taylor series around k and assume that the momentum matrix element p_{cv} is weakly dependent on k. Then we get (Glembocki *et al.*, 1989)

$$\kappa(\varepsilon,F,\Gamma) = 1 + \frac{iA}{\varepsilon^2}\int_{BZ} d^3k \int_0^\infty dt \exp\left\{\frac{i}{\hbar}\left[(\varepsilon-\varepsilon_{cv})t - (\Omega t)^3\right] - \frac{\Gamma t}{\hbar}\right\} \ , \qquad (41)$$

where $\Omega = e^2F^2/8\mu\hbar$, $A = 4\pi^2e^2\hbar|e\cdot p_{cv}|^2/m^2$ and μ is the reduced interband mass in the direction of the field and is related to $\nabla_{ki}\varepsilon_{cv}$ (i = x,y,z). If the field is small or the broadening factor Γ large, we can expand the exponent containing Ω, to get $\exp[-i(\Omega t)^3/3\hbar] \approx 1 - i(\Omega t)^3/3\hbar$. With this approximation, we obtain Aspnes' low field limit (Aspnes *et al.*, 1972)

$$\kappa(\varepsilon,F,\Gamma) \equiv 1 + \frac{iA}{\varepsilon^2}\int_{BZ} \frac{d^3k}{\varepsilon-\varepsilon_{cv}+i\Gamma} + \frac{2iA}{\varepsilon^2}\int_{BZ} \frac{(\hbar\Omega)^3\, d^3k}{(\varepsilon-\varepsilon_{cv}+i\Gamma)^4} \ . \qquad (42)$$

In (42), the first two terms represent the unperturbed dielectric function, while the last term is the field dependent perturbation and is directly related to the well known Aspnes third derivative functional form

$$\Delta\kappa(\varepsilon,F,\Gamma) = \frac{\hbar^3\Omega^3}{3\varepsilon^2}\frac{\partial^3}{\partial\varepsilon^3}\left[\varepsilon^2\kappa(\varepsilon,\Gamma)\right] \ , \qquad (43)$$

where the unperturbed dielectric function is given by

$$\kappa(\omega,\Gamma) = -\frac{Q}{\pi\omega^2}\int \frac{d^3k}{\hbar\omega-\varepsilon_{cv}+i\Gamma} \ , \quad Q = \frac{e^2\,|e\cdot p_{cv}|^2}{\pi m^2} \ . \qquad (44)$$

The physical meaning of the low-field modulation, third-derivative functional form, is as follows (Pollak, 1990): Electrons and holes get energy in the field F, which is $\varepsilon_{gain} = e^2F^2t^2/2\mu$, where μ is the reduced mass of electron and hole. The dielectric function is written as

$$\kappa = \kappa(\varepsilon - \varepsilon_G + i\,\Gamma), \qquad (45)$$

and the electric field induced change in the dielectric function is therefore written as

$$\Delta\kappa = \kappa(\varepsilon - \varepsilon_G + \varepsilon_{gain} + i\,\Gamma) - \kappa(\varepsilon - \varepsilon_G + i\,\Gamma), \qquad (46)$$

If we assume low field case, the energy gain is small, $\varepsilon_{gain} << \Gamma$, and then a Taylor expansion gives

$$\Delta\kappa = \varepsilon_{gain}\frac{\partial}{\partial\varepsilon}\kappa(\varepsilon-\varepsilon_G+i\Gamma) \ . \qquad (47)$$

Noting that time t is an operator in quantum mechanics, $t \approx i\hbar(\partial/\partial\varepsilon)$, the field induced change in the dielectric function is written as

$$\Delta\kappa \approx (\hbar\Omega)^3\frac{\partial^3}{\partial\varepsilon^3}\kappa(\varepsilon-\varepsilon_G+i\Gamma) \ , \qquad (48)$$

which agrees with (43) qualitatively.

Combining the results shown in Table 1 and (43), we obtain the well known Aspnes' third derivative functional form for the low field modulation spectroscopy.

$$\frac{\Delta R}{R} = \Re\left[\, Ce^{i\theta}(\varepsilon-\varepsilon_G+i\Gamma)^{-m}\right] \ , \qquad (49)$$

where C and θ are the amplitude and phase factor which vary slowly with ε and hence essentially may be considered energy independent for small change in ε. The value of m depends on the type of critical points and given by

$$m = \begin{cases} \frac{5}{2} , & \text{for 3D critical point,} \\ 3 , & \text{for 2D critical point,} \\ \frac{7}{2} , & \text{for 1D critical point.} \end{cases} \quad (50)$$

In the case of two-dimensional electron system we find that the optical transition is proportional to the square of

$$I = \int F^{*}_{cn}(z) F_{vm}(z)\, dz , \quad (51)$$

where $F_{jn}(z)$ are the envelope functions of electron (j = c) and hole (j = v) for the subband index n or m. This result indicates that a selection rule exists for optical transition, $n = m$. However, we have to note that the application of electric fields results in a change in the wave functions of electron and hole, and therefore we can expect a change in dielectric function induced by the change in the wave functions. Taking into account this effect the field induce change in the dielectric function is given by (Shanabrook *et al.*, 1987)

$$\Delta\kappa = \left[\frac{\partial\kappa}{\partial\epsilon_G}\frac{\partial\epsilon_G}{\partial F} + \frac{\partial\kappa}{\partial\Gamma}\frac{\partial\Gamma}{\partial F} + \frac{\partial\kappa}{\partial I}\frac{\partial I}{\partial F} \right] \Delta F . \quad (52)$$

The expression (52) is the most general form of the modulated dielectric function for a confined system with field along the confinement direction. In addition, Glembocki *et al.* (1989) found that the electric field induced change (electric field induced change in the gap) gives rise to the first derivative of the dielectric function. In other words, the first term of (52) is proportional to the first derivative of the dielectric function.

MODULATION SPECTROSCOPY IN QUANTUM WELLS AND SUPERLATTICES

One of the most powerful methods to investigate optical properties of superlattices is photoreflectance spectroscopy. The method is contactless and provides well resolved structure of the modulated reflectance. In the photoreflectance spectroscopy, incident light (normally laser with photon energy higher than the gap), chopped with a frequency around 10 to several 100 Hz, produces electron-hole pairs in the region of surface, resulting in neutralization of the surface states and thus in a surface field change. The change in the surface field induces a reflectivity change as described in the previous section, which is monitored by a probe light dispersed by a monochromator and reflected light beam is detected by photodetector and a lock-in amplifier. An example of the experimental setup for photoreflectance measurement is shown in Fig. 8.

Figure 9 represents experimental results on photoreflectance measurements of a multiple quantum well, $Al_{0.33}Ga_{0.67}As/GaAs$, with GaAs well width 25 Å and AlGaAs barrier width 200 Å. As we can see in Fig. 9, the photoreflectance spectra are rich in structures in the photon energy region 1.4 to 2.0 eV, where we find photoreflectance signals from the GaAs substrate, subband transitions, and higher minibands transitions. In Fig. 9, dotted curves are experimental data and the solid curves are best fits to the data by using Aspnes' third derivative functional form. The doublet structure around 1.42 eV in Fig. 9 was first assigned as fundamental absorption edge of GaAs and band gap of AlGaAs with a small mole fraction of Al formed at the initial stage of buffer layer growth (Al sputtered from the shutter of Al effusion cell). In the present analysis, we obtained three transition energies labeled by E_{01}, E_{02}, and E_{0A}, which are assigned as fundamental absorption edge of GaAs substrate or buffer layer, strained layer of GaAs just above the GaAs substrate and the fundamental edge of AlGaAs (Al mole fraction is about 0.02, formed by sputtering effect), respectively. In contrast, Shen *et al.* (1986), assigned E_{02} as AlGaAs formed by sputtering. In the photon energy region of 1.6 to 1.7eV, we obtained two transition energies labeled C1-H1 at 1.643 eV and C1-L1 at 1.675 eV, which are in good agreement with the calculated ones between the subbands in the conduction and valence bands of GaAs well of 25 Å, where C1-H1 and C1-L1 correspond to the transitions from the heavy-hole and light-hole ground subbands to the lowest conduction band subband, respectively. These energies are calculated under the assumption of the conduction band offset 57% and the effective mass of electron 0.665 m, heavy-hole 0.34 m and light-hole 0.094 m (Miller *et al.*, 1984).

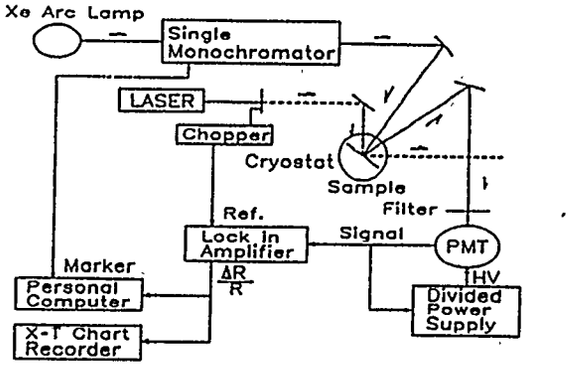


Fig. 8 Experimental setup of photoreflectance spectroscopy. The laser beam from an Ar ion laser is chopped at a frequency of 210 Hz. Signals are detected by a photomultiplier tube and amplified by a lock-in-amplifier and recorded by a personal computer. The sample temperature is controlled by a cryostat.

In the photon energy region from 1.8 to 1.9 eV, we find complicated structures. However, the best fit curves give five transition energies, which are assigned as follows. The transition at 1.792 eV labeled $E_0+\Delta_0$ in Fig. 9 arises from spin-orbit splitting of the GaAs valence band. The other four transition energies are ascribed to the transitions between minibands in the valence and conduction bands. This assignment is supported by calculations of Kronig-Penney model which are shown by C_m-H_n and C_m-L_n with m and n greater than unity, where C_m, H_n and L_n stand for the miniband edges in the conduction, heavy-hole and light-hole bands, respectively. As stated, photoreflectance spectra due to transition between the confined electronic states should be expressed by the first derivative functional form (Glembocki *et al.*, 1989) and therefore the above analysis should be modified. However, it is well known that the transition energy is affected little by the assumption of the functional form, third or first derivative. Therefore we did not re-analyze the experimental data by using the first derivative functional form. Recently, modulated reflectance spectra are found to be well analyzed by using Gaussian form for the dielectric function (Shen *et al.*, 1987; Garland *et al.*, 1988). We will not discuss the functional form in detail here. In the short period superlattices which will be discussed below, electrons and holes are extended in the SL and thus their wave functions are three-dimensional-like. Therefore, the experimental data of PR in SLs are analyzed by using (49) with $m = 2.5$ (three-dimensional critical point).

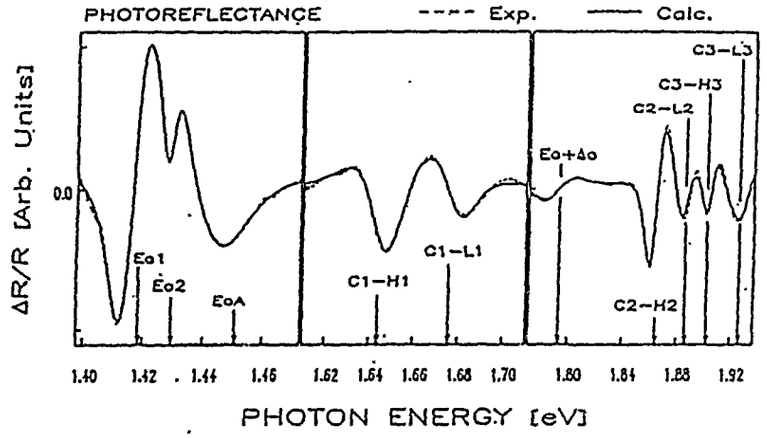


Fig. 9 Photoreflectance spectra of (AlGaAs/GaAs) multiple quantum well with barrier/well widths of 200/25 Å at room temperature. The dashed curves are experimental spectra and the solid curves are best-fitted to the experimental spectra by using the third derivative functional formula of modulated reflectance. The transition energies determined from the best-fitting are shown by the arrow.

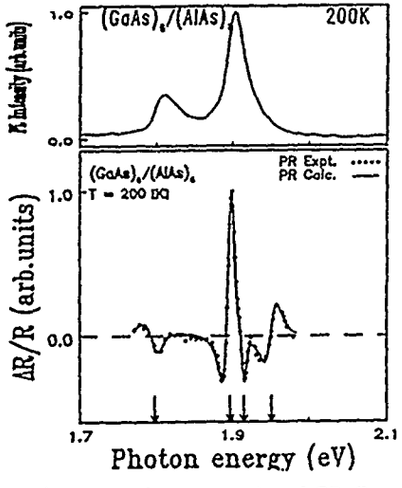


Fig. 10 Photoluminescence (upper trace) and PR (lower trace) spectra observed in $(GaAs)_8/(AlAs)_8$ at 200 K. The solid curve in the lower trace is calculated from Aspnes' third derivative formula so as to best fit to the experimental reflectance curve and the vertical arrows indicate the transition energies estimated from the best fitting.

In SLs, the Brillouin zone is folded and mini-bands are formed, resulting in many critical points in a narrow region of photon energy. This means that the PR spectra consist of a combination of many critical points. The best fitting procedure was carried out by adopting the method described elsewhere. (Nakazawa *et al.*, 1989, 1990)

Figure 10 shows a typical result of photoreflectance (PR) and photoluminescence (PL) spectra in $(GaAs)_8/(AlAs)_8$ SL at 200 K, where the solid circles in the lower trace represent experimental data and the solid curve is a best fit to the data using Aspnes' third derivative formula. The experimental data of PR are plotted with a suitable interval so that a comparison between the experimental and best fit curves becomes clear. In Fig. 10 we find that the PL consists of two main peaks, while the PR spectrum exhibits a very complicated structure due to the existence of many critical points in the energy range 1.75 to 2.0 eV. The PL peaks in the photon energy region 1.8 to 2.0 eV seem to correspond to the PR structure in this region. We find very weak emission at lower energy in PL spectra. In SLs with smaller n, the emission on the low energy side, where we have no structure in PR, becomes stronger and exceeds the emission on the higher energy side. We do not know the origin of the low energy side emission. The vertical arrows in Fig. 10 indicate the critical point energies obtained from the best fit. It is very important to point out that the PL peak at 1.815 eV coincides with the critical point energy 1.797 eV of the weak structure in the PR spectra. Taking into account the fact that the PR spectra arise from the critical points of the joint density of states and thus from the direct transition process, the SL of $(GaAs)_8/(AlAs)_8$ has a direct gap at E_g = 1.797 eV at 200 K, but this transition is very weak compared to the transition at 1.897, 1.915 and 1.951 eV. It is very interesting to recall that the spectra of PR have fine structures and the analysis of the best fit provides accurate determination of the critical point energies. Similar features are observed in other SLs investigated in the present work.

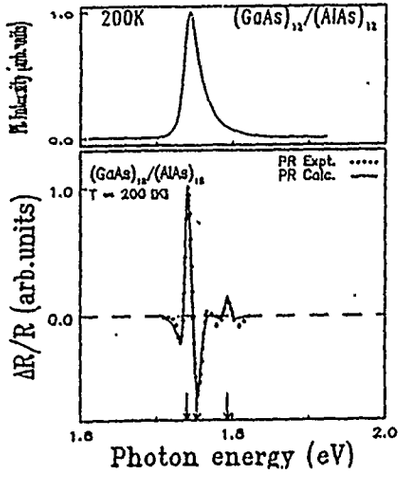


Fig. 11 Same as Fig. 10 but for $(GaAs)_{12}/(AlAs)_{12}$ at 200 K. The PR spectra consists of three structures.

These features have been observed in our previous work at room temperature, where a weak structure was observed only in $(GaAs)_5/(AlAs)_5$ at lower energy side.(Fujimoto, 1990) Two possibilities are proposed for the lower energy peak; (1) direct transition at the Γ point associated with the zone-folded conduction band and (2) indirect transition between the lowest

conduction band located at X or at a point in the Brillouin zone other than the Γ point and the valence band at the Γ point. As stated before, the PR spectra arise not from a indirect transition but from a direct transition because the indirect transition is higher order perturbation compared to the direct transition. As we can see in Fig. 10, the weak structure in the PR spectra gives the critical point energy corresponding to the PL peak, the lower energy structure may be assigned to the weakly allowed direct transition between the zone folded conduction band and the valence band. In our previous paper (Fujimoto *et al.*, 1990), we assigned the lower energy peak of the PL peaks not to the direct transition but to the indirect transition. This is because the weak structure at lower energy region in the PR spectra has been observed only in the SL $(GaAs)_5/(AlAs)_5$. In the present work, however, we carried out similar experiments very carefully by changing the temperature and improving the sensitivity of the detection, and found the weak structures in almost all the samples we investigated.

Figure 11 presents PL and PR spectra for $(GaAs)_{12}/(AlAs)_{12}$ at 200 K, where we find that only one main emission is observed in the PL spectrum and the weak structure of PR in the lower energy region merges into the main structure, where we find three transition energies in a narrow photon energy region from 1.7 to 1.85 eV, whereas the PL signals consist of a main peak and a weak shoulder at higher energy side. The present results are summarized in Fig. 12, where we plott transition energies determined from the PR spectra by solid and open circles, and the PL peak energies by the crosses. In Fig. 12, the transition energy for the weak structure in the PR signals is shown by the solid circles. Although the weak structures at lower energy side are not observed in all the samples investigated, it is very interesting to point out that the lowest PL peak agrees well with the transition energy of the weak structure in the PR at the low energy side. Another interesting feature is that the higher energy peak of the PL agrees well with the transition energy of the strongest structure of the PR. Calculated results are also shown in Fig. 12 which will be discussed later.

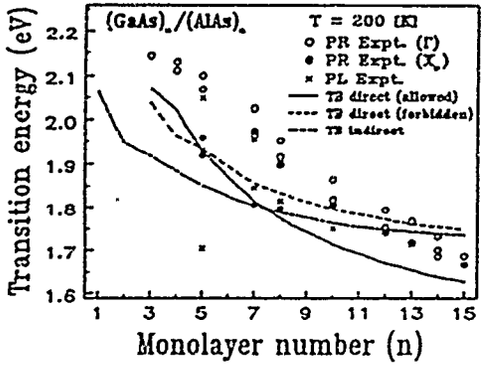


Fig. 12 Transition energies obtained from the PR and PL experiments at 200 K are plotted as a function of the atomic layer number n, along with the calculated curves. The strong (main) structures in the PR spectra are shown by open circles(referred as PR Expt. (Γ)), the weak structure by solid circles (referred as PR Expt. (X_Z)) which is ascribed to the zone-folded weak transition, and PL peaks by crosses (referred as PL Expt.). The calculated results from the tight-binding method are shown by the solid curve for the direct allowed transition, by the dashed curve for the weakly allowed transition which arises from the zone-folding effect (referred as TB direct (forbidden)), and the lowest indirect transition (referred as TB indirect).

Up to now many papers have been published on the energy band calculations of short period SLs by using various methods (Schulman *et al.*, 1979; Harrison, 1981, Nakayama *et al.*, 1985; Drummond *et al.*, 1987; Yamaguchi, 1987, Jian-Bai-Xia, 1988; Wei *et al.*, 1988; Eppenga *et al.*, 1988; Gopalan *et al.*, 1989). Most of these show that the lowest conduction band of $(GaAs)_n/(AlAs)_n$ with small n forms an indirect band. Our calculations based on the empirical tight-binding method reveals that the SLs are indirect for $n < 8$ (Nakazawa *et al.*, 1989).

We have to note here that the absolute values of the calculated critical point energies of the SLs depend strongly on the parameters assumed for the calculations, such as band parameters of GaAs and AlAs, and also the conduction band or valence band discontinuity. The accuracy of the energy band calculation is in a range of 0.1 to 1 eV (accuracy is worse at higher energy critical points), and thus a detailed comparison between the experimental and calculated results is impossible. In our previous paper we concluded that the lowest energy peak of the PL spectra arises from the indirect transition for $(GaAs)_n/(AlAs)_n$ with n less than 10, although we observed a weak structure in PR spectra in $(GaAs)_5/(AlAs)_5$ which corresponds to the lowest energy peak of the PL. If this weak structure is ascribed to the zone-folded direct transition, the conclusion drawn in the previous paper is subject to a change. As stated earlier, the main subject of the present work is to clarify whether the low energy peak of PL and the weak structure of the PR arise from the zone-folded direct transition or from the indirect transition. When the weak structure is observed in every SL, it may be ascribed to the zone-folded direct transition because only the direct transition gives rise to the PR signals. In the present work we observed the weak structure in almost all superlatticies with $4 \leq n \leq 13$.

In order to discuss the magnitude of the optical transition probability in SLs we carried out energy band calculations using the tight-binding method, (Nakazawa *et al.*, 1989; Yamaguchi, 1987) and calculation of momentum matrix elements (Fujimoto *et al.*, 1989, 1990; Drummond *et al.*, 1987). Our method is based on the empirical-tight-binding theory, which takes into account the second nearest neighbor interactions in addition to the nearest neighbor interactions of sp^3s^* orbitals as done by Yamaguchi (1987). In the present work, we use the same computer program prepared by our group and the same parameters (Nakazawa *et al.*, 1989; Fujimoto *et al.*, 1989).

The momentum matrix element between the valence and conduction bands is defined as, using the wave functions given previously,

$$M_X = \sum_{b,n,n'} \langle n'bk| C^*_{bn'} p_X C_{bn} |nbk\rangle \quad , \tag{53}$$

where p_X is the x-component of momentum operator, and the coefficients C_{bn} and $C_{bn'}$ are those defined by the right hand side of (21) and given by $\langle nbk|k\lambda\rangle$ and $\langle n'bk|k\lambda\rangle$, respectively. Calculation of the momentum matrix element is straightforward. It is evident from the symmetry consideration of the SLs that the momentum matrix elements M_X and M_Y are identical, and therefore we calculated M_X only.

The results are shown in Fig. 13, where squared matrix elements between the top valence band and lowest three conduction bands at the Γ point are plotted as a function of the atomic layer number n. The solid, dashed and dot-dashed curves, respectively, represent the squared matrix elements for the lowest, second-lowest, and third-lowest conduction bands. As seen in Fig. 12, the transition to the lowest conduction band is allowed for $n > 4$, whereas the transition to the second lowest conduction band is allowed for $n < 4$. In addition, the matrix elements oscillate in the region of small number of n. These results clearly indicate an importance of the zone-folding effect, because the zone-folding effect depends on the layer number, even or odd. As the layer number is decreased, the lowest transition 1-1 becomes very weak while the second lower transition 1-2 becomes strong. This feature is very similar

to the experimental observation, in which we found that the weak structure in the PR signals appear only in the region of small n .

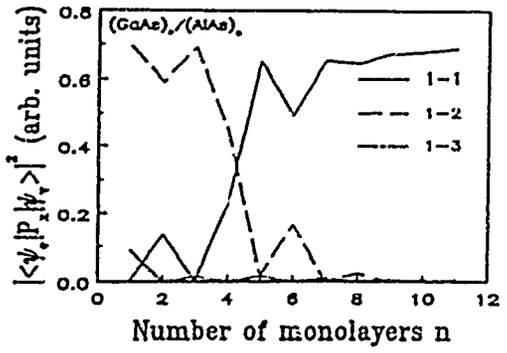


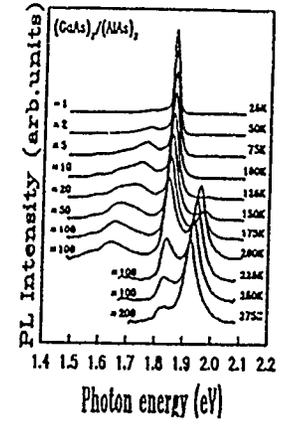
Fig. 13 Squared momentum matrix elements as a function of atomic layer number for the transition between the top valence band and the lowest, second lowest and third lowest conduction band are shown by the solid (1 - 1), dashed (1 - 2), and dot-dashed curve (1 - 3), respectively.

The calculated transition energies are shown in Fig. 12 along with the experimental result, where the energy of the direct allowed transition is plotted by the solid curve, the energy of the weak transition by the dashed curve, and the lowest indirect gap by the dot-dashed curve. Although the agreement between the present calculations and experiments is not good, the general feature is consistent. The crossover between the strong direct transition and weak direct transition is expected to appear at about $n = 5$, while the present experiment reveals that the crossover occurs at about $n = 10$.

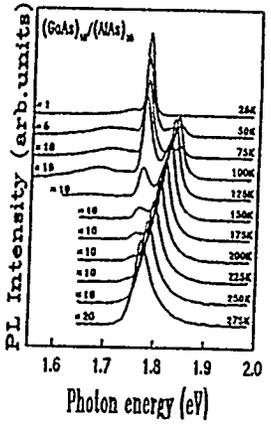
The temperature dependence of the PL spectra for $(GaAs)_7/(AlAs)_7$ is shown in Fig. 14(a). Based on the discussion stated above, the higher- and lower-energy peaks are assigned to be the direct and pseudodirect, respectively. Figure 14(b) shows the temperature dependence of the PL spectra for $(GaAs)_{10}/(AlAs)_{10}$, in which the assignment of the two peaks is the same as those in Fig. 14(a), direct and pseudodirect (weakly allowed direct) transitions. In Figs. 14(a) and (b), the intensity of the PL peak at higher energy side decreases and the other peak at lower energy side increases as the temperature decreases. It is very important to point out that the temperature at which the intensity of the two peaks becomes equal is higher for $(GaAs)_n/(AlAs)_n$ with smaller n. This feature indicates that the separation of the energy gaps between the pseudodirect and direct band gaps increases with decreasing the monolayer number n.

Intensity of emission depends on the transition probability and number of carriers excited. In the following, we assume the emission is governed by the two factors and neglect the effect of absorption. With this approximation, we are able to estimate the transition probability and the energy separation from the data shown in Figs. 14(a) and (b). Under this assumption the PL intensity is proportional to the product of electron density at the conduction band and the transition probability, and the ratio of the PL emission between the direct gap (E_Γ and the pseudodirect band gap (E_X) is written as

$$\frac{I(\Gamma)}{I(X)} = \frac{W_\Gamma}{W_X} \exp\left(\frac{E_X - E_\Gamma}{k_B T}\right), \qquad (54)$$



(a)



(b)

Fig. 14 PL spectra of (a) $(GaAs)_7/(AlAs)_7$ and (b) $(GaAs)_{10}/(AlAs)_{10}$ are plotted at various temperatures (from 25 K to 275 K) for comparison in the photon energy region where emissions of the pseudodirect and direct transitions appear.

where $I(\Gamma)$, E_Γ and W_Γ are PL intensity, transition energy and transition probability for the direct allowed transition (wave function of the conduction band has predominantly the character of the Γ band of GaAs), and $I(X)$. E_X and W_X for the pseudodirect transition (wave function of the conduction band has predominantly the character of the X band of AlAs, and referred to as X_Z-like state in this lecture), k_B is the Boltzmann constant. In (54) the factor of the effective mass or the density of states is ignored, but we have to remember that the effective mass of the X_Z-like band is very heavy as seen in the energy band calculations (Fujimoto *et al.*, 1990).

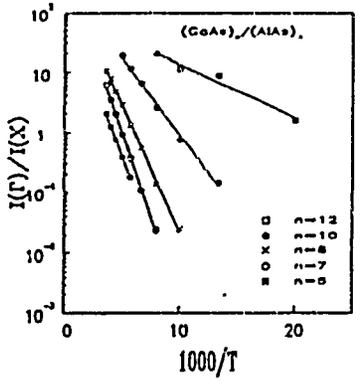


Fig. 15 The ratio of the PL peak intensity of the direct transition to the pseudodirect transition is plotted as a function of inverse temperature for the $(GaAs)_n/(AlAs)_n$ SLs with n = 5, 7, 8, 10, and 12, where the slope gives the energy separation and the intercept of the curve at zero gives the ration of the transition probabilities (see text in detail).

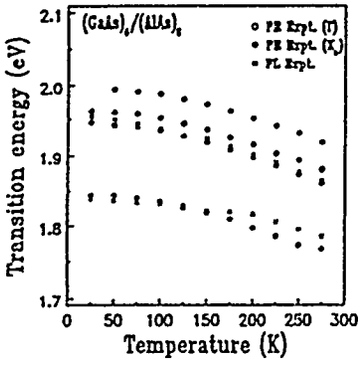


Fig. 16 Temperature dependence of the transition energies obtained by PR and PL measurements in the $(GaAs)_8/(AlAs)_8$ SL, where the open (higher transitionenergies, allowed direct) and solid circles (lower transition energies, pseudodirect) are the energies determined from the PR and the crosses are PL peak energies.

Figure 15 shows the intensity ratio I(Γ)/I(X) as a function of inverse temperature obtained from the present experiments shown in Figs. 14(a) and (b) together with the least square fit lines. These lines for $(GaAs)_n/(AlAs)_n$ with n = 5, 7, 8, 10, and 12, give the energy separation $E_\Gamma - E_X$ = 100 meV, 109 meV, 83 meV, 50 meV, and 18 meV, respectively. From the intercept of the line the ratio of transition probabilities W_X/W_Γ is estimated to be 7.7×10^{-3}, 1.8×10^{-3}, 2.9×10^{-3}, 3.1×10^{-3}, and 6.9×10^{-3} for n = 5, 7, 8, 10, and 12, respectively. The energy separation is in reasonable agreement with the transition energy obtained earlier (within about 10% difference except $(GaAs)_{12}/(AlAs)_{12}$). In addition the ratio of the transition probability suggests that the transition at the lower energy side is much weaker (by two or three orders of magnitude, or more if we take into account the difference in the effective mass). Therefore, these results strongly suggest that the lower-energy band gap is ascribed to a pseudodirect band gap at the Γ-point, resulting from the zone-folding and thus electronic state is associated with the X_Z-like state. The ratio of the transition probability obtained in the present work is consistent with the ratio of photoluminescence decay time between the two emission bands (Finkmann *et al.*, 1986; Minami *et al.*, 1987; Feldmann *et al.*, 1987; Kato *et al.*, 1989).

Figure 16 shows the temperature dependence of the transition energies in $(GaAs)_8/(AlAs)_8$ SL, where transition energies determined from the PR and PL are plotted by the circles (open and solid circles) and crosses, respectively. We find that the three transition energies obtained from the PR analysis decrease monotonically with increasing temperature and the shift is about 100 meV in the temperature range from 25 to 300 K, while the PL peaks at the lower energy side, ascribed to the pseudodirect transition, exhibits about 70meV shift in the same temperature range. The energy shift of the transition energies determined from the PR experiments is quite similar to that of the fundamental energy gap of bulk GaAs. The transition energy of the PR main structure agrees well with the PL peak at the higher energy side in the temperature range. On the other hand, there exists a slight difference in the temperature dependence between the lowest transition energy determined from PR analysis and the low-energy side peak of PL, and the shift of the PL peak energy is smaller. The difference is higher at higher temperatures and about 20 meV at 250 K. We assign this transition as the pseudodirect transition due to the zone-folding effect and thus the transition energies are expected to agree with each other within the range of phonon energy. Crossing of the transition energies at about 120 K is not understood at the present stage.

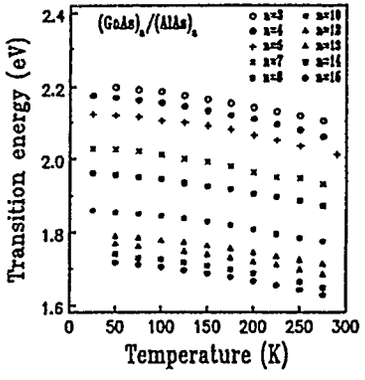


Fig. 17 Temperature dependence of the lowest direct allowed transition energy obtained by PR measurement are plotted as a function of temperature in the $(GaAs)_n/(AlAs)_n$ SLs with n = 3, 4, 5, 7, 8, 10, 12, 13 14, and 15.

Similar behavior is observed in other $(GaAs)_n/(AlAs)_n$. In Fig. 17 we plot only the transition energy corresponding to the main structure of the PR signals as a function of temperature for $(GaAs)_n/(AlAs)_n$ with n = 3 to 15. All the samples exhibit a quite similar feature in their temperature dependence. Taking into account the fact that the conduction band associated with this transition reflects the nature of the conduction band of GaAs at the Γ point, its temperature dependence is expected to be similar to that of GaAs.

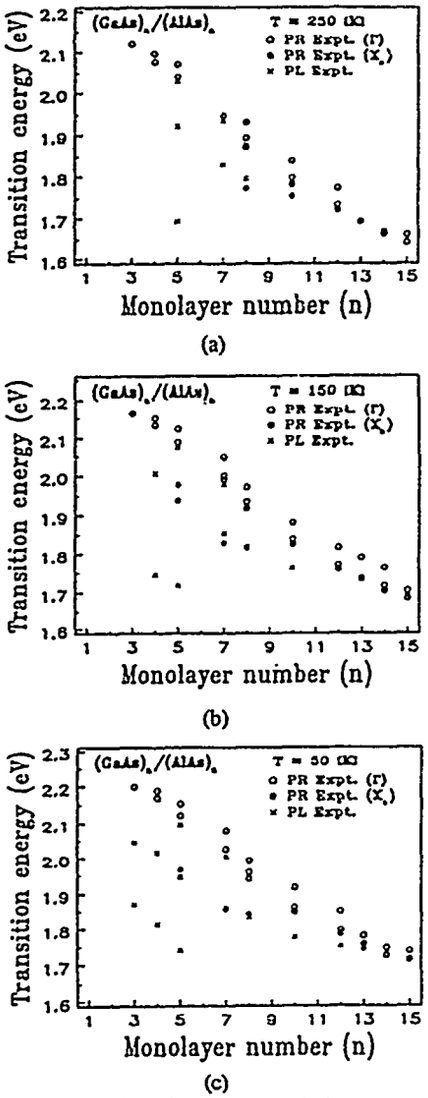


Fig. 18 Monolayer number dependence of the transition energies of $(GaAs)_n/(AlAs)_n$ SLs obtained from the present PR and PL experiments (a) at 250 K, (b) at 150 K, and (c) at 50 K, where the open and solid circles represent the allowed direct band gaps and pseudodirect band gap obtained from the PR experiments, respectively, and the crosses indicate the PL peak photon energies.

In Figs. 18(a), (b) and (c), we plot the monolayer number (n) dependence of the transition energies determined from the PR and PL measurements at 250 K, 150 K and 50 K, respectively, in $(GaAs)_n/(AlAs)_n$ with n=1-15, where the open (notation PR Expt.(X_Z)) and solid circles (notation PR Expt.(Γ)) represent transition energies determined from the PR analysis, and the crosses are the PL peak energies. The notation (Γ) and (X_Z) is used for the reasons stated earlier, where the higher transition energies determined from the PR is understood to be due to the allowed direct transition at the Γ point and the lower weak structure of the PR arises from the pseudodirect transition (the zone-folded conduction band with the wave function of the X state of AlAs). The pseudodirect transition appears at 50 K in $(GaAs)_n/(AlAs)_n$ with $n < 14$, and at 150 and 250 K in $(GaAs)_n/(AlAs)_n$ with $n < 11$. This suggests that the direct-pseudodirect crossover depends on temperature and at lower temperatures the crossover occurs in $(GaAs)_n/(AlAs)_n$ with larger n. This temperature dependence of the crossover will explain the difference between our previous work (Fujimoto *et al.*, 1990) and the work by Kato *et al.* (1989). One of the most probable possibilities for this temperature dependence may be the difference in the temperature dependence of the energy gaps of GaAs and AlAs, where the temperature dependence of the gap in AlAs is slightly smaller than that of GaAs. Noting that the pseudodirect gap reflects the nature of the conduction band at the X-point of AlAs, its temperature dependence is expected weaker compared to that of the allowed direct band gap. In Fig. 16, we find that the pseudodirect gap obtained from the PR experiments behaves quite similarly to that of the direct gap of GaAs, although the gap determined from the PL data behaves similarly to that of indirect band gap of AlAs. However, some of the $(GaAs)_n/(AlAs)_n$, with small n, exhibit difference in the temperature dependence between the pseudodirect and allowed direct gaps. It should be noted that the weak structure in PR spectra depends strongly on the temperature.

In Fig.18, we find PL peak energies, in the photon energy region around 1.7 to 1.75 eV, which are well below the pseudodirect gap. The energy is higher than the band gap of GaAs and thus the PL signals are not ascribed to the emission from the substrate. If this emission band arises from the superlattice layers, the signals are assigned to the transition from the conduction band associated with the X_{XY}-like state to the top valence band at the Γ-point, because the X_{XY}-like state is expected to be lower than the X_Z-like state for small n. This band alignment is shown by the tight-binding calculation.(Ihm, 1987)

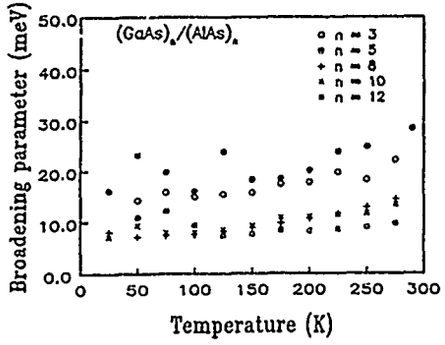


Fig. 19 Temperature dependence of the broadening parameter of PR spectra for $(GaAs)_n/(AlAs)_n$ SLs with n = 3, 5, 8, 10, and 12, where the broadening parameter for the lowest direct gap is plotted.

The analysis of PR spectra using the third-derivative formula (Aspnes, 1972, 1973) gives the broadening parameter for each critical point. Temperature dependence of the broadening parameter of the main PR signals (lower critical point for the allowed direct transition) for $(GaAs)_n/(AlAs)_n$ with n = 3, 5, 8, 10, and 12 are shown in Fig. 19. These broadening parameters are almost independent of the temperature but depend on the monolayer number n. The broadening parameters decrease with increasing n, except $(GaAs)_5/(AlAs)_5$. Since the

broadening parameters are almost independent of the temperature, the broadening arises not from the electron-phonon interactions but from other effects. We pointed out in our previous paper (Fujimoto,1990) that the broadening is induced by the fluctuation of monolayers during the growth of the SLs. Although the monolayer number was carefully controlled by observing RHEED oscillations, it is very difficult to obtain uniform periodicity of the monolayer number in the whole region of the samples used in the PR measurements. This layer number fluctuation also results in a considerable change in the energy gaps as shown in our paper (Hamaguchi, 1990), where we calculated energy band structures of a long period superlattice such as $(GaAs)_n/(AlAs)_n/(GaAs)_{n+1}/(AlAs)_n$ and the lowest direct gap is about 100 meV lower than $(GaAs)_n/(AlAs)_n$ for $n = 4$.

REFERENCES

Ando, T., Fowler, A. B., and Stern, F., 1982, Electronic Properties of two-dimensional systems, Rev. Mod. Phys. 54: 437.

Aspnes, D. A., 1973, Third-derivative modulation spectroscopy with low-field electroreflectance, Surface Sci. 37: 418.

Aspnes, D. A., 1980, Modulation spectroscopy/electric field effects on the dielectric function of semiconductors, in "Handbook on Semiconductors", Vol. 2, ed. by T. S. Moss (North Holland, N. Y.), p. 109 and references therein.

Aspnes, D. A., and Rowe, J. E., 1972, Resonant Nonlinear Optical Susceptibility: Electroreflectance in the low-field limit, Phys. Rev. B, 5: 4022.

Bastard, G., 1981, Superlattice band structure in the envelope-function approximation, Phys. Rev. B, 24: 5693.

Bastard, G., 1982, Theoretical investigations of superlattice band structure in the envelope-function approximation, Phys. Rev. B, 25: 7584.

Cardona, M., 1969, "Modulation Spectroscopy", in "Solid State Physics", Suppliment 11, Ed. by Seitz, F., Turnbull, D., and Ehrenreich, H. (Academic Press, New York), p.1.

Drummond, T. J., Jones, E. D., Hjalmarson, H. P., and Doyle, B. L., 1987, $GaAs/In_xAl_{1-x}As$ ($0 \le x \le 0.006$) Indirect Bandgap Superlattices, in "Proc. Int. Symp. GaAs and Related Compounds, Las Vegas, Nevada, 1986", Inst. Phys. Conf. Ser. No.83, Chapter 6, pp.331-336.,

Esaki, L., and Tsu, R., 1970, Superlattice and negative differential conductivity in semiconductors, IBM J. Res. Dev., 14: 61.

Eppenga, R., and Schuurmans, M. F. H., 1988, Thin [001] and [$1\bar{1}0$] GaAs/AlAs Superlattices: Distinction Between Direct and Indirect Semiconductors, Phys. Rev. B, 38: 3541.

Feldmann, J., Peter, G., Gobel, E. O., Dawson, P., Moore, K., Foxon, C., and Elliott, R. J., 1987, Linewidth Dependence of Radiative Exciton Lifetimes in Quantum Wells, Phys. Rev. Lett., 59: 2337.

Finkman, E., Sturge, M. D., and Tamargo, M. C., 1986, X-point Excitons in AlAs/GaAs Superlattices, Appl. Phys. Lett., 49: 1299.

Fujimoto, H., Hamaguchi, C., Nakazawa, T., Taniguchi, K., and Imanishi, K., 1989, Crossover of Direct and Indirect Transitions in $(GaAs)_m/(AlAs)_5$ Superlattices (m = 1 - 11), J. Phys. Soc. Jpn., 58: 3727.

Fujimoto, H., Hamaguchi, C., Nakazawa, Taniguchi, K., Imanishi, K., Kato, H., and Watanabe, Y., 1990, Direct and Indirect Transition in $(GaAs)_n/(AlAs)_n$ Superlattices with n = 1 - 15, Phys. Rev. B, 41 (to be published).

Garland, J. W., Abad, H., Viccaro, M., and Raccah, M., 1988, Line Shape of the optical dielectric function, Appl. Phys. Lett., 52: 1176.

Glembocki, O. J., and Shanabrook, B. V., 1989, Electromodulation Spectroscopy of Confined Systems, Superlatt. Microstruct., 5: 603.

Gopalan,S., Christensen, N. E., and Cardona, M., 1989, Band-Edge States in Short-Period $(GaAs)_m/(AlAs)_n$ Superlattices, Phys. Rev. B, 39: 5165.

Hamaguchi, C., Nakazawa, T., Matsuoka, T., Ohya, T., Taniguchi, K., Fujimoto, H., Imanishi, K., Kato, H., and Watanabe, Y., 1990, Direct and Indirect Transition in $(GaAs)_n/(AlAs)_n$ Superlattices with n = 1 - 15, "SPIE Int. Conf. on Modulation Spectroscopy" (in press).

Harrison, W. A., 1981, Total Energies in the Tight-Binding Theory, Phys. Rev. B, 23: 5245.

Ihm, J., 1987, Effects on the Layer Thickness on the Electronic Character in GaAs-AlAs Superlattices, Appl. Phys. Lett., 50: 1068.

Jian-Bai-Xia, 1988, Theoretical Analysis of Electronic Structures of Short-Period Superlattices $(GaAs)_m/(AlAs)_n$ and Corresponding Alloys $Al_n/(m+n)Ga_m/(m+n)As$, Phys. Rev. B, 38: 8358.

Kato, H., Okada, Y., Nakayama, M., and Watanabe, Y., 1989, Γ-X Crossover in GaAs/AlAs Superlattices, Solid State Commun., 70: 535.

Miller, R. C., Kleinman, D. A., and Gossard, A. C., 1984, Energy-gap discontinuities and effective masses for $GaAs-Al_xGa_{1-x}As$ quantum wells, Phys. Rev. B, 29: 7085.

Minami, F., Hirata, H., Era, K., Yao, T., and Masumoto, Y., 1987, Localized Indirect Excitons in a Short-period GaAs/AlAs Superlattices, Phys. Rev. B, 36: 2875.

Nakayama, T., and Kamimura, H., 1985, Band Structure of Semiconductor Superlattices with Ultrathin Layers $(GaAs)_n/(AlAs)_n$ with n = 1,2,3,4, J. Phys. Soc. Jpn., 54: 4726.

Nakazawa, T., Fujimoto, H., Imanishi, K., Taniguchi, K., Hamaguchi, C., Hiyamizu, S., and Sasa, S., 1989, Photoreflectance and Photoluminescence Study of $(GaAs)_m/(AlAs)_5$ (m = 3 - 11) Superlattices: Direct and Indirect Transition, J. Phys. Soc. Jpn., 58: 2192.

Nakazawa, T., Matsuoka, T., Ohya, T., Taniguchi, K. Hamaguchi, C., Kato, H., and Watanabe, Y., 1990, Temperature Dependence of the Energy Gaps of $(GaAs)_n/(AlAs)_n$ Superlattices, "SPIE Int. Conf. on Modulation Spectroscopy" (in press).

Newman, K. E., and Dow, D. J., 1984, Theory of deep impurities in silicon-germanium alloys, Phys. Rev. B, 30: 1929.

Pollak, F. H., 1990, Modulation spectroscopy characterization of semiconductors and semiconductor microstructures, Short Course Notes, "SPIE's 1990 Symposium on Advances in Semiconductors and Superconductors: Physics Toward Device Applications" (in press).

Schulman, J. N., and McGill, T. C., 1979, Electronic Properties of the AlAs-GaAs (001) Interface and Superlattice, Phys. Rev. B, 19: 6341. See also references therein.

Seraphin, B. O., 1972, Electroreflectance, "Semiconductors and Semimetals", Vol. 9, ed. by R. K. Willardson and A. C. Beer (Academic Press, New York), pp. 1-149.

Shanabrook, B. V., Glembocki, O. J., and Beard, W. T., 1987, Photorefrectance modulation mechanisms in $GaAs-Al_xGa_{1-x}As$ multiple quantum wells, Phys. Rev. B, 35: 2540.

Shen, H., Parayanthal, P., Pollak, F. H., Tomkiweicz, M., Drummond, T. J., and Schulman, J. N., 1986, Photoreflectance study of GaAs/AlAs superlattices: Fit to electromodulation theory, Appl. Phys. Lett., 48: 653.

Shen, H., Pan, S. H., Pollak, Fred H., Dutta, M., and AuCoin, T. R., 1987, Conclusive evidence for miniband dispersion in the photorefrectance of a $GaAs/Ga_{0.74}Al_{0.26}As$ coupled multiple-quantum-well structure, Phys. Rev. B, 36: 9384.

Vogl, P., Hjarmarson H. P., and Dow, J. D., 1983, A semi-empirical tight-binding theory of the electronic structure of semiconductors, J. Phys. Chem. Solids, 44: 365.

Wei, S.-H., and Zunger, A., 1988, Electronic Structure of Ultrathin $(GaAs)_n/(AlAs)_n$ [001] Superlattices and the $Ga_{0.5}As_{0.5}$ Alloy, J. Appl. Phys., 63: 5794.

Yamaguchi, E., 1987, Theory of the DX Centers in III-V Semiconductors and (001) Superlattices, J. Phys. Soc. Jpn., 56: 2835.

21

MICROWAVE STUDIES OF QUASI-ONE DIMENSIONAL WIRES

F.Kuchar[1], J.Lutz[1], K.Y.Lim[1], R.Meisels[1], G.Weimann[2], W.Schlapp[3], A.Forchel[4], A.Menschig[4], D.GrÄtzmacher[5], P.Beton[6], S.P.Beaumont[7], and C.D.W.Wilkinson[7]

[1]Institut für Festkörperphysik, Universität Wien
A-1090 Vienna, Austria, and
Ludwig Boltzmann Institut für Festkörperphysik
Kopernikusg.15, A-1060 Vienna, Austria

[2]Walter Schottky Institut, München, West Germany

[3]Forschungsinstitut der Deutschen Bundespost, Darmstadt, West Germany

[4]4.Physikalisches Institut, Universität Stuttgart, Stuttgart, West Germany

[5]Institut für Halbleitertechnik, RWTH, Aachen, West Germany

[6]Department of Physics, University of Nottingham, United Kingdom

[7]Nanoelectronics Research Centre, University of Glasgow, United Kingdom

INTRODUCTION

In the past, the electronic transport in quantum wires has mostly been studied using d.c. or low-frequency measuring techniques (Washburn and Webb, 1986; Heinrich *et al.*, 1988). The use of high-frequency electric fields, however, is of particular interest if the product $\omega\tau_\phi \sim 1$, where $1/\tau_\phi$ is the phase-breaking rate which determines the phase coherence of the wave function (Altshuler *et al.*, 1981; Khmelnitskii, 1988). In quantum wires, high-frequency effects were studied using narrow silicon MOSFETs and GaAs epitaxial layers (Bykov *et al.*, 1989) and metal films (Liu *et al.*, 1988; Lindelof *et al.*, 1987). In the first experiments on heterostructure quantum wires the influence of a microwave field on the weak electron localization (WEL) and the universal conductance fluctuations (UCF) was studied (Kuchar *et al.*, 1989, 1990).

The Hall effect of quantum wires is strongly influenced by the potential probes when using standard d.c. techniques. At microwave frequencies a contactless technique can be applied which does not disturb the dimensionality of the electronic system (Kuchar *et al.*, 1986; Volkov *et al.*, 1986). It was used to study the Hall effect of large 2DEG samples particularly in order to investigate the role of localization in the integer Quantum Hall Effect, IQHE (Lim *et al.*, 1990, Kuchar *et al.*, 1990).

In this paper we present results of two kinds of microwave experiments. The first regards the effect of microwave radiation on the d.c. conductivity of quantum wires made from InGaAs/InP heterostructures and n+-GaAs epitaxial layers. The theoretical description of the high-frequency effect is based on the introduction of a frequency-assisted inelastic scattering rate which replaces the phase coherence rate τ_ϕ at high frequencies. Together with the diffusion constant D, τ_ϕ determines the phase coherence length $L_\phi=(D\tau_\phi)^{1/2}$, which is the parameter giving the size of a mesoscopic system and also determines the effective sample size in the d.c. IQHE (Wei *et al.*, 1988; Lim *et al.*, 1990). The second type of experiment regards the contactless measurement of the Hall conductivity of narrow stripes of AlGaAs/GaAs.

EXPERIMENTAL TECHNIQUES

The effect of high-frequency radiation can be investigated in principle in two ways: (a) The "photoconductivity" technique records the change of the d.c. or low-frequency resistance due to the irradiation, or (b) a "transmission"-type technique, where the sample sees only the high-frequency field. We have applied technique (a) to the experiments on quantum wires made from InGaAs/InP and n+-GaAs epitaxial layers. They were mounted at the end of a rectangular waveguide with the high-frequency field parallel to the d.c. field. The d.c. voltage drop across the wire was much smaller than kT/e. Lock-in detection allowed to vary the modulation frequency of the d.c. current and/or the microwave radiation.

A transmission technique (b) was applied for the study of the microwave Hall effect where the Hall conductivity σ_{xy} can be measured (Kuchar *et al.*, 1986; Volkov *et al.* 1986). A Ka-band bridge (26.5 - 40 GHz) as described by Kuchar (1988) was used in our experiments. The essential part of the experimental set-up consists of crossed rectangular waveguides which act as polarizer and analyser. In this way the component of the wave transmitted through the 2DEG with a polarization perpendicular to the incident wave can be selected (Kuchar, 1988; Meisels *et al.*, 1987). Its electric-field amplitude is proportional to σ_{xy}/N, with $N=(1+\sigma_{xx}Z/2)^2+(\sigma_{xy}Z/2)^2$. Z is the impedance of the waveguide. N approaches 1 at high magnetic fields, i.e. far above the cyclotron resonance. In that case the microwave power transmitted through the sample is proportional to σ_{xy}^2.

The GaAs/GaAlAs single heterostructures were grown by MBE. For the microwave Hall effect experiments, an array of 3 µm-wide stripes, spaced by 1 µm-wide groves was produced by electron-beam lithography (Forchel *et al.*, 1988). The groves were edged down into the GaAs buffer layer. For the InGaAs/InP wires, MOVPE grown high-quality modulation-doped single quantum well structures with $n=8\times10^{11}$ cm^{-2} were used as the starting material. Long wires (width of 120 nm) were produced by electron-beam lithography and a combination of dry and wet etching (HCl:HNO3:H2O=1:1:2) (Menschig *et al.*, 1990). The processing allowed a variation of the electron mobility from 170000 in the starting material to 15000 cm^2/Vs in the narrowest wire. Electron-beam lithography was also used for the structurization of the n+-GaAs layers ($n= 1.3\times10^{18}$ cm^{-3}, thickness 50 nm) (Beaumont *et al.*, 1988). For both kinds of experiments (a) and (b), room temperature radiation was filtered out by a black polyethylene foil mounted in the waveguide in front of the sample.

TRANSPORT IN QUASI-ONE DIMENSIONAL WIRES UNDER MICROWAVE IRRADIATION

Theoretical Background

Several effects can cause a photoconductivity signal due to the microwave irradiation. Several of these are:

(a) "Dephasing": The high-frequency field destroys the constructive interference between electron waves back-scattered along time-reversed paths (Altshuler *et al.*, 1981). This implies that the negative WEL magneto-resistance of the wire should disappear. The universal conductance fluctuations (UCF) are not expected to be destroyed by this dephasing since there are no contributions of time-reversed paths at fields above the negative-magnetoresistance and a phase change should only change the UCF pattern but not the amplitude of the fluctuations.

(b) "Heating": Electron as well as lattice heating can destroy the WEL and the UCF as soon as the change of the energy of occupied conduction band levels is larger than the correlation energy E_c (Washburn *et al.*, 1986). We note that dephasing in high-frequency fields, process (a), can occur even if the heating is negligible, i.e. if the increase in electron energy is much less than the correlation energy E_c.

(c) "Photovoltaic Effects" can be of different origins. Rectification of microwaves can be caused by mesoscopic junctions (Falko, 1989) or by the contacts (if non-ohmic). Hot electron effects can cause a nonlinear current-voltage characteristic which can lead to a rectification at non-zero bias.

(d) "Thermoelectric Effects" due to asymmetric heating of the contacts.

The WEL contribution to the conductivity under the influence of a high-frequency electric field was calculated by Altshuler *et al.* (1981). The theoretical results are only given in the form of a complicated integral. Approximations for (a) low field amplitudes E_ω and high frequencies and (b) high field amplitudes and low frequencies are given by Altshuler *et al.* (1981). The two cases are distinguished by the value of a high-frequency parameter $\alpha=2e^2DE_\omega^2/\hbar^2\omega^3$ which contains the high-frequency field amplitude E_ω and the frequency ω. D is the diffusion constant. Case (a) corresponds to $\alpha<<1$, case (b) to $\alpha>>1$. The results are applicable to the frequency range where the effective inelastic scattering rate $1/t_0$ is larger than $1/\tau_\phi$, where $1/t_0$ is connected with the absorption of the high-frequency radiation and is given by

$$t_0 = \frac{1}{\omega\alpha} , \qquad \alpha << 1 , \tag{1a}$$

$$t_0 = \frac{1}{\omega}\left(\frac{45}{2\alpha}\right)^{0.2} , \qquad \alpha >> 1 . \tag{1b}$$

The frequency dependent contribution to the conductivity, due to a two-dimensional WEL is

$$\Delta\sigma = \frac{e^2}{\pi h}\ln(t_0/\tau) , \tag{2}$$

where τ is the elastic scattering time. For one-dimensional systems, $\Delta\sigma$ depends on frequency and field strength as follows: (a) for alpha<<1:

$$\Delta\sigma \propto \frac{\omega}{E_\omega}\ln\left(\frac{E_\omega}{\omega}\right)^2 , \tag{3a}$$

and (b) for a>>1:

$$\Delta\sigma \propto (\omega E_\omega)^{0.2} . \tag{3b}$$

The quasi-d.c. case is reached when $t_0>>\tau_\phi$. Then $\Delta\sigma\sim(e2/h)(D\tau_\phi)^{1/2}$, which is independent of ω and E_ω.

Experimental Results

The dephasing and the electronic heating were experimentally distinguished from the other effects by a proper choice of the modulation of the d.c. current and the microwave radiation. The microwave field strength E_ω in the wire can be only roughly estimated from the incident microwave power since E_ω very sensitively depends on the coupling into the sample at the open end of the waveguide. This yields $E_\omega\approx1$ V/cm for the highest field strength applied which certainly is an overestimate because of inefficient coupling and because of the large metallic contact areas attached to the wires.

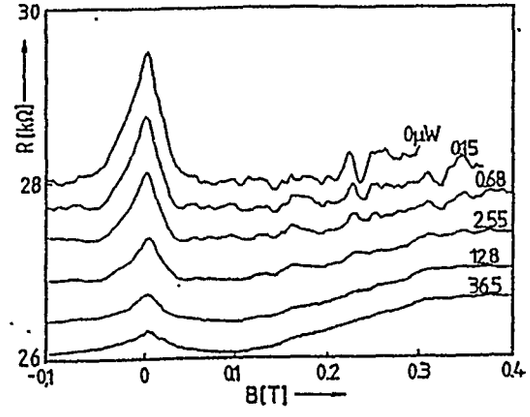


Fig.1. Magnetoresistance of a 260 nm wide InGaAs/InP wire with the incident microwave power (measured in the waveguide) as the parameter. (μ=56000 cm^2/Vs, ν=35 GHz, T=1.5 K.)

Figure1 shows experimental results concerning the WEL and the UCF in a 260 nm wide wire of InGaAs/InP (length L=50 m, n_s=8x10^{11} cm^{-2}, μ=5.6x10^4 cm^2/Vs at T=1.5 K). The UCFs are weaker in wires with lower mobility. In these wires, we observe the WEL peak at B=0 as well as a magnetoresistance peak followed by a very strong negative magnetoresistance at fields of the order 1 T (Fig.2). This kind of magnetoresistance was also observed in AlGaAs/GaAs and attributed to a geometrical effect (Thornton *et al.*, 1989; Menschig *et al.*, 1990). At the magnetoresistance peak, the cyclotron radius l_c roughly fits into the wire width causing a maximum of the interaction with the side walls when W/l_c=0.55. This interaction is drastically reduced when shrinking the cyclotron radius by increasing the

magnetic field. When increasing the wire width the "geometrical" peak in the magnetoresistance shifts towards zero field and eventually overlaps with the weak localization peak as observed in our samples. Strong microwave fields totally suppress the WEL peak and the SdH oscillations but leave the geometrical effect essentially unchanged. This shows that the latter phenomenon is not sensitive to electron heating (as long as $\omega_c\tau > 1$) and that it actually is of geometrical origin.

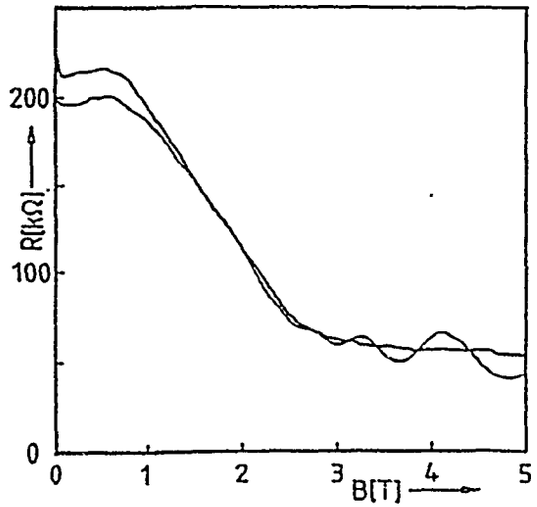


Fig.2. D.c. magnetoresistance of the 120 nm wire with μ=15000 cm^2/Vs. The curve without the SdH oscillations was recorded under microwave irradiation (power comparable to that for the lowest trace in Fig.1), with T=1.5 K.

The dependence of the WEL contribution to the resistance at B=0, ΔR, on the microwave field is plotted in Fig.3 for two InGaAs/InP wires and one n$^+$-GaAs wire. ΔR is defined as $R(B=0)-R_{min}(B)$. In the upper range of microwave fields, the data can be described by a dependence on field as $\Delta R \propto E_w^s$, with s=-0.55 (straight lines in Fig.3).

From our estimate of the electric field amplitude E_ω, we calculate that $\alpha<1$ in the whole field range. This indicates that $t_0 \approx \tau_\phi$ at the field strength where ΔR becomes strongly dependent on E_ω (Fig.3). ΔR is independent of E_ω if $t_0 >> \tau_\phi$. In the 260 nm wire, the value of τ_ϕ obtained from the fit to the d.c. WEL magnetoresistance is 50 ps. Setting this value equal to t_0 and using (1b), we obtain α=0.1 (E_ω=0.6 V/cm), justifying our assumption that $\alpha<<1$ over the whole experimental field range. The value of 0.6 V/cm is in order of magnitude agreement with the estimate given above. Our results provide the first - at least qualitative - evidence that the theoretical treatment by Altshuler *et al.* (1981) gives a reasonable description of the influence of high-frequency fields on the quantum contributions to the d.c. transport in heterostructure quantum wires, viz. that the deviations from ΔR_{dc} occur when $1/t_0$ becomes larger than $1/\tau_\phi$. However, we attribute these deviations to an electron heating effect and not to the pure dephasing effect without heating. The evidence for this conclusion is the observation that the UCF pattern does not change, only the UCF amplitude is reduced with increasing E_ω.

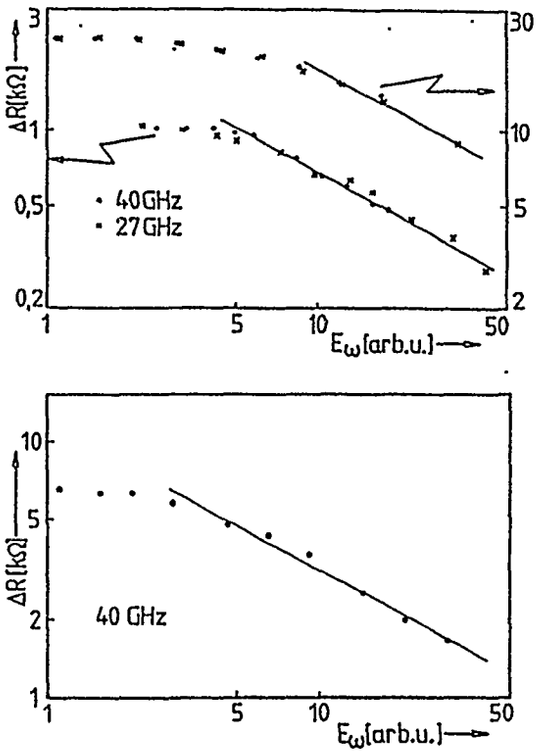


Fig.3. Weak localization contribution ΔR (see text) as a function of the microwave field strength for two InGaAs/InP wires, length L= 50 μm (upper diagram) and an n$^+$-GaAs wire, length L= 3 μm, width W= 0.5 μm (lower diagram). The maximum applied high-frequency field strength is approximately 1 V/cm. The horizontal scales for the different sets of data are adjusted so that the steep sections can be directly compared. Lower set of the InGaAs/InP data: wire with W= 260 nm, μ=56000 cm^2/vs; upper set: 120 nm, 15000 cm^2/Vs at T= 1.5 K.

The WEL data on the n$^+$-GaAs wire are evaluated in the same way and also plotted in Fig.3. The general dependence on E_ω is the same as for the heterostructure wires. In particular, the exponent in the power-law dependence at high E_ω is also close to -0.55. Also, the UCF amplitude decays faster with E_ω than ΔR does. This can be understood by considering the electron heating mentioned above. At low temperatures, it can be assumed that L_ϕ is not determined by electron-phonon scattering. Therefore, it decreases with increasing electron energy as it does with increasing lattice temperature ($L_\phi \propto T^{-p/2}$). According to the

theoretical results given by Washburn and Webb (1986), the WEL contribution ΔR is $\propto L_\phi$, the UCF contribution is $\propto L_\phi^{3/2}$. With increasing microwave power and consequently increasing electron energy a stronger decay of the UCF amplitude than of ΔR is expected - as observed experimentally in both kinds of quantum wires. The increase in electron energy is only effective if it is larger than the correlation energy. This condition is equivalent to $t_0<\tau_\phi$.

MICROWAVE HALL EFFECT OF AN ARRAY OF AlGaAs/GaAs STRIPES

The microwave Hall effect of large homogenuous AlGaAs/GaAs samples has been studied experimentally by two groups (see Kuchar, 1988). The general shape of the magnetic field dependence of the Hall conductivity at microwave frequencies of about 30 GHz (Ka-band) corresponds to the classically expected σ_{xy}^2 behavior (Fig.4). On expanded scales, the integer quantization of the Hall conductivity around the filling factor i=4 is clearly visible. The width of the plateaus and their temperature dependance were explained within the localization model (Wei *et al.*, 1988; Lim *et al.*, 1990; Huckestein *et al.*, 1990).

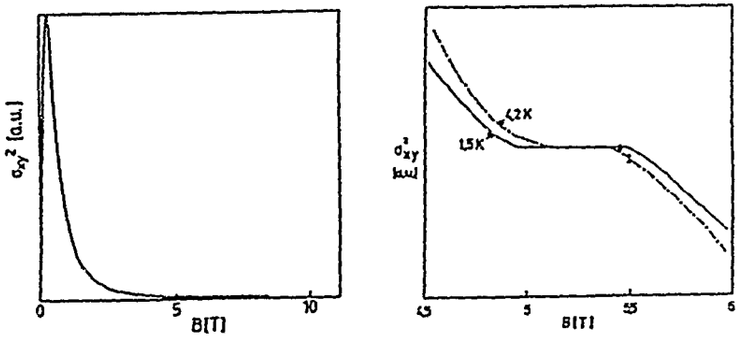


Fig.4. Crossed-waveguide transmission ($\propto\sigma_{xy}^2$) of a large sample of AlGaAs/GaAs (n_s=5.2x10^{11} cm^{-2}) at ν=35 GHz and T= 4.2 K (left). Right: Expanded section around the i=4 plateau.

Experiments on many parallel wires are more difficult then on large homogenuous samples because of the wires acting as antennas. Already, a very small deviation of the wire axis from the direction of the microwave field couples a field component proportional to σ_{xx} into the crossed waveguide (the analyser). Figure 5 shows experimental results obtained on an array of about 500 3-µm-wide stripes at 38.7 GHz and 2 K. The two curves shown correspond to two settings of the phase shifter in the microwave bridge which differ by 90°. By a proper choice ("0°") of the microwave phase on the i=4 plateau (at about B= 7 T), a signal was observed which shows the typical σ_{xy}^2 shape but sits on a background of approximately the same height. Superimposed are Hall plateaus as in large samples. After a phase shift of 90°, a typical σ_{xx}^2 curve with Shubnikov-de Haas oscillations is observed. The width of the stripes is not small enough to produce quantum size effects (Hansen and Kotthaus, 1988), but shows depolarization effects at microwave frequencies (Kuchar, 1988).

This leads to a drastically reduced conductivity far away from the depolarization eigen-frequency $\nu_d=(n_s e^2/4\pi^2 W d m^* \varepsilon_d \varepsilon_0)^{1/2}$ of a single stripe. W is the width of the stripe, ε_d the average dielectric constant of the medium in which the electron gas is imbedded and the surrounding medium. With the parameters of our stripes (n_s=6.4x10^{11} cm^{-2}, ε_d=6.5, W=3 µm), we obtain ν_d=450 GHz. Nevertheless, at the measuring frequency of 38.7 GHz, a strong signal could be observed. This is attributed to the narrow spacing of the stripes (1 µm) which allows a coupling of the electric fields between them. These experiments demonstrate that the Hall effect on quasi-one dimensional wires can be observed without using current contacts and potential probes.

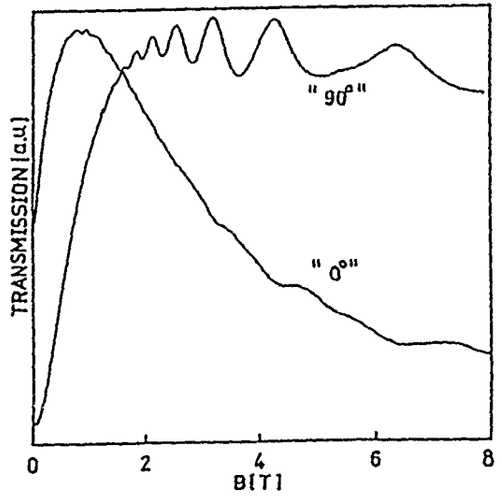


Fig.5. Crossed-waveguide transmission of the array of stripes in AlGaAs/GaAs. The dimension of one conducting stripe is 3 µm x 2.3 mm, ν=38.7 GHz,T=2.1 K.

ACKNOWLEDGEMENT

This work was supported by the "Österreichische National-bank, Jubiläums-fonds-projekt 3555", Austria, and by "Stiftung Volkswagenwerk", West Germany.

REFERENCES

Altshuler, B.L., Aronov, A.G., and Khmelnitskii, D.E., 1981, Sol. State. Comm., 39:619.

Bykov, A.A. et al., 1989, JETP Letters, 49:13.

Falko, V., 1989, Europhysics Letters, 8:785.

Forchel, A., Leier, H., Maile, B.E., and German,R., 1988, Adv. Solid State Phys., 28:99.

Galchenkov, .A., Grodnenskii, .M., Kostovetskii, .V., and Matov, O.R., 1988, JETP Letters, 46:542.

Hansen, W., and Kotthaus, J.P., 1988, "Springer Series in Solid State Sciences", Eds. Heinrich,H., Bauer,G., and Kuchar, F., 83:187.

Heinrich, H., Bauer, G., and Kuchar, F. , 1988, "Physics and Technology of Submicron Structures," 83.

Huckestein, B., Apel, W., and Kramer, B., 1990, "Springer Series in Solid State" (in print).

Khmelnitskii, D.E., 1984, Physica, 126:B235.

Kramer, B., Ono, Y., and Ohtsuki, T., 1989, "Springer Series in Solid State Sciences", 87:24.

Kuchar, F., 1988, Adv. Sol. State Phys., 28:45.

Kuchar, F., Meisels, R., Weimann, G., and Schlapp, W., 1986, Phys.Rev. B, 33:2965.

Kuchar, F., Meisels, R., Lim, K.Y., Pichler, P., Weimann, G., and Schlapp, W., 1987, Physica Scripta, 19:79.

Kuchar, F., Lutz, J., Lim, K.Y., Weimann, G., Schlapp, W., Forchel, A., and Menschig, A., 1989, Extended Abstracts, Symposium "New Phenomena in Mesoscopic Structures", Hawaii, p.200.

Kuchar, F., Lutz, J., Lim, K.Y., Meisels, R., Weimann, G., Schlapp, W., Forchel, A., Menschig, A., and GrÅtzmacher, D., 1990, "Quantum Coherence in Mesoscopic Systems" (ed. Kramer,B.), Plenum (in print).

Lim, K.Y., Auer, I., Kuchar, F., Weimann, G., Schlapp, W., Forchel, A., and Menschig, A., 1990, Surf. Science (in print).

Lindelof, P.E., and Wang, S., 1987, Phys. Rev. Letters, 59:1156.

Liu, J., and Giordano, N., 1988, "Springer Proc. in Physics", 28:159.

Meisels, R., and Kuchar, F., 1987, Z.Phys., 67:443.

Menschig, A., Forchel, A., Roos, B., Germann, R., Pressel, K., Heuring, W., and GrÅtzmacher, D., 1990, (submitted to Appl. Phys. Letters).

Thornton, T.J., Roukes, M.L., Scherer, A., van de Gaag, B.P., 1989, Phys. Rev. Letters, 63:2128.

Volkov, V.A. et al., 1986, JETP Letters, 43:328.

Washburn, S., and Webb, R.A., 1986, Adv. in Physics, 35:375.

Wei, H.P., Tsui, D.C., Paalanen, M.A., and Pruisken, A.M.M., 1988, Phys. Rev. Letters, 61:1294.

(22)

NON-EQUILIBRIUM CARRIER TRANSPORT IN SMALL STRUCTURES

Kenji Taniguchi and Chihiro Hamaguchi

Department of Electronic Engineering
Osaka University, Suita, Osaka 565, Japan

INTRODUCTION

Current VLSI fabrication technology has progressed rapidly and is pushing toward deep submicron dimension devices. In such small devices, high-field effects become more pronounced, which are velocity saturation, non-stationary transport (ballistic transport and velocity overshoot), impact ionization, and hot-carrier induced degradation of devices. However, many fundamental problems on non-equilibrium carrier transport still remain to be clarified, which makes it difficult to establish a proper guideline for designing small geometry devices. Especially, detailed understanding of high-field effects in future deep submicron devices strongly requires a precise knowledge on non-equilibrium carrier transport.

In the next section, we first attempt to lay a conceptual framework for an essential physics for small devices and present simple analytical expressions necessary to characterize these devices, including carrier transport equations derived from Boltzmann transport equation, which conserve both momentum and energy of carriers. Then, we discuss non-stationary transport phenomena such as ballistic transport and velocity overshoot. The remaining part of this latter section is devoted to an assessment of the validity of the analytical expression through comparison with the results calculated by more accurate iterative method. In the third section, from a practical view point, we introduce electrical characteristics of a small geometry MOSFET and try to explain hot-electron related phenomena inherent to such small devices: velocity saturated I_d-V_d characteristics, substrate current, gate current characteristics. The final two sections are devoted to our experimental and modeling results relating to non-equilibrium carrier transport, such as the determination of energy relaxation time by optical measurement techniques, an impact ionization model, and extended drift-diffusion model for non-equilibrium carrier transport.

NON-EQUILIBRIUM TRANSPORT EQUATIONS

The Boltzmann equation

In electron-transport theory, we are concerned with average effects produced by many electrons such as carrier density, average velocity and average energy. In order to study these macroscopic physical quantities, it is convenient to introduce a distribution function f(k,r,t) expressing the density of electrons in space r and momentum k at time t. Then, all the physical quantities of interest may be expressed in terms of f(k,r,t): The electron density n(r,t) and carrier drift velocity v_d(r,t) are given by

$$n(r,t) = \int f(k,r,t)\, dk \ , \tag{1}$$

$$v_d(r,t) = \int v(k) f(k,r,t)\, dk \ . \tag{2}$$

The following Boltzmann transport equation is derived from the fact that the derivative of the distribution function with respect to time along a particle trajectory r, k vanishes in the entire phase space as:

$$\frac{df(r,k,t)}{dt} = 0 \ , \tag{3}$$

which expands to yield

$$\frac{df}{dt} = \frac{\partial f}{\partial t} + \frac{\partial f}{\partial k} \cdot \frac{dk}{dt} + \frac{\partial f}{\partial r} \cdot \frac{dr}{dt} = 0 \ . \tag{4}$$

For an external electric field E, we can write

$$\frac{dk}{dt} \cdot \nabla_k f = -\frac{q}{h} E \cdot \nabla_k f \ . \tag{5}$$

Similarly, we can write

$$\frac{dr}{dt} \cdot \nabla f = v \cdot \nabla f \ . \tag{6}$$

The steady-state Boltzmann transport equation ($\partial f/\partial t = 0$) is therefore

$$-\frac{q}{h} E \cdot \nabla_k f + v \cdot \nabla f = \left(\frac{df}{dt}\right)_{coll} \ . \tag{7}$$

The form of $(\partial f/\partial t)_{coll}$ depends on the nature of the specific scattering process. A rigorous solution of the above equation is formidable, with no closed form solutions available.

Energy and momentum conservation equations

A simple solution of the Boltzmann equation (7) becomes possible when the effects of scattering can be described in terms of a relaxation time. By multiplying (7) by the group velocity v of each carrier and integrating the equation over the entire momentum space, the following momentum conservation equation is obtained (Snowden, 1986)

$$v_d \cdot \nabla v_d - \frac{qE}{m^*} + \frac{1}{m^* n} \nabla (n k_B T_e) = -\frac{v_d}{\tau_m} \ , \tag{8}$$

where m^* is the effective mass of carriers, k_B Boltzmann's constant, T_e electron temperature, E electric field, v_d drift velocity of carriers, τ_m the momentum relaxation time. Under the assumption of a displaced Maxwellian distribution, the Boltzmann equation (7) may be multiplied by the kinetic energy ($v^2/2m^*$) and integrated over momentum space. This yields the following energy conservation equation:

$$v_d \cdot \nabla \langle \varepsilon \rangle - q v_d \cdot E + \frac{1}{n} \nabla (n v_d k_B T_e) + \nabla \cdot S = -\frac{\langle \varepsilon \rangle - \langle \varepsilon_o \rangle}{\tau_e} \ , \tag{9}$$

where S is the energy flux. The average electron energy $\langle \varepsilon \rangle$ is given by

$$\langle \varepsilon \rangle = \frac{m^* v_d^2}{2} + \frac{3k_B T_e}{2} , \tag{10}$$

where τ_e is the energy relaxation time. Both time constants, τ_m and τ_e, are considered to be functions of average electron energy. The momentum conservation equation (8) is further simplified by assuming that the term $v_d \cdot \nabla v_d$ is small compared with other terms because the thermal energy $3k_BT_e/2$ is typically an order of magnitude greater than the kinetic energy (Baccarani and Wordeman, 1985).

After the simplifications described above, *a set of the carrier transport equations* become

$$v_d = \frac{\tau_m}{m^*}\left[-qE - \frac{k_B}{n}\nabla(nT_e) \right] , \tag{11}$$

$$\frac{5}{2}k_B v_d \cdot \nabla T_e = -q v_d \cdot E - \frac{3k_B(T_e - T_o)}{2\tau_e} , \tag{12}$$

where T_o is the lattice temperature. A comparison of (11) with the classical equation (drift-diffusion equation) reveals that the equation has additional terms for spatial variation in the electron temperature, that is, macroscopic carrier flow caused by the difference in electron gas pressure. This temperature gradient term becomes significant in sub-micron devices, which will be discussed later. Equation (12) indicates that, at high electric field, the electron temperature is always higher than the lattice temperature. The carrier transport under $T_e > T_o$ is refered as non-equilibrium carrier transport hereafter.

Basic Physical parameters

Before explaining the non-stationary transport, basic physical parameters used in the momentum and energy conservation equations described above should be evaluated. In the following, a uniform and homogeneous semiconductor is assumed to make the derivation simple.

(1) Momentum relaxation time

In a non-degenerate semiconductor, the generalized Einstein relation is given by

$$\frac{D}{\mu} = \frac{k_B T_e}{q} , \tag{13}$$

where D is diffusion coefficient and μ is mobility. Since Monte Carlo simulation reveals that the diffusion coefficient of electrons only weakly depends on electric field (Jacoboni *et al.*, 1977), the diffusion coefficient can be assumed to be independent of electric field. Therefore, (13) leads to the relation

$$\frac{\mu(E)}{\mu_o} = \frac{T_o}{T_e(E)} . \tag{14}$$

In a uniform and homogeneous semiconductor, the momentum conservation equation (11), is further simplified as

$$v_d = \frac{q\tau_m}{m^*} E . \tag{15}$$

From (14) and (15), the momentum relaxation time is obtained (Baccarani and Wordeman, 1985) as

$$\tau_m(E) = \frac{m^* \mu_o}{q}\left(\frac{T_o}{T_e} \right) , \tag{16}$$

since $\mu = q\tau_m/m^*$. The momentum relaxation time is therefore inversely proportional to the electron temperature because of a high scattering rate for high energy electrons. The relation (16) also leads to the saturation of carrier drift velocity at high electric field as shown in Fig.1.

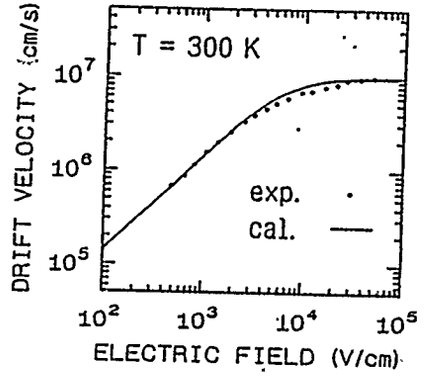


Fig.1 Comparison of computed hot electron drift velocity as a function of the driving field (solid lines) with the experimental data (Jacoboni *et al.*, 1977). The drift velocity was calculated by using (15), (16) and (20). Theoretical drift velocity saturates in the electric field above 1×10^4 V/cm.

(2) Energy relaxation time

Under the uniform and homogeneous condition, spatial variation term in (12) can be also neglected. Then,

$$T_e = T_o - \frac{2q\tau_e}{3k_B} v_d \cdot E . \tag{17}$$

Substituting (15) into (17), the energy relaxation time is obtained as

$$\tau_e = \frac{3k_B m^*(T_e - T_o)}{2\tau_m q^2 E^2} . \tag{18}$$

Using the Caughey-Thomas expression (Caughey and Thomas, 1967), given by

$$v_d = \mu(E)E = \frac{\mu_o E}{\sqrt{1 + \left(\frac{\mu_o E}{v_s}\right)^2}} \quad (19)$$

and (14), we find the relation between the applied electric field and electron temperature as

$$\frac{T_e}{T_o} = \sqrt{1 + \left(\frac{\mu_o E}{v_s}\right)^2} \quad (20)$$

where v_s is the saturation velocity (1×10^7cm/s) and μ_o is the low field mobility (1,400 cm^2/Vs). From (18) and (20), the following energy relaxation time as a function of electron temperature is derived (Baccarani and Wordeman, 1985):

$$\tau_e(T_e) = \frac{3k_B T_o \mu_o}{2qv_s^2}\left(\frac{T_e}{T_e + T_o}\right) \quad (21)$$

The energy relaxation time increases with electron temperature, showing a distinct difference from the momentum relaxation times as described in (16). Figure 2 shows the dependence of τ_m and τ_e upon electron temperature at room temperature (Baccarani and Wordeman, 1985).

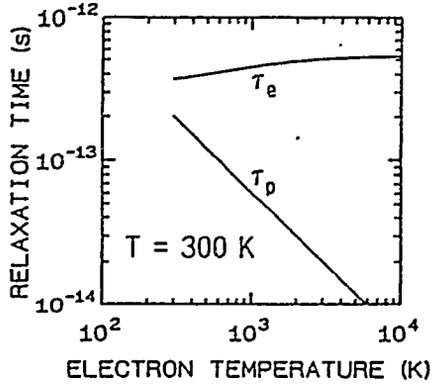


Fig.2 Momentum and energy relaxation times as a function of electron temperature at room temperature The momentum relaxation time is inversely proportional to electron temperature, while the energy relaxation time is less dependent on electron temperature.

NONSTATIONARY TRANSPORT

The advent of high resolution lithography techniques makes it possible to fabricate very small devices in which individual feature sizes might well be below the scale of 0.1 micron. It should be noted that transport phenomena in the small size devices may still be based on the Boltzmann equation, but the features characterizing carrier transport can be very different from those related to steady state: Carrier transport is often in non-stationary conditions characterized by strong spatial non-uniformity of electric field. Therefore, to gain the proper insight into electrical behavior of the small scale devices, understanding of carrier transport in such devices is required.

In this section, we describe the physics of nonstationary transport, which is categorized into two regimes: (1) ballistic regime and (2) velocity overshoot regime.

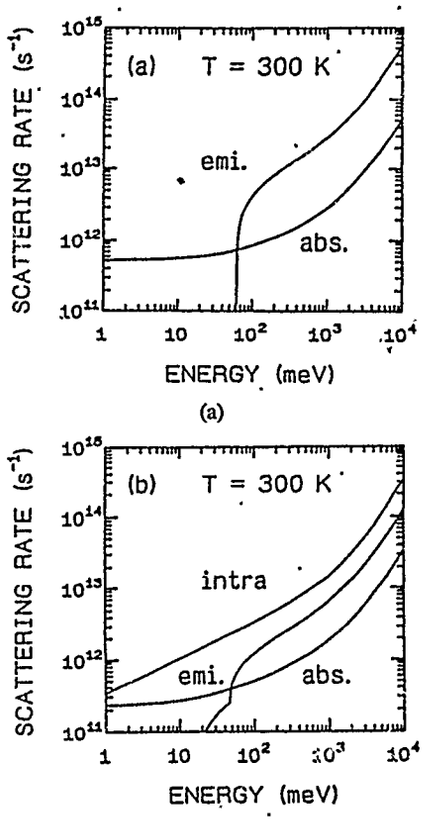


Fig. 3 Electron-phonon scattering rate for silicon at room temperature. (a) Scattering due to optical phonons, in which optical phonon emision rate significantly increases with electron energy above 60 meV and (b) Scattering due to acoustic phonons; intra-valley scattering dominates over inter-valley scattering.

Carrier transport in the ballistic regime

Monte Carlo simulation reveals that, in the ballistic transport regime, carriers having low energy are mainly scattered by acoustic phonons at room temperature. A ballistic mean free path, Λ, is simply given by (Rosencher, 1981)

$$\Lambda = \frac{qE\tau_m^2}{2m^*} \quad . \tag{22}$$

At low electric field, kinetic energy of electrons gained from the electric field during free flight is very small compared with thermal energy according to (20), meaning a constant momentum relaxation time. Therefore, the ballistic mean free path Λ linearly increases with the applied electric field.

The mean free path , however, saturates at high electric field ($E > 10^4$V/cm), because the scattering rate due to optical phonons is roughly proportional to the square root of kinetic energy for high energy electrons as shown in Fig.3. Note that Λ is defined as the product of average carrier velocity during free flight and momentum relaxation time

$$\Lambda \approx \text{constant} \quad . \tag{23}$$

The saturated mean free path Λ is only 20 nm at room temperature while it increases to 35 nm at liquid nitrogen temperature. This means that the contribution of the ballistic transport to the electrical characteristics will be significant only for sub-tenth μm Si devices.

In the ballistic transport regime, the convective term ($m^*v_d^2/2$) in (10) can be of the same order of magnitude as the thermal energy ($3k_BT_e/2$) so that the use of the momentum and energy conservation equations derived above are not suitable for the description of the ballistic transport.

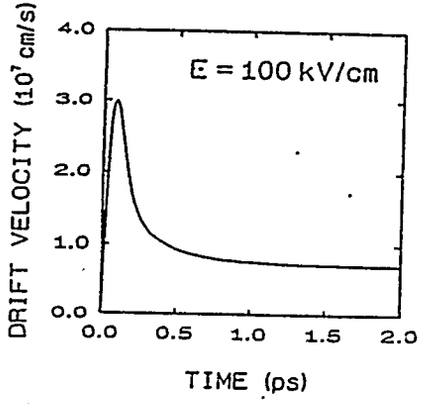


Fig. 4 Temporal velocity overshoot characteristics at room temperature. Velocity overshoot up to 3×10^7 cm/sec.is observed in the initial short period.

Carrier transport in velocity overshoot regime

Figure 4 shows a typical nonstationary transport characteristics in Si. As shown in the figure, the end of ballistic transport coincides closely with the maximum of the velocity. Therefore, we can identify the range after the maximum to the point of reaching the steady-state velocity as the overshoot range, although the separation of ballistic and overshoot range is not always as clear as indicated above. The overshoot range following the ballistic transport regime is the time period which is necessary for the momentum distribution to interact and arrive at a steady state in high electric field. In the following, we first describe a temporal velocity overshoot followed by velocity overshoot in real space.

(1) Temporal velocity overshoot

For a homogeneous system in the relaxation approximation, the time dependent momentum conservation equation is obtained from (11) as

$$\frac{dv_d}{dt} = -\frac{qE}{m^*} - \frac{v_d}{\tau_m} \quad . \tag{24}$$

In order to get physical insight into the nonstationary transport, we first assume constant momentum relaxation time. Under this assumption, the solution of (24) is simply given by

$$v_d(t) = -\frac{qE\tau_m}{m^*}\left[1 + \exp\left(-\frac{t}{\tau_m}\right)\right] \quad , \tag{25}$$

which shows an initial rapid rise of the drift velocity with time, on a temporal scale τ_m and the drift velocity ultimately reaches the steady state velocity.

If we consider more realistic transport, by assuming τ_m to vary inversely with electron energy as shown in (16), the momentum and energy conservation equations can be expressed as (Ferry, 1985)

$$\frac{dv_d}{dt} = -\frac{qE}{m^*} - \frac{v_d}{\tau_{mo}}\left(\frac{T_e}{T_o}\right) \quad , \tag{26}$$

$$\frac{dT_e}{dt} = -\frac{2q}{3k_B}v_d \cdot E - \frac{T_e - T_o}{T_o} \quad , \tag{27}$$

where τ_{mo} is independent of electron temperature. The second term on the right-hand side of (26) indicates friction acting on the electron gas system. In the case of $\tau_e >> \tau_m$, which is physically realized in most semiconductors, v_d grows rapidly on the momentum relaxation time scale, whereas electron temperature grows more slowly than v_d on the energy relaxation time scale. In this case, the velocity initially rises to a value much larger than the steady state and then damps to the steady-state velocity for times of the order of the energy relaxation . The velocity overshoot is a direct consequence of the slower rise of electron temperature. It is, however, worthwhile to note that the temporal velocity overshoot would not occur in normally operating devices because the RC delay time, due to parasitic capacitance, is much larger than the energy relaxation time. Furthermore, it should be noted that the velocity overshoot effect is only observed in devices with low electric field and high impurity concentration.

(2) Spatial velocity overshoot

In order to investigate the *spatial* velocity overshoot, the time derivative in (26) and (27) should be replaced by spatial derivative as

$$m^*v_d \cdot \frac{dv_d}{dx} + \frac{m^*v_d}{\tau_m(T_e)} = -qE \quad , \tag{28}$$

$$\frac{d(T_e - T_o)}{dx} + \frac{m^*(T_e - T_o)}{q\tau_e\tau_m(T_e)E} = -\frac{2qE}{3k_B} \quad . \tag{29}$$

Since these two equations are decoupled, (29) is solved for a given E and the resulting $T_e(x)$is

then substituted into (28), which is solved for drift velocity. It should be noted that this simplification is allowed only to provide a qualitative description of nonstationary transport because spatial derivative terms in (11) and (12) are ignored.

Validity of the analytical expression

(1) Local electric field dependent carrier mobility model

More realistic solutions have been performed by using direct iterative solutions of the Boltzmann transport equation (Rees, 1969; Sonoda *et al.*, 1989). We discuss the important features of the spatial velocity overshoot effect. Figure 5 shows the simulated results on a 1.0 μm long sample having a field discontinuity at x=0.5 μm by incorporating periodic boundary conditions. On the left side, the electric field is assumed to be 9 kV/cm, while it changes abruptly to 9.9 kV/cm on the right side. In the figure, velocity calculated from carrier mobility is also shown (dotted lines) as a reference.

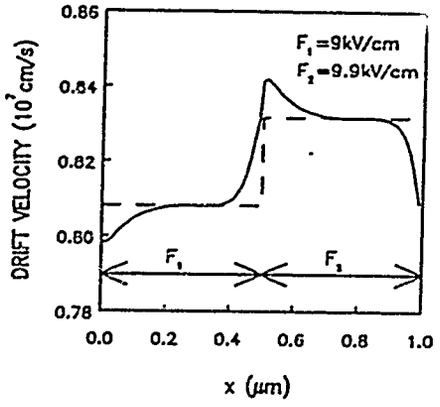


Fig. 5 Spatial velocity overshoot characteristics at room temperature calculated by the direct iterative solution of the Boltzmann transport equation. Velocity overshoot and undershoot are observed at the low--high field and high-low field interfaces, respectively

The calculated velocity curve exhibits a maximum (velocity overshoot) in the vicinity of the interface between the two regions, while the minimum (velocity undershoot) is at the high-low field boundary. In a rapidly-varying electric field in the space domain, as shown in Fig.5, carrier velocity increases even in the low-field region before experiencing the large field, and reaches a maximum at the interface due to the existence of a diffusion component. This result indicates that carrier velocity is not a unique function of the local electric field and the gradient of the field, although a carrier velocity model, including not only local field but also the gradient of the field, is correct in the first order of approximation and is more accurate than any local field models.

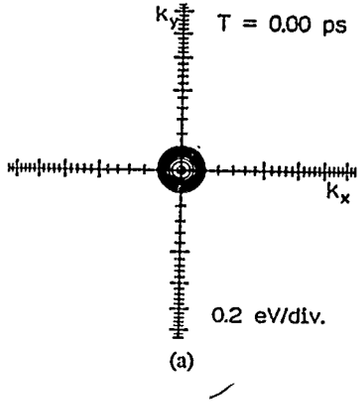
(2) Time evolution of distribution function during non-stationary transport

In the following, in order to obtain the physical image of the non-stationary transport, we describe the time evolution of electron distribution function after the application of high electric field. Figure 6 shows the time evolution of the distribution function in momentum space at room temperature. The initial distribution function shifts along the electric field, while keeping the original structure, during the initial stage of the time evolution up to 0.05 psec (<< τ_m at low electron temperature). We call this a shifted distribution function as a drifting peak, which mainly contains ballistic carriers. After 0.05 psec, electrons scatter randomly into momentum space so that the distribution function begins to expand, while there still exists the drifting peak containing only small number of electrons. The drifting peak disappears after 0.25 psec because most of the ballistic electrons reach high energy and emit optical phonons because of their high scattering probability. The electrons scattered by optical phonons lose their energy and return back to the center of the momentum space, resulting in the increase of the electron concentration near the center. High energy electrons are also scattered by acoustic phonons without losing their energy, which greatly expands the distribution function in the momentum space as shown in Fig.6(c), which corresponds to the steady state carrier distribution.

(3) Momentum relaxation time vs. mean electron energy

Based on the discussion described above, we investigated the momentum and energy relaxation times in more detail by using the iterative method. As is described earlier, both relaxation times are considered to be a unique function of average electron energy. Numerical calculation under various electric field conditions reveals that the energy relaxation time is indeed a unique function of electron energy, while the momentum relaxation time depends on the electric field history acting on the electron gas system. Figure 7 shows a typical example of momentum relaxation time as a function of mean energy, which is calculated from the transient drift velocity characteristics after switching off various applied electric fields. The resulting momentum relaxation time is quite short at high energy state, while it sharply increases with decreasing mean energy. All the curves come to close and are merged into a universal curve at low mean energy.

Momentum relaxation time characteristics shown in Fig.7 can be explained as follows. Since high energy electrons have high scattering rates, they lose their memory of direction quite rapidly. Therefore, the contribution of high energy electrons to the drift velocity becomes small except an initial short period. So, after shutting off the electric field, the drift velocity is mainly controlled by low energy electrons which have low scattering rate. This leads to large relaxation time in spite of high mean energy.



(a)

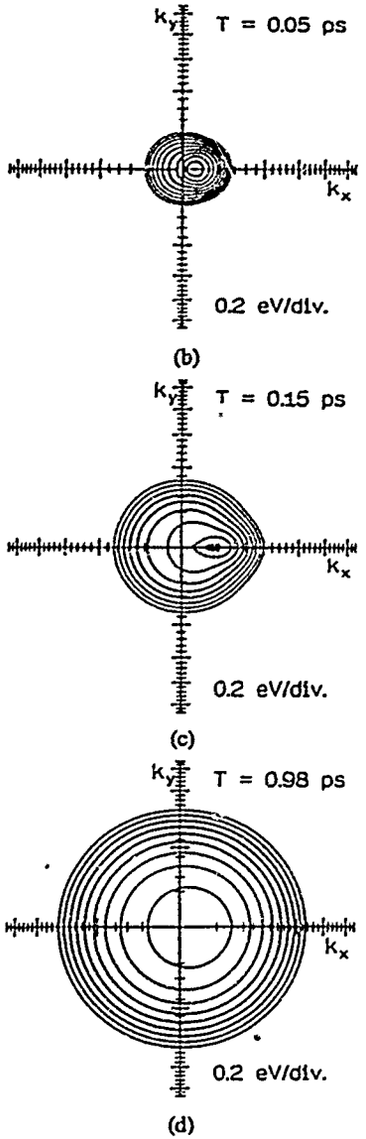


Fig. 6 Time evolution of distribution function of electrons in silicon at room temperature after the application of high electric field. (a) Thermal equilibrium distribution function at room temperature, (b) distribution function at 0.05 psec after applying high electric field of 100 kV/cm and (c) steady state distribution function at the electric field of 100 kV/cm.

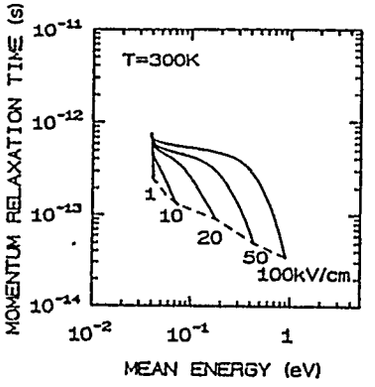


Fig. 7 Momentum relaxation time as a function of mean electron energy: relaxation time during the recovery after shutting off the electric field (solid lines) and steady-state relaxation time (dotted line).

NON-EQUILIBRIUM CARRIER TRANSPORT IN SMALL SIZE MOSFETS

With decreasing size of MOSFETs for both high-speed and high-packing density, electric field inside the device increases. This increase is caused by the use of a supply voltage that is not scaled for constant field to make the chip voltage compatible with that of other system components. The net result is higher fields that create hot carriers, which in turn cause long-term instability. In this section, we first explain the electrical characteristics of short channel MOSFETs. Then, we extract analytical models relating to hot carrier effects inherent to small geometry devices: substrate current and gate current.

Drain current model including velocity saturation effect

It is well known that the basic "gradual channel" approximation approach for long-channel MOSFETs is inadequate for short-channel devices. One of the most significant factors limiting the drain current in short-channel devices is the velocity saturation effect, which is associated with non-equilibrium transport. Many empirical velocity-field models have been published. However, to avoid a prohibitively complicated derivation of the drain current, a proper choice of the velocity-field models is required: Caughey-Thomas model is not suitable for the derivation of analytical expression. In the following, a two-region piecewise model (Ko, 1989) is used to circumvent the problem.

$$v_d = \begin{cases} \dfrac{\mu_o E}{1+\dfrac{E}{E_o}}, & E < E_o \ , \qquad (30a) \\ v_s \ , & E > E_o \ . \qquad (30b) \end{cases}$$

In this model, the carrier will saturate at E_s. This piecewise model well predicts the experimental data over a wide range of electric field while slightly overestimating the carrier velocity near E_o. Saturation electric field is 4×10^4 V/cm for silicon substrate.

Using (30a), the drain current of MOSFETs at position y in the channel can be expressed as

$$I_d = WC_o[V_g - V_t - \phi(y)] \frac{\mu_o E(y)}{1 + E(y)/E_o} \quad (31)$$

where y is the distance from the source junction, W the channel width, C_o the gate capacitance per unit area, V_g the gate voltage, V_t the threshold voltage, and $\phi(y)$ the channel potential. After integrating (31) from y = 0 to y = L, with $\phi(0) = 0$ and $\phi(L) = V_d$, we obtain

$$I_d = \frac{W\mu_o C_o}{L}\left[V_g - V_t - \frac{V_d}{2}\right]\frac{V_d}{1 + \frac{V_d}{E_o L}} \quad (32)$$

where $1/[1 + (V_d/E_oL)]$ is a current reduction factor due to velocity saturation. In this model, drain current saturates when the carriers at the drain are velocity saturated. The saturated drain current is given by the product of the carrier density and the saturated velocity. Therefore, the current is

$$I_d = v_s WC_o(V_g - V_t - V_{ds}) \quad (33)$$

Equating (32) and (33) at $V_d = V_{ds}$, we obtain

$$V_{ds} = \frac{E_o L(V_g - V_t)}{V_g - V_t + E_o L} \quad (34)$$

Substituting (34) into (33), the saturation drain current is obtained

$$I_{d,} = v_s WC_o \frac{(V_g - V_t)^2}{V_g - V_t + E_o L} \quad (35)$$

If $E_oL >> V_g - V_t$, the drain current is given by

$$I_{ds} = \frac{\mu_o WC_o}{2L}(V_g - V_t)^2 \quad (36)$$

because $E_o = 2v_s/\mu_o$. Note that even a very short channel device exhibits long-channel behavior at a low enough gate voltage. On the other hand, if $E_oL << V_g - V_t$, the saturation drain current is linearly proportional to gate voltage as

$$I_{ds} = v_s WC_o(V_g - V_t) \quad (37)$$

In the above equation, carrier velocity is saturated in the whole channel region because $C_o(V_g - V_t)$ represents the number of inversion carriers per unit area. Although the drain current should be limited by (37), which is independent of channel length, some of the reported drain current of ultra-small MOSFETs (Sai-Halasz *et al.*, 1988) well exceed the value given by (37). The deviation of the experimental data from the analytical model prediction is a clear sign of velocity overshoot. A Monte Carlo simulation including the full band structure (Fischetti and Laux, 1988) reveals that the carrier velocity well exceeds twice as much as saturation velocity near the drain edge in the 0.07 μm device. The Monte Carlo simulation also represents that to achieve higher drain current than (37), the carriers must reach a high velocity already at the source edge. It should also be noted that experimental observation of velocity overshoot requires an increase in the low-field mobility due to deliberate reduction in the channel doping (Shahidi *et al.*, 1988).

Electric field distribution in the channel

In the small size MOSFET devices, velocity of carriers near the drain is saturated. Calculation of the electric field distribution of short channel MOSFETs requires the simultaneous solution of current continuity equation and Poisson's equation. In order to derive the field distribution in the channel, the following four assumptions are introduced (Ko *et al.*, 1981): (1) carrier drift velocity is saturated in the drain depletion region, (2) current flow is confined to no deeper than the junction depth x_j, (3) the drain is assumed to be a perfect conductor, and (4) electric field toward substrate at the depth of x_j from the interface is negligibly small. The last assumption is reasonable since the field lines originating near the drain depletion region run more or less along the channel.

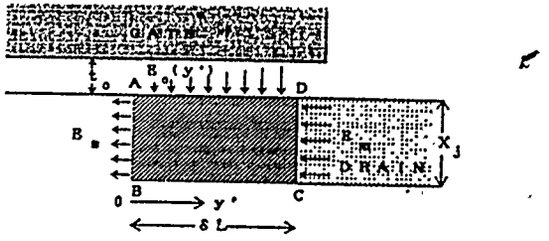


Fig. 8 Schematic diagram to illustrate the electric field distribution near the drain edge. Saturated carrier velocity is assumed in the hatched region.

The coordinate system in Fig.8 is chosen such that y' = 0 at point A where carrier velocity saturation occurs and y' = δL at the drain edge D. We apply Gauss' law to the box ABCD.

$$-E_o x_j + E(y')x_j + \frac{\varepsilon_o}{\varepsilon_s}\int_0^{y'} E_N(0,u)\,du = \frac{qN_A x_j y'}{\varepsilon_s} + \frac{qnx_j y'}{\varepsilon_s} \quad (38)$$

where E(y) and E_N are the electric field along and normal to the Si/SiO_2 interface, respectively. Differentiating (38) with respect to y', we have

$$\frac{dE(y)}{dy}x_j + \frac{\varepsilon_o E_N(0,y)}{\varepsilon_s} = \frac{qx_j}{\varepsilon_s}(N_A + n) \quad (39)$$

where n is the free carrier concentration. After some algebra, using the boundary condition, we obtain

$$\frac{dE(y)}{dy} = \frac{\phi(y) - V_{ds}}{\lambda^2}, \qquad \lambda^2 = \frac{\varepsilon_s x_j}{C_o} \quad (40)$$

λ is the characteristic length in the drain depletion region, which is independent of channel length and only depends on the junction depth and the gate oxide thickness. After integrating (40) along the channel, under the condition of E(0) = 0, we get

$$E(y) = E_o \cosh\left(\frac{y}{\lambda}\right) \quad (41)$$

The above equation indicates that the channel field grows almost exponentially toward the drain. The electric field has maximum value at the drain end of the channel as

$$E_m = E_o \cosh\left(\frac{\delta L}{\lambda}\right), \tag{42}$$

or

$$E_m = \sqrt{\frac{(V_d - V_{ds})^2}{\lambda^2} + E_o^2}\ , \tag{43}$$

where

$$\delta L = \lambda \sinh^{-1}\left(\frac{V_d - V_{ds}}{E_o L}\right). \tag{44}$$

Note that those analytical expressions are derived using quite simple assumptions. Two-dimensional simulation results and an empirical fit to the experimental data provide more accurate solution to the characteristic length (Chan *et al.*, 1985) as

$$\lambda = 0.22\, x_j^{1/2}\, t_o^{1/3}\ \text{cm}\ , \tag{45}$$

which is proved to be valid over wide ranges of x_j, t_o and N_A. It is found that from two-dimensional simulation and experimental results, N_A has little effect on λ.

Substrate current model

(1) Local electric field based model

Channel electrons in an n-channel MOSFET experience a very large electric field near the drain. The high field can cause impact ionization, and additional electrons and holes are generated by avalanche multiplication. The generated electrons are attracted to the drain, adding to the channel current, while holes are collected by the substrate contact, resulting in a substrate current.

The impact ionization coefficient, indicating the number of electron-hole pairs produced by one carrier per unit length, is expressed analytically as (Crowell and Sze, 1966)

$$\alpha(E) = A \exp\left(-\frac{B}{E}\right)\ , \tag{46}$$

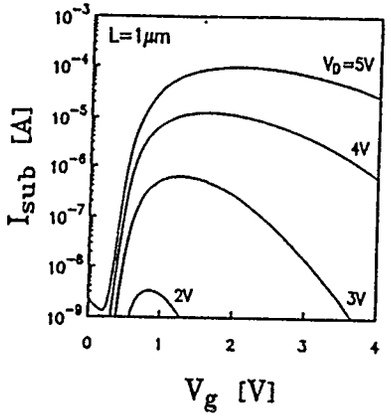
where A and B are constants. The generated current near the drain is calculated by integrating the product of the impact ionization coefficient and the drain current along the channel, as

$$I_{sub} = \int_0^{\delta L} I_d A \exp\left(-\frac{B}{E}\right) dy\ . \tag{47}$$

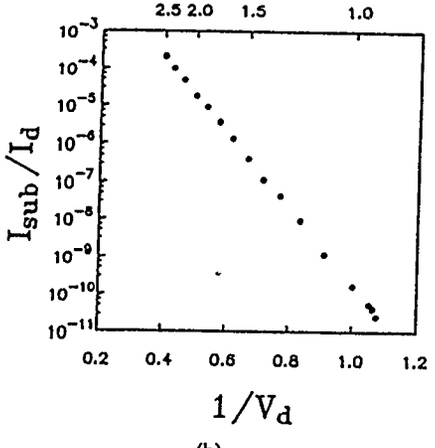
Replacing dy in (47) with $-E^2(dy/dE)d(1/E)$ and using $dE/dy \sim E/\lambda$ obtained from (41), (47) becomes

$$I_{sub} \approx \frac{A I_d (V_d - V_{ds})}{B} \exp\left(-\frac{B\lambda}{V_d - V_{ds}}\right). \tag{48}$$

Figure 9(a) shows a typical substrate current, which is plotted against gate voltage with drain voltage as a parameter. The initial rise is due to the increase of drain current and the eventual fall is due to the decrease of the channel electric field as expected from (48). For a given drain voltage, the maximum electric field decreases as V_g increases, but I_d increases monotonically. The main features are its bell shape and its maximum occurring at $V_g \sim V_d/2$. Figure 9(b) shows the ratio of substrate to drain current as a function of inverse drain current, indicating that the substrate current can be detected even at the drain voltage below silicon band gap energy (ε_g=1.1eV) and follows quite well with (48). This means that, even at low applied drain voltage, appreciable numbers of high energy electrons exist in the channel region and cause impact ionization. In other words, electron energy distribution in the channel has a high energy tail so that the high energy tail of the distribution function brings about impact ionization. These high energy electrons are considered to originate from either Auger process or electron-electron interaction.



(a)



(b)

Fig. 9 Substrate current characteristics as a function of (a) V_g with V_d as a parameter and (b) $1/V_d$ for devices with effective channel length of 1 μm.

(2) Non-local electric field based model

Analysis of substrate current in n-channel MOSFETs has been performed assuming a local field-dependent ionization coefficient given by (46). Since (46) was originally derived under a homogeneous electric field, the model is not applicable to the case of strongly inhomogeneous fields, where it has little physical justification. One approach to incorporate nonlocal hot-electron effects is to introduce corrections to well-known device equations. For example, Thornber (1982) gave a modified version of the drift-diffusion equation with a term proportional to the gradient of the electric field. A similar approach can be applied to the nonlocality of the ionization coefficient in MOSFETs. To include the nonlocal effect, Higman *et al.* (1988) investigate the dependence of the ionization coefficient on not only the local field but also on the derivatives of the field.

As is shown in (41), the electric field distribution near the drain is expressed by a hyperbolic cosine form, which is further approximated by an exponential field:

$$E(y) = E_N \exp(\frac{y}{\lambda}) \ . \qquad (49)$$

where λ is the length parameter. To investigate the dependence of the impact ionization coefficient on the length parameter, Higman *et al.* performed Monte Carlo simulations. They clearly demonstrated that no single set of parameters is valid for modeling impact ionization in the highly inhomogeneous fields in MOSFETs. They also showed that, in the practical range of electric field, an ionization coefficient of the form

$$\alpha(y,\lambda) = \gamma\lambda \exp\left(-\frac{B}{E(y)}\right) , \qquad (50)$$

is deduced, where γ is a constant parameter ($=3\times10^{10}cm^{-2}$) and the product $\gamma\lambda$ simply takes the place of A in (46). The product $\gamma\lambda$ approaches A when the length parameter becomes large. It should be emphasized that the parameters chosen for (50) are only valid for the fields represented by (49). The deduced impact ionization coefficient, expressed as a function of electric field, has been shown to depend linearly on the length parameter.

The substrate current can be calculated by integrating the space-dependent ionization coefficient over the high field region of the device. If (50) is used in place of (48), the expression for the substrate current is given by

$$I_{sub} \approx \frac{\gamma\lambda I_d(V_d - V_{ds})}{B} \exp\left(-\frac{B\lambda}{V_d - V_{ds}}\right) . \qquad (51)$$

The length parameter is linearly proportional to the substrate current so that it should be taken into account in the calculation of hot-electron effects, especially for ultra-small MOSFETs.

Gate Current

Gate current results from hot carriers, either channel electrons or avalanche-induced holes, that possess sufficient energy to surmount the $Si\text{-}SiO_2$ interface barrier. The typical characteristics of gate current are shown in Fig.10 for n-channel MOSFETs, with gate oxides below 20 nm, showing unusual structures which are not observed for thicker oxides.The gate current characteristics at low gate voltage is caused by the increase in the electric field near the drain, which promotes avalanche multiplication. Impact ionization supplies hot electrons and hot holes which are injected into the gate. These phenomena are called the *drain avalanche hot carrier injection* abbreviated to DAHC (Takeda *et al.*, 1983). At low gate voltage ($V_g < V_d/2$), low gate current due to hot hole injection is observed., which is quite small because of the higher injection barrier for holes (4.9 eV) than for electrons (3.1 eV). The electric field is, however, favourable for hole injection at low gate voltage, resulting in positive gate current. With increasing gate voltage, we observe a polarity change of the gate current followed by increase of the negative gate current. This is due to the fact that the opposing field for electrons in the oxide decreases with gate voltage and thus the effective barrier height for electron injection into the gate is decreased as (at the drain edge)

$$\phi_b = \phi_{bo} + V_d - V_g \ . \qquad 52)$$

The above equation is valid only for $V_g < V_d$, in which the gate current is found to be quite independent of V_d.

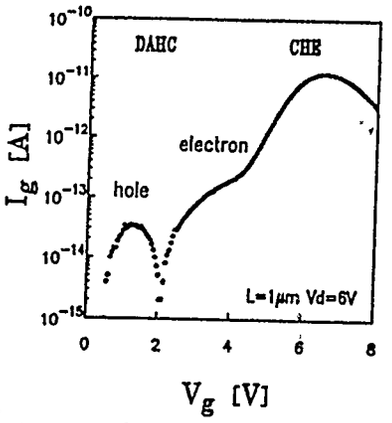


Fig. 10 Typical gate current characteristics in MOSFETs with gate oxide thickness of 11 nm and channel length of 1 μm. Solid circles indicate positive gate current, while solid circles negative current.

The other source of hot electrons in the gate is due to those electrons of the channel current flowing in the channel from source to drain, which gain enough energy from high fields existing near the drain junction, to overcome the interfacial barrier. This *channel hot electron* effect abbreviated to CHE is maximum when the MOSFET is operating at $V_d \sim V_g$, which is distinctly different from the peak of the substrate current. The CHE injected current is made up of "lucky electrons." Beyond the gate current peak ($V_g > V_d$), the MOSFET is driven toward the linear operating region and the maximum field in the channel decreases. This field decrease results in less energetic electrons, reducing the gate current in spite of a rise in channel current.

In the case of $V_g > V_d$, taking into account barrier lowering by image force and tunneling effects, the barrier height for electron injection is expressed as (Ning *et al.*, 1977)

$$\phi_b = \phi_{bo} - CE_N^{1/2} - DE_N^{2/3} \quad , \quad E_N > 0 \ , \qquad (53)$$

where C is calculated as $2.6\times10^{-4}(cm\text{-}V)^{1/2}$ and D is obtained by fitting experiment as 10^{-5} $(cm^2V)^{1/3}$.

Among several reported theoretical gate current models, a thermionic emission approach is the most easy to handle. The concept of the electron temperature can be applied to the thermoionic emission current: The higher electron temperature increases the number of electrons in the high-energy tail of the distribution. The gate current density is given by

$$J_g = \frac{q n_s k_B T_e}{2\pi m^*} \exp\left(-\frac{q\phi_b}{k_B T_e}\right), \qquad (54)$$

where n_s is the surface carrier density. The electron temperature T_e is a function of the field parallel to the current. For the high electric field, where velocity overshoot is achieved, T_e can be expressed from (12) as

$$T_e = T_o + \frac{2qv_s\tau_e E}{3k_B}, \qquad (55)$$

where v_s is the saturation velocity ($\sim 10^7$cm/s for $E > 10^5$ V/cm), τ_e the energy relaxation time. For the bias condition of $V_g < V_d$, carriers which have energies greater than the potential drop in the oxide, ($\sim V_d - V_g$), arrive at the gate electrode, and contribute to the gate current. The total gate current I_g is calculated by integrating (54) along the Si/SiO_2 interface.

EXPERIMENTS ON NON-EQUILIBRIUM CARRIER TRANSPORT

Evaluation of energy relaxation time

For the study of hot carrier effects in MOSFETs, such as substrate current, gate current, and velocity overshoot effect, an accurate electron temperature to describe electron energy distribution is required, which is directly related to the local electric field as shown in (55). In the following, we demonstrate a way to evaluate electron temperature in MOSFETs by means of optical method and we discuss the energy relaxation time extracted from the experimental data.

Since the emitted photons carry the physical information on hot carriers and are a useful probe for understanding the high-field effects in MOSFETs, photon emission spectra from n-channel MOSFETs have been investigated in detail (Lanzoni *et al.*, 1989; Herzog and Koch, 1988; Toriumi *et al.*, 1987). It is considered that the emitted photon energy is closely related to the hot electron energy.

(1) Experiment:

The devices used in this study were conventional poly-silicon-gate n-channel MOSFETs connected in parallel to obtain a high S/N ratio. The total channel width is 1.2 mm and the length is 2 μm. The gate oxide thickness is 15 nm. Figure 11 shows the experimental configuration for measuring the emitted photons from MOSFETs. The photons from the MOSFETs were collected through a lens and supplied to a monochromator. The monochromatic light was then converted to an electric signal by a photomultiplier. The signal, in a pulsed form, was fed into a lock-in amplifier.The drain was biased at a constant voltage, while the gate was biased with square wave pulses and the synchronizing reference pulses were supplied to the lock-in amplifier. Measured light intensity was automatically calibrated after measurement by a desk top computer to compensate the optical characteristics of the system.

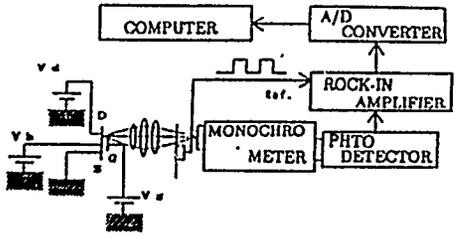


Fig. 11 Experimental configuration for measuring the energy spectrum of emitted photons from MOSFETs

(2)Results and discussion:

Figure 12 shows the photo-emission spectra as a function of photon energy, which decreases exponentially with photon energy. The number of photons per unit energy I(v) can be expressed as:

$$I(v) = C_1 \exp(-C_2 hv), \qquad (56)$$

where C_1 and C_2 are constants. Although no direct evidence to clarify the emission mechanism has been reported, the bremsstrhlung is at present the most probable photon emission in Si MOSFETs. The number of emitted photons due to the bremsstrahlung can be expressed as

$$N(v) = \int_{hv}^{\infty} P_{hv}(\varepsilon)\, f(\varepsilon)\, d\varepsilon, \qquad (57)$$

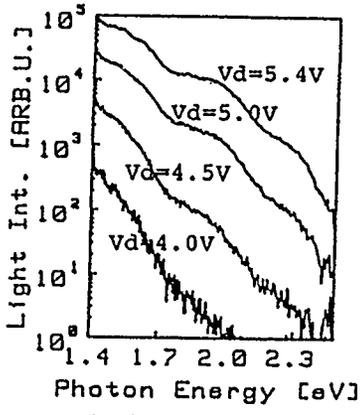


Fig. 12 Energy spectrum of emitted photons from a MOSFET at V_g=2 V with V_d as a parameter. Slight oscillations in the spectrum are caused by the interference of emitted light in the poly-silicon gate.

where $P_{hv}(\varepsilon)$ represents the emission probability that an electron with energy ε emits a photon with energy $h\nu$ and $f(\varepsilon)$ is the energy distribution of the hot electrons. The lowest energy of the integral in (57) is $h\nu$ because the electron energy must always be higher than the emitted photon energy. Here, since $P_{hv}(\varepsilon)$ would consist of the transition probability in the momentum space and the density of states, it would not vary as much as the distribution function $f(\varepsilon)$. Therefore, when the photon energy is high enough, $P_{hv}(\varepsilon)$ can be assumed to be constant as

$$N(\nu) \approx P_{h\nu}^{0} \int_{h\nu}^{\infty} f(\varepsilon)\, d\varepsilon \ . \tag{58}$$

Comparing (56) and (58), it can be concluded that the energy distribution function $f(\varepsilon)$ is expressed as

$$f(\varepsilon) \approx C_3 \exp(-C_4\varepsilon) \ . \tag{59}$$

Since the energy distribution of hot electrons in an n-channel MOSFET is expressed by a Maxwell-Boltzmann distribution in the first order of approximation, (59) can be deduced to be

$$f(\varepsilon) = C_3 \exp\left(-\frac{\varepsilon}{k_B T_e}\right), \tag{60}$$

which means that the electron temperature can be determined from the numerical constant C_2, which corresponds to the slope of the line in Fig.12.

According to the above interpretation on the emitted photon spectra, both the drain bias and the gate bias dependences of the slope of the spectra can be understood as a change in the electron temperature. Figure 13 shows the electron temperature as a function of maximum electric field calculated by using (43), which increases almost linearly with the electric field. The maximum electron temperature was evaluated with (43). The electron temperature decreases monotonically with the gate bias at a given drain voltage. Considering the energy balance equation, (12), the following equation is obtained:

$$\frac{3}{2}k_B T_e = \frac{3}{2}k_B T_o + q v \tau_e E_{max} \ . \tag{61}$$

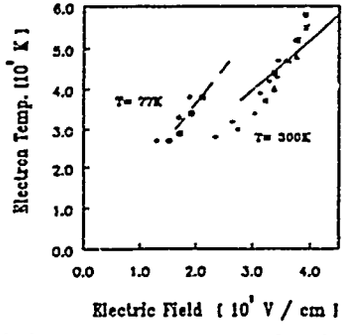


Fig. 13 Measured electron temperature as a function of the maximum electric field in an MOSFET for 77 K and room temperature.

Since the electron velocity saturates and has a constant v_s under the electric field of interest, (61) can be described as

$$T_e = T_o + 0.77\times10^{11}\, \tau_e E_{max} \ , \tag{62}$$

where we assume $v_s = 1\times10^7$ cm/s. The linear dependence in Fig.13 can be explained well by (62), when a constant energy relaxation time is assumed. This fact shows that the energy relaxation time can be determined experimentally by the slope of the line in Fig.12. In our experiment, the value of 1.4×10^{-13} sec (at room temperature) is obtained.

It should be noted that the measured energy relaxation time thus obtained is much smaller than that obtained by Monte Carlo method. The energy relaxation time obtained from other experimental methods is also quite small: $\tau_e=6\times10^{-14}$ sec. (Toriumi *et al.*, 1987), $\tau_e=1.0\times10^{-13}$ sec. (Takeda *et al.*, 1982). The difference between theoretical and experimental values is mainly due to the fact that Monte Carlo calculations reported so far neglect several factors that affect the energy relaxation time, such as the contribution of L valley and other higher lying energy band structures to the carrier scattering.

The reason why the energy relaxation time is independent of carrier energy can be roughly explained as follows. The momentum relaxation time τ_m is strongly dependent on carrier energy, which is a time of the order of mean free time between collisions. On the other hand, the energy loss after one collision is independent of the carrier energy, since it is determined by the phonon energy involved in the scattering. Thus, more collisions are needed to relax higher carrier energy, so that τ_e is not as dependent on the carrier energy as on τ_m.

Hot carrier induced degradation of MOSFET

Interface states and trapped oxide charge induced by the high electric field are of special interest due to their role in the degradation of the MOSFETs and especially their probable role in the breakdown of the insulator. It is known that induced charges have serious deleterious effects on threshold voltage, subthreshold slope, and surface carrier mobility. These effects are more important in scaled devices as the electric field does not scale with device length, and device operation often takes place close to the surface. The problems associated with hot carriers in MOSFETs have been recognized as one of the major constraints in device scaling. Although it is accepted that the hot carriers generated in the high drain field cause the most severe degradation of MOSFETs, the mechanism for the degradation is still a controversial issue. Recent studies (Heremans *et al.*, 1988; Hu *et al.*, 1985) indicate that the generation of interface states is the dominant cause for the degradation of n-channel MOSFETs.

The degradation of MOSFETs caused by hot carriers is considered to originate from the following two mechanisms: (1)Direct interactions of channel hot carriers (electrons or holes) with interfacial region and (2) interaction during passage of carriers through the gate oxide. Mechanism (1) may be correlated with the substrate current, and mechanism (2) with the gate current. Although a large number of degradation mechanisms have been reported so far, there are still no well accepted degradation models reported. Therefore, in this section, we focus our attention to the interface state generation induced by hot electrons from an experimental view point. We try to shed light on one of the physical mechanisms of hot-carrier induced interface state generation, using uniform hot electron injection from the substrate into the channel (Verwey, 1973).

The samples used are polysilicon gate n-channel MOSFETs fabricated in a p-well with 3 μm depth and surface acceptor concentration of $2\times10^{16}/cm^3$. The resistivity of the n-type

substrate is 1–2 Ωcm. The gate oxide is 20 nm thick. The transistors with W = 10 μm and L = 0.9 ~ 10 μm are connected together in parallel. The total channel area is $2.5 \times 10^{-6} cm^2$.

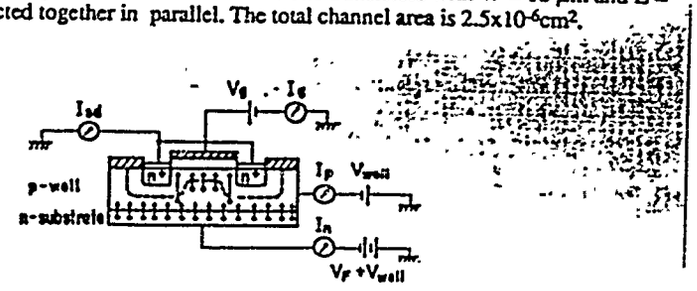


Fig.14 Cross-section of a sample for the hot-electron injection measurement.

The electron injection is carried out by applying a forward bias between the p-well and n-substrate. During the injection, a reverse bias is applied to the p-well with respect to the source and drain. The electrons diffusing into the p-well from the substrate are accelerated by the electric field in the depletion region. Some of them with high energy at the interface are injected into the gate oxide. The gate voltage applied during the electron injection is varied in the range from 2.5 to 14 V while keeping the p-well voltage at -4 V. The injected electron density (N_{inj}) is evaluated from the total gate current density, using the following equation

$$N_{inj} = \frac{1}{qA_c} \int I_g(t)\, dt \ ,$$

where I_g is the gate current and A_c is the total channel area.

We use the charge-pumping technique to measure the interface-state density because of its high accuracy and insensitivity to the fixed-oxide charge. The charge-pumping current is measured intermittently during the electron injection. The generated interface-state density is expressed as

$$\Delta N_{it} = \frac{\Delta I_{cp}}{qA_c f}$$

where f is the frequency of the gate pulse, and I_{cp} is the increased charge-pumping current after the carrier injection.

Figure 15 shows the generated interface-state density (ΔN_{it}) versus the injected electron density (N_{inj}). The generated interface-state density at a small injected electron density ($N_{inj} < 10^{17} cm^{-2}$) is expressed as

$$\Delta N_{it} = A N_{inj}^{n}$$

where A is a constant, and n is the slope of the interface-state generation. Figure 16 shows the slope n as a function of the gate voltage applied during the carrier injection. The slope becomes unity at a high gate voltage, while it is reduced to a half at a low voltage. Between the two extreme values, a clear transition of the slope is observed around $V_g = 7$ V, indicating the existence of two types of interface-state generation mechanisms: type I (n = 1.0) and type II (n = 0.5).

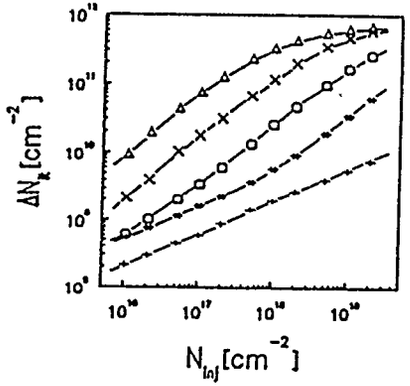


Fig. 15 Generated interface-state density (ΔN_{it}) versus injected electron density (N_{inj}). The reverse bias applied to the p-well is -4 V. The gate voltages during the electron injection are: 2.5V (+), 6V (*), 8V (X) and 14V (D).

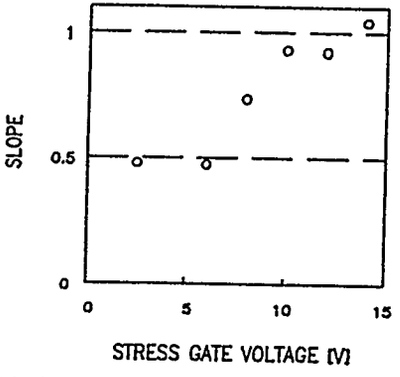


Fig. 16 Slope of the interface-state generation (n) versus the stressing gate voltage (V_g). The slope n is evaluated from the expression $\Delta N_{it} = A N_{inj}^{n}$ ($N_{inj} < 10^{17} cm^{-2}$).

Type I interface states are generated under a high oxide field. This suggests that the hot electrons in the oxide are responsible for the generation of type I interface states. To confirm this hypothesis, we have carried out Fowler-Nordheim tunneling injection in which electrons have a small energy near the interface. The interface-state generation in F-N tunneling injection is characterized as follows: (1) The generated interface-state density is linearly proportional to the injected electron density, and (2) mid-gap states are generated. These experimental results indicate that the interface states generated in the hot-electron injection and F-N tunneling injection are quite similar by nature, in spite of the significant difference in electron energy near the interface. Hence, it is the carrier heating in the oxide that causes the type I interface states. It is considered that the type I interface states originate from the capture of holes at the interface, because the interface-state generation during F-N injection are considered to be caused by holes (Muto *et al.* 1989; Toriumi *et al.*, 1988) which are generated either by impact ionization in the oxide or by the decay of surface plasmons at the gate/oxide interface (Fischetti, 1985).

The generation rate of the type II interface states (n=0.5) is well explained by the hydrogen release model (Hu *et al.*, 1985). Interface states are created by breaking chemical Si–H bonds. Hence we conclude that the balance between the diffusion and recombination of species such as hydrogen (Jeppson and Svensson, 1977) control the generation rate of the type II interface states.

MODELING OF NON-EQUILIBRIUM CARRIER TRANSPORT

Impact Ionization Coefficient

In the "lucky-drift" theory, the momentum relaxation time τ_m is separated from the energy relaxation time τ_e and it is assumed that $\tau_m << \tau_e$. Therefore, the electron will gain energy as it drifts in the field before it is scattered energetically and may occasionally pick up sufficient energy to make an impact ionization possible. Impact ionization is a three-particle process, where a high-energy electron of the conduction band generates a electron-hole pair. The minimum energy for impact ionization is known as the threshold energy and initially "hard" thresholds were used, in that the probability of impact ionization was taken as zero below threshold and unity above it. The value of threshold is nominally 3/2 times the band gap for the simple case where all the particles concerned have equal masses. In practice, however, it is found that for $\varepsilon > \varepsilon_T$, the probability is not unity but is a function of ε and ε_T; this is refereed to as the "soft" threshold case. The most frequently used "soft" impact ionization model is a simple Keldysh formula, which provides the rate for impact ionization for an electron of energy ε:

$$P_{ii}(\varepsilon) = P_o \frac{(\varepsilon - \varepsilon_T)^2}{\tau_{ph}(\varepsilon_T)} , \qquad (63)$$

where ε_T is the threshold energy for impact ionization, $1/\tau_{ph}(\varepsilon_T)$ is the total phonon scattering rate at the ionization threshold energy and P_o is a constant. Note that this is a very simple formula, inconsistent with the band structure of silicon because direct- and isotropic- parabolic band structure are assumed. To calculate impact ionization coefficient in silicon precisely, requires a detailed knowledge of the band structure.

In the "lucky drift" model, the momentum relaxation time is considered to be much shorter than the energy relaxation time so that there are many momentum relaxing collisions for each energy relaxation. Thus the direction of the momentum is continually changing and it would seem appropriate that with this randomization, the impact ionization rate should be averaged over all directions, meaning the isotropic threshold energy. Under this condition, the impact ionization rate is calculated by taking into account the effect of anisotropic parabolic bands. The transition probability per unit time that an electron of wavevector k_1 will impact ionize as in figure 14 is written as

$$P(k_1) = C \int |M|^2 \Theta \, \delta(\varepsilon_1 + \varepsilon_2 - \varepsilon_{1'} - \varepsilon_{2'}) \, dk_{1'} dk_{2'} , \qquad (64)$$

where C is composed of fundamental constants, M is the transition matrix element, Θ is the probability of vacancies in state 1' and 2' and occupancy in state 2 involved in the transition.

By taking into account the anisotropic and nonparabolic conduction band structure of silicon, the direction-averaged impact ionization probability rate near threshold is found to be approximately given by (Sonoda *et al.*, 1989)

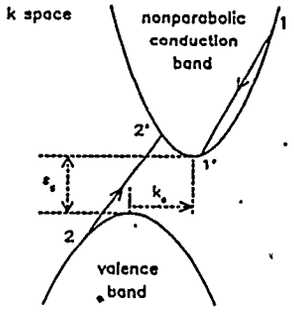


Fig. 17 Schematic diagram of energy band structure for indirect transition semiconductor together with carrier transitions involved in impact ionization.

$$P_{ii}(\varepsilon) = P_o \frac{(\varepsilon - \varepsilon_T)^{2.5}}{\tau_{ph}(\varepsilon_T)} . \qquad (65)$$

Note that the power of the impact ionization rate is 2.5, while that of Keldysh model is only 2.0. This gives a higher ionization rate than the Keldysh model for the higher energy region. The present model also gives an additional "softness" around the threshold energy, which is caused by the indirect and anisotropic band structure. The validity of the new impact ionization model is verified by comparing Monte Carlo simulation using the impact ionization rate described above and experimental data. There exist two independent experimental data sets for the verification; (1) quantum yield and (2) impact ionization coefficient.

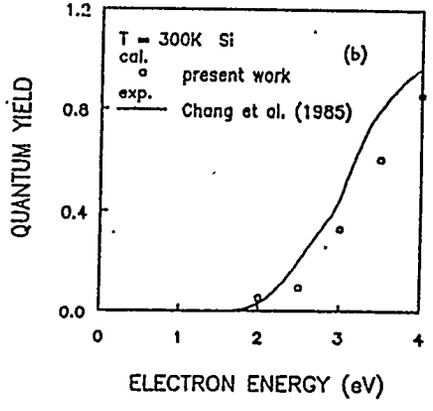


Fig. 18 Quantum yield of electron as a function of electron energy: experimental (Chang *et al.*, 1985: solid line) and simulational (open squares) results calculated by Monte Carlo method

Using p-channel Si-gate MOSFETs of very thin gate oxide, Chan *et al.*, (1985) measured quantum yield of electron impact ionization, which is the number of generated electron-hole pairs per incident electron injected into silicon from the polysilicon gate by

tunneling. The measured quantum yield (solid line) is given in Fig.18 together with our calculated result. Good agreement between them is observed all the way up to 4 eV. In this calculation, we use a parameter set of $\varepsilon_T = 1.7$ eV and $P = 9\times10^{12}$ s^{-1}.

Using the same physical parameter set describe above, we calculated impact ionization coefficient of silicon. In Fig. 19, we show the calculated electron impact ionization coefficient together with other experimental data which are plotted against the inverse field strength. The calculated values are in good agreement with the experimental data. The agreement for two independent experimental data give strong support to the validity of our model given in (65).

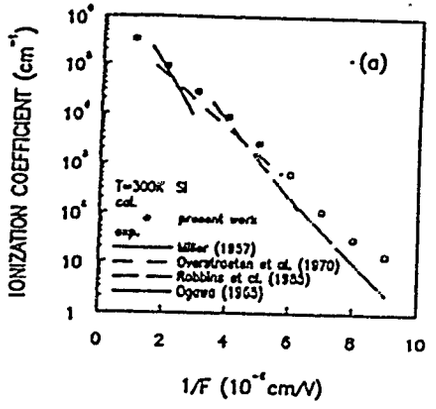


Fig. 19 Impact ionization coefficient as a function of electric field at room temperature: experimental results and calculated results (open squares).

Physical model for deep submicron device simulation

The majority of submicron MOSFETs are subject to regions of high electric field which gives rise to non-equilibrium transport. Traditional device simulators based on drift-diffusion equations are not capable of modeling accurately future deep submicron MOSFETs. One approach to overcome this difficulty is to use an advanced energy relaxation model which includes momentum and energy relaxation effects of carriers. However, numerical calculation of the energy relaxation model under a two-carrier condition requires an excessive amount of computational time and limits the usefulness of relaxation model from a software engineering point of view. The other approach is to use the extended drift-diffusion model (Price, 1988). The model parameters, however, are not well investigated so that one has to obtain accurate parameters before using the extended model for device simulation.

It has been recognized that the carrier velocity is not well expressed as a function of the local electric field in the presence of a large electric field gradient. In order to include non-local effects in the conventional drift-diffusion model, Artaki (1988) and Price (1988) added a term proportional to the field gradient with a phenomenological length constant $L_e(E)$, which leads to

$$v_d = v_{do}(E)\left[1 + \frac{L_e(E)}{E}\frac{dE}{dy} - \frac{D(E)}{n}\frac{dn}{dy}\right], \qquad (66)$$

where v_{do} is the drift velocity corresponding to the homogeneous field E. Artaki (1988) estimated the length constant by means of Monte Carlo simulation using a simplified band structure: a single spherical and parabolic conduction band. The Monte Carlo simulation showed that $L_e(E)$ becomes significant only for fields above 30 kV/cm and flattens out with increasing field at about 40 nm. In the field region below 30 kV/cm, $L_e(E)$ is negligible and the data points scattered in a wide range because of stochastic noises due to the Monte Carlo simulation. For that reason, the data at low field are highly qualitative. We investigated the length coefficient more precisely by using the iterative method which provides quite smooth carrier velocity vs. space curve without stochastic noise. Figure 20 shows the length coefficient, $L_e(E)$, evaluated using the iterative method. In the low electric field region, the calculated $L_e(E)$ linearly increases with field, while it saturates to the value of 40 nm at 30 kV/cm. In order to clarify the physical meaning of $L_e(E)$, the length coefficient is estimated by a different approach using analytical formulas. In the saturation velocity regime, we can assume constant carrier density so that the equation (11) is expressed as

$$v_d = \frac{\tau_m}{m^*}\left[-qE - k_B\nabla T_e\right] = v_{do}\left(1 + \frac{k_B|\nabla T_e|}{qE}\right). \qquad (67)$$

From (17), we obtain

$$\nabla T_e = \frac{2q\tau_e v_s}{3k_B}\nabla E\ . \qquad (68)$$

Substituting (68) into (67), and comparing with (66), the length coefficient is found to be represented as

$$L_e(E) = \frac{2\tau_e v_{ds}}{3}\ . \qquad (69)$$

Therefore, it is reasonable to define the length coefficient as the distance where electrons move without losing their energy in high electric field, which is independent of the applied electric field.

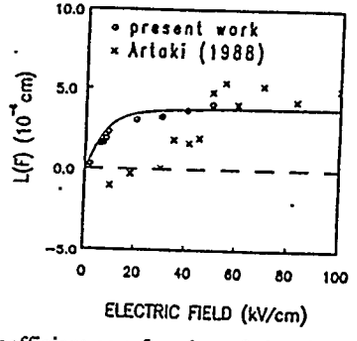


Fig. 20 Length coefficient as a function of electric field calculated from iterative method (o) and Monte Carlo (x) (Artaki, 1988).

Carrier transport characteristics of submicron devices are investigated from the practical point of view. Figure 21 shows the velocity profile in the MOSFET along the channel. Equation (66) indicates that, in a field increasing along the current flow, the carrier velocity exceeds v_S, meaning the velocity overshoot, while in a field decreasing along current flow, (66) predicts the velocity undershoot as shown in the figure. This fact means that the advantage of non-equilibrium carrier transport cannot be easily taken of the velocity overshoot because both accelerating and decelerating fields coexist in the actual devices. Note that the benefit of velocity overshoot in the channel of a MOSFET cannot be realized until the carriers are overshooting very near their injection point from the source into the channel. For that reason,

the non-equilibrium transport effect should be taken into account for deep submicron device simulation by adopting the extended drift-diffusion model together with accurate model parameters.

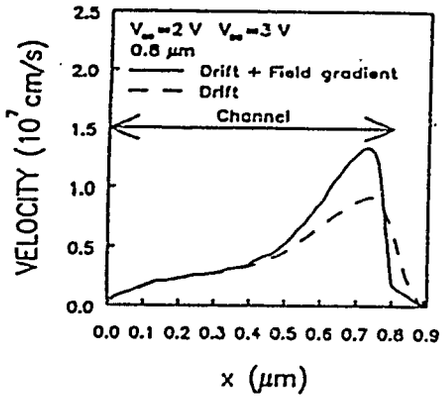


Fig. 21 Electron velocity parallel to the interface at 300 K in Si-MOSFET

ACKNOWLEDGEMENTS

The authors would like to express his appreciation to N. Yasuda, K. Sonoda, M. Sekido and K. Kitatani for many helpful discussions and comments. The authors also express their sincere gratitude to VLSI Research Center of Toshiba Co. and LSI Research & Development Laboratory of Mitsubishi Electric Co. for supplying the samples.

REFERENCES

Artaki, M., 1988, Hot-electron in an inhomogeneous field, Appl. Phys. Lett., 52:141.

Baccarani, G., and Wordeman, M. R., 1985, An investigation of steady-state velocity overshoot in silicon, Solid State Electron, 28:407.

Beattie, A.R., 1988, Impact ionization rate and soft energy thresholds for anisotropic parabolic band structure, Semicond. Sci. Technol., 3:48

Caughey, D, M, and Thomas, R. E., 1967, Carrier mobilities in silicon empirically related to doping and field, Proc. IEEE, 55:2192.

Chan, C., Hu, C., and Brodersen, R. W., 1985, Quantum yield of electron impact ionization in silicon, J. Appl. Phys., 57:302.

Chan, T. Y., Ko, P. K., and Hu, C., 1985, Dependence of channel electric field on device scaling, IEEE Electron Device Lett., 6:551.

Crowell, C. R., and Sze, S. M., 1966, Temperature dependence of avalanche multiplication in semiconductors, Appl. Phys. Lett., 9:242.

Ferry, D. K., 1985, in "Gallium Arsenide Technology," Edited by D. K. Ferry, Howard W Sams & Co., Inc., Chapter 12.

Heremans, G., Bellens, R., Groeseneken, G., and Maes, H. E., 1988, Consistent model for the hot-carrier degradation in n-channel and p-channel MOSFET's, IEEE Trans. Electron Devices, 35:2194.

Herzog, M., and Koch, F., 1988, Hot-carrier light emission from silicon metal-oxide-semiconductor devices, Appl. Phys. Lett., 53:2620.

Higman, J. M., Kizilyalli, I. C., and Hess, K., 1988, Nonlocality of the electron ionization coefficient in n-MOSFET's: an analytic approach, IEEE Electron Device Lett., 9:399.

Hu, C., Tam, S. C., Hsu, F. C., Ko, P. K., Chan, T. Y., and Terrill, K. W., 1985, Hot electron-induced MOSFET degradation - Model, monitor and improvement, IEEE Trans. Electron Devices, 32:375.

Jacoboni, C., Canali, C., Ottavaniani,G., and Quaranta, A. A.,1977, A review of some charge transport properties of silicon, Solid State Electron., 20:77.

Jeppson, K. O., and Svensson, C. M., 1977, Negative bias stress of MOS devices at high electric fields and degradation of MNOS devices, J. Appl. Phys., 48:2004.

Keldysh, L. V., 1965, Sov. Phys. JETP, 21:1135.

Ko, P. K., Muller, R. S., and Hu, C., 1981, A unified model for hot-electron currents in MOSFETs, in "Tech. Dig. Int. Electron Devices Meet.", 81:600.

Ko, P. K., 1989, Approaches to Scaling, in "VLSI Electronics Microstructure Science: Vol.18, Advanced MOS Device Physics", Ed. by Einspruch, N. G., and Gildenblat, G.S., Academic Press, Inc.

Lanzoni, M., Manfredi, M., Selmi, L., Sangiorgi, E., Capelletti, R., and Ricco, B., 1989, Hot-Electron-Induced Photon energies in n-channel MOSFET's operating at 77 and 300 K, IEEE Electron Device Lett., 10:173.

Miller, S. L. , 1957, Ionization rates for holes and electrons in silicon, Phys. Rev., 105:1246.

Muto, H., Fujii, H., Nakanishi, K., Shibuya, Y., J. Mitsuhashi and T. Matsukawa, 1989, The effects of oxide thickness and temperature on the generation of silicon dioxide/silicon interface states at high electric fields, Trans. IEICE, J72-C-II:174.

Ning, T. H., Osburn, C. M., and Yu, H. N., 1977, Emission probability of hot electrons from silicon into silicon dioxide, J. Appl. Phys., 48:286.

Ogawa, T., 1965, Avalanche breakdown and multiplication in silicon pin junctions, Jpn. J. Appl. Phys., 4:473.

Rees, H. D., 1969, Calculation of distribution functions by exploiting the stability of the steady state, J. Phys. Chem. Solids, 30:643.

Price, P. J., 1988, On the flow equation in device simulation, J. Appl. Phys., 63:4718.

Robbins, V. M., Wang, T., Brennan, K. F., Hess, K., and Stillman, G. E., 1985, Electron and hole impact ionization coefficients in (100) and in (111) silicon, J. Appl. Phys., 58:4614.

Rosencher, E., 1981, Ballistic transport in semiconductors: a displaced Maxwellian formulation, J. de. Physique, 42:C7-351.

Sai-Halasz, G., Wordeman, M. R., Rishton, S., Ganin, E., and Kern, D. P., 1988, High transconductance and velocity overshoot in NMOS devices at the 0.1 μm gate length level, IEEE Electron Device Lett., 9:464.

Shahidi, G. G., Antoniadis, D. A., and Smith, H. I., 1988, Electron velocity overshoot at room and liquid nitrogen temperatures in silicon inversion layer, IEEE Electron Device Lett., 9:94.

Snowen, C. M., 1986, "Introduction to Semiconductor Device Modelling," World Science Publishing Co. Ltd, Chapter 2.

Sonoda, K., Inoue, Y., Taniguchi, K., and Hamaguchi, C., 1989, Physical model for deep submicron device simulation, in "Ext. Abstract of 21th Conf. on Solid State Devices and Materials".

Takeda, E., Kume, H., Toyabe, T., and Asai, S., 1982, Submicronmeter MOSFET structure for minimizing hot-carrier generation, IEEE Trans. Electron Devices, 29:611.

Thornber, K. K., 1982, Current equations for velocity overshoot, IEEE Electron Device Lett, 3:69.

Toriumi, A., Yoshimi, M., Iwase, M., Akiyama, Y., and Taniguchi, K., 1987, A study of Photon emission from n-channel MOSFET's, IEEE Trans. Electron Devices, 34:1501.

van Overstraeten, R., and de Man, H., 1970, Measurement of the ionization rates in diffused silicon p-n junctions, Solid State Electron., 13:583.

Verwey, J. F., 1973, Nonavalanche injection of hot carriers into SiO_2, J. Appl. Phys., 44:2681.

23

SOME CONSIDERATIONS RELATED TO THE QUANTIZATION OF CHARGE IN MESOSCOPIC SYSTEMS

A. J. Leggett

Department of Physics
University of Illinois
Urbana, Illinois 61801

INTRODUCTION

Over the last few years there have been a large number of theoretical and experimental papers concerned with the consequences, in various types of mesoscopic systems, of the fact that electric charge is quantized in units of e. In this lecture I do not intend to go into the details of the calculations of these effects—something for which there are several people at this School much better qualified than me. Rather, I would like to take a hard look at some of the fundamental conceptual aspects of this type of problem, and in particular at the question: To what extent do the familiar concepts of macroscopic electrical engineering such as voltage, current, capacitance etc. have a meaning when we describe our system in quantum mechanical terms? In particular, just how far do we have to take our book-keeping as regards, say, the actual nature of a "current source" we intend to use to drive our quantum mechanical system? To many working in this area, perhaps these questions seem trivial and the answers to them obvious. As far as I am concerned, they are far from trivial, and while a sustained attempt to answer them may not alter the results of existing calculations, or at least not qualitatively, I believe it will give us significant understanding that may be useful elsewhere in mesoscopic physics.

ELECTRIC CHARGE

Let's start with the most basic and familiar notion of all, that of electric charge. We all know that electric charge comes quantized in units of the electron charge e (at least if we do not go down to the quark level!); what could possibly be mysterious about that? Perhaps nothing, but there is one point which needs some emphasis: let us consider a large volume which we know to contain a certain number of electrons, and cut out in our minds an arbitrary but well-defined closed surface within it. We then ask: Is the amount of charge contained within this surface constrained to be a multiple of the electron charge e? This question needs some explication. What we mean is: If we were to somehow measure the total charge within the surface, are we guaranteed to get the results Ne, where N is some integer? Needless to say, even if the answer is yes, this is perfectly compatible with the fact that the time average of the charge, or even its expectation value in a given quantum-mechanical state, may be quite arbitrary. In fact of course the answer *is* yes, both as seen in experiment (e.g. in electron diffraction experiments where discrete charges are counted, and from fundamental principles of quantum field theory. As to the latter, we have the fundamental commutation relation (for fermions: I ignore spin for present purposes).

$$\{\Psi(x), \Psi^+(x')\} = \delta(x - x') . \tag{1}$$

Now, let us introduce a complete orthonormal set of functions $\varphi_n(x)$ which cover the volume enclosed within the surface (with no restrictions on their behavior at the surface itself); then we can write

$$\Psi(x) \equiv \sum_n a_n \varphi_n(x) , \qquad \Psi^+(x) \equiv \sum_n a_n^+ \varphi_n(x) , \tag{2}$$

with

$$\{a_n, a_{n'}^+\} = \delta_{nn'} , \tag{3}$$

and the total charge contained within the surface is

$$Q = e \int \Psi^+(x) \Psi(x) \, d^3x = e \sum_n a_n^+ a_n . \tag{4}$$

But from the commutation relation (3) it may be shown in the standard way that the only eigenvalues of the operator $a^+_n a_n$ are 0 and 1, from which it immediately follows that the only possible eigenvalues of Q are Ne, N=0,1,2... This point may seem completely trivial and obvious, but it is as well to state it explicitly. Note in particular that the moment we "smear" Q with any nonconstant function, i.e. define

$$Q_s \equiv e \int \Psi^+(x) \Psi(x) f(x) \, d^3x , \quad \int f(x) \, d^3x = 1 , \; f(x) \neq \text{const.} \tag{5}$$

the result breaks down: Q_S can have any values whatever.

POTENTIALS

Next consider the concept of voltage. We have to distinguish between the concept of *electrostatic potential*, which is a quantum-mechanical operator, and *electrochemical potential*, which is an essentially thermodynamic concept. The definition of electrostatic potential (call it $V_{e.s.}$) is straightforward:

$$\hat{V}_{e.s.}(r) \equiv \int \frac{\hat{\rho}(r')}{4\pi\varepsilon_c |r - r'|} \, d^3r' , \tag{6}$$

where $\hat{\rho}(r')$ is the total charge density operator. It is important to note that in the general case the eigenvalues of $V_{e.s.}(r)$ for given r need not form a continuous, or even approximately continuous, spectrum. For example, in the situation shown in Fig. 1, it is clear that the eigenvalues of the electrostatic potential at point r must occupy bands between $Ne^2/(r+a)$ and $Ne^2/(r-a)$; for small enough N this leaves gaps with no eigenvalues. However, in most situations of practical interest, and in particular for most of the ways in which in practice one applies an "external voltage" to one's sample, whether a.c. or d.c., the charge density in question corresponds to a large enough assembly of charges, distributed over a wide enough spatial region, that it can be treated as a continuous variable. This does not, of course, necessarily mean that we can immediately treat it as a *classical* variable: e.g. the discussion below.

It should be remarked that the operation of applying an electrostatic potential difference is actually a somewhat awkward way of inducing a current to flow, and that whenever we have a closed circuit it is usually much more convenient to induce an e.m.f. around it by Faraday's

law, that is by applying a time-varying magnetic flux through the circuit. The advantage of this is that there is a conceptually clear-cut way of describing the effect quantum-mechanically, namely the standard replacement $p_i \to p_i - e A(r_i : t)$ where $A(r_i : t)$ is the vector potential [which of course must for consistency include induced as well as external terms: cf. e.g. Bloch (1970)]. The great convenience of this is that $A_{ext}(r,t)$, which may be derived e.g. from the motion of a large permanent-iron magnet, may be treated as a c-number without inducing any conceptual complications.

The electrochemical potential, for a bulk system, is usually defined as $(\partial E/\partial N)_S$, where E is the mean energy of the system and N the number of particles (electrons) in it. For a finite system at zero temperature it seems most natural to define it as the (mean) energy necessary to add one more electron to the system; that is

$$\mu(N) \equiv E_0(N+1) - E_0(N) , \tag{7}$$

where $E_0(N)$ is the groundstate energy of the system with N electrons. At finite temperature the definition is a bit more ambiguous. Depending on the physics of the situation, it may be more appropriate to use the macrocanonical or the grand canonical ensemble. In a situation where energy exchange of the system with its environment is efficient but the transfer of particles is slow (e.g. a grain separated from bulk metal by a thick oxide barrier but exchanging heat via phonons, etc.) the former description is the more natural, and then it seems most natural to define $\mu(N)$ as

$$\mu(N;T) \equiv F(N+1,T) - F(N,T) , \; F(N) \equiv -kT \ln \sum_n \exp[-\beta E_n(N)] , \tag{8}$$

where $E_n(N)$ denotes an energy level of the N-particle system. By the usual arguments, $\mu(N)$ so defined can also be written $E(N+1,S) - E(N,S)$ with S the entropy of the system. On the other hand, if the exchange of particles as well as of energy with the environment is efficient on the time scale of interest, then the natural definition is from the standard Gibbs formula for the occupation probabilities, $P(N,n) = \text{const.} \cdot \exp\{-\beta [E_n(N) - \mu N]\}$.

The electrochemical potential, as its name implies, results from both chemical (or Pauli-principle) and electrostatic effects. Indeed, under certain circumstances it may be a matter of taste whether we call a given part of the interaction "chemical" or "electrostatic" (see below). We will generally count as electrostatic those effects which result from "distant" charges. Of course, if nothing else is altered then a *change* in electrostatic potential results in an identical change in the electrochemical potential of the system.

As is well known, the electrochemical potential governs the rate of evolution of the relative phase of the many-body wave functions corresponding to N and N+1 particles. By this I mean the following: Imagine that we have a small metallic grain which is separated by a thick oxide barrier from bulk metal, and that at zero temperature there is a finite but small matrix element for tunnelling of an electron through the barrier:

$$\hat{T} = \sum_{mn} T_{mn} a^+_m b_n + \text{c.c.} , \tag{9}$$

where a^+_m creates energy eigenstate m within the grain and b^+_n state n in the bulk metal. (Strictly speaking, T_{mn} and a^+_m should have a label N as well, since the labelling of the states may be different for different N, but we neglect this for notational simplicity). If we start from a state with definite numbers of electrons on the grain (N) and in bulk (N_b) and treat $\hat{T}$ as a perturbation, then after a time short compared to $\hbar/|T|$ (where $|T|$ is a typical value of T_{mn}) the many-body wave function will be a linear superposition of the original state (N,N_b) and various states (N+1, m: N_b-1, n) and (N-1, m: N_b+1, n). If for the moment we arbitrarily assume that the states m and n, as well as the initial state, both represent the groundstate for the number of particles in question, then it is clear that the energy difference E(N+1, m; N_b-1, n) - E(N, N_b) is $\mu(N)$ and the difference E(N, N_b) - E(N-1, m: N_b+1, n) is $\mu(N-1)$, and thus the relative phases evolve as $\mu(N)t$ and $\mu(N-1)t$ respectively; note that in general the two rates are not equal (although in some simple cases of subsequent interest they may be taken so). If m and n are not the groundstates, then, of course, in general the phases evolve at a different rate. The case of tunnelling of Cooper pairs (between a superconducting grain and bulk superconductor) is of special interest in this context: since the gap for single-particle excitations is approximately the BCS energy gap Δ both for the grain and for the bulk for processes involving energy less than this (eg. Josephson tunnelling of pairs) m and n are indeed the groundstate.

CAPACITANCE

Finally, let us turn briefly to the notion of capacitance. The very simplest system of all is an empty hollow dielectric sphere in vacuum. We ask: How much energy does it take to bring up charge Q from infinity and place it on the sphere, allowing it to distribute itself as it thinks best? The answer by definition is $E = Q^2/2C$ where C is the self-capacitance of the sphere. Note that the definition strictly speaking only makes sense if we take the limit $Q >> e$; obviously, to bring up the first electron costs nothing, and the energy to add the second, while finite, is not numerically equal to $e^2/2C$ as C is defined above. (For large Q the charge is spread more or less over the surface of the sphere, while for Q=2e it is clear that this is energetically unfavorable and the system prefers to correlate the two electrons so that they sit predominantly at one another's antipodes). Actually this feature is something of a pathology of the empty sphere and has rather little significance, c.f. below.

The next simplest situation is the parallel-plate capacitor or a topologically equivalent object: see Fig. 2. Here the relevant question is: How much energy does it take to *transfer* a given amount of charge Q from one side of the barrier to another, keeping the total charge constant and allowing the charge to do what it likes best *within* each of the two plates? Again, the answer is $E = Q^2/2C$, where C is the capacitance of the system as traditionally defined. Note that in both this case and that of the isolated sphere we have simply $E(Q) = \int V(Q)dQ$, where V(Q) is the voltage difference for the given value of Q, so $V(Q) = dU/DQ = C^{-1}Q$ and we could equally well define C as dQ/dV (or simply Q/V).

The case of an experimentally realistic metal grain in contact with a bulk metal is very similar; in this case, as distinct from the sphere in vacuum, we must remember to allow the electrons to adjust themselves in the bulk metal as well as in the grain. If d is the thickness of the dielectric shell separating the two and ε its dielectric constantt, it is clear that the order of magnitude of the capacitance (in SI units) will be the larger of the quantities $4\pi\varepsilon_0 R$ and $4\pi\varepsilon_0 R^2/d$. (As a mnemonic, it is useful to remember that the "capacitance" of the H atom ($R \sim a_0 \sim 0.5$ Å) is $e^2/4$ Rydbergs, i.e. approximately $2 \cdot 5 \times 10^{21}$ F).

As soon as we have more than two regions of interest (or wish, in the parallel-plate case, to change the *total* charge) the definition needs a little more care. Suppose for example we have three regions (Fig. 3). The most natural definition is to start with the system in equilibrium and ask for the energy $E(Q_1, Q_2, Q_3)$ to place extra charge Q_i are the i-th region allowing it to distribute itself *within* the regions as it likes best. Then we write

$$E(Q_1,Q_2,Q_3) = \sum_{i=1}^{3}\sum_{j=1}^{3} \left(\frac{1}{C}\right)_{ij} Q_i Q_j , \tag{10}$$

where C^{-1} is the inverse capacitance *matrix*. Since this is by its definition symmetric, it has 6 independent elements; however, if we insist on the condition of total charge neutrality

$$\sum_{i=1}^{3} Q_i \equiv 0 ,$$

only two of the Q_i's are independent and hence C^{-1} in effect becomes a 2 x 2 symmetric matrix with only 3 independent elements. Generally we can see that with n different regions the capacitance matrix will have n(n+1)/2 independent elements if total charge can be changed, or n(n-1)/2 if it cannot. Since the work done in charging up the system can be written as

$$\sum_{i=1}^{3} \int V_i(Q_j)\, dQ_i ,$$

we have

$$V_i(Q_j) = \frac{\partial E}{\partial Q_i} = \sum_{i=1}^{3} \left(\frac{1}{C}\right)_{ij} Q_j , \tag{11}$$

and hence

$$Q_i = \sum_{j=1}^{3} C_{ij} V_i \tag{12}$$

when C is the matrix inverse of C^{-1}. Bring a symmetric real matrix, C_{ij} can clearly be diagonalized.

We finally note a more general definition of capacitance which will be useful when discussing circuits later on. Imagine a system of arbitrary shape, containing arbitrary barriers etc. which starts with expectation value of charge density zero everywhere. For any (operator) disturbance in the charge density $\rho(r)$, we have

$$E_{Coul} = \frac{1}{2} \int\int \frac{\rho(r)\rho(r')}{|r - r'|} d^3r\, d^3r' . \tag{13}$$

We ask now: Suppose we wish to produce a certain value of the charge density weighted with a certain function f(r), that is, we want a specified value of the quantity

$$A[f] \equiv \int f(r)\, \delta\rho(r)\, d^3r , \tag{14}$$

but otherwise put no restrictions on the adjustment of the system. Then we define a generalized capacitance C(f) by the statement that the necessary energy E[f] is given by

$$E[f] \equiv \frac{1}{2} C^{-1}[f]\, A^2[f] . \tag{15}$$

Actually this is only a special case of a yet more general definition: consider a whole set of orthogonal functions f_i, and define

$$A_i \equiv \int f_i(r)\, \delta\rho(r)\, d^3r . \tag{16}$$

Then we can define the minimum energy necessary to create a specified set of A_i's by

$$E = \frac{1}{2} \sum_{ij} \left(\frac{1}{C}\right)_{ij} A_i A_j . \tag{17}$$

It is clear that the cases of discrete regions considered above correspond to special case of the above generalized approach, with a restriction to only a few of the A_n. In these special cases (but *not* in the more general case) the eigenvalues of the Q_i are quantized in units of e. Finally, we note that all these relations remain valid when we regard $\delta\rho(r)$, and hence A_n, and E as *operators.*

To classify recent work on problems associated with the discreteness of the electron charge (or perhaps more accurately with small capacitances—see below) it is convenient to focus on a specific physical system, namely a metallic grain separated by thick oxide barriers from two bulk metals (Fig. 4). In practice, the grains studied experimentally range in size from around 10 Å up to say 1000 Å, (corresponding to capacitance energies in the range ~0.1 meV - 0.1 eV) (Bloch, 1970) and in some cases one of the "bulk metals" may be the tip of an STM, so that the relevant oxide layer is partly or wholely replaced by vacuum. We assume that at the start of the process we consider the grain is in equilibrium and approximately electrically neutral. Thus, it will contain N^+_0 positive ions and N_0 electrons, where N_0 is some integer close to N^+_0 (but not necessarily equal to it, because of chemical effects). Even for the smallest grains of experimental interest, N_0 is a large number, ≥ 300. We also note that the spacing of single-particle energy levels is of order $N_0^{-1/2}(\hbar^2/mR^2$ where R is the radius of the grain. The crucial point, now, is that if we add one more electron to the grain, then the total energy of the universe will be increased by the extra electrostatic energy $e^2/2C$ where C is the mutual capacitance of the grain and the relevant bulk metal; since C is of order $4\pi\varepsilon_0 R$, this is much greater than the spacing of the single-particle levels (recall that for $R \sim a_0$, the Bohr radius, the two terms are of the some order of magnitude). Thus, the energetics of charge transfer into the grain, and hence the conduction from one bulk metal to another, is overwhelmingly determined by the Coulomb (capacitance) energy.

It might at first sight seem puzzling that there is a finite capacitance energy even when the grain was originally electrically neutral, since we saw that for an empty sphere in vacuum no such energy was associated with the first electron transferred. The answer is that both in the grain and in the bulk metal the electron produces a polarization of the pre-existing charge, which results in the usual Thomas-Fermi type screening. As a result, a cloud of positive charge (i.e. a reduced density of electrons) of net charge +e and of radius ~ k_{FT}^{-1} ~1Å (inverse Thomas-Fermi wave vector) forms around the electron, and the missing negative charge is pushed to "infinity", or rather to the boundaries of the system (Giaver and Zeller, 1968). The "short-range" gain in correlation energy so achieved is the same in the grain and in the bulk metal, but the Coulomb energy associated with the repelled charge of -e produces an electrostatic energy $e^2/2C$; for the bulk metal this is essentially zero since C is very large and the remainder is just what we would have naively calculated (Kulik and Shekhter, 1975).

Suppose now that we start as above with a definite number of electrons on each side and on the grain, and consider the process of tunnelling of an electron from one of the metals into the grain. Let Ψ_{Ni} denote the different rates of the grain with N particles, and χ_{Nbj} similarly those of the bulk metal; let 0 denote the original state of each. Then, if for the moment we focus on a process in which the electron vacates a given single particle state in the bulk and tunnels into a definite state on the grain, and neglect all processes of interaction within the grain and the bulk, we have schematically after time t for the many-body wave function $\Psi(t)$ of the relevant "universe" (grain plus the bulk metal in question)

$$\Psi(t) \equiv a(t)\, \Psi_{N0}\chi_{N_b 0} + b(t)\, \Psi_{N+1,m}\chi_{N_b - 1,n} . \tag{18}$$

Now we ask: Do we need to keep track of the phase coherence between these two terms, or could we at this stage simply apply standard Golden Rule perturbation theory in the tunnelling matrix element and say that there is simply a certain probability per unit time that the electron

has tunnelled from the state corresponding to m in the bulk metal to that corresponding to n in the grain?

At first sight one would think that under most practical conditions the answer would be yes. In the first place, we normally actually have not a single second term but a superposition of a large number of such terms corresponding to different values of m and n, and one's instinct is to think that in this situation most operators we would want to measure would be such that the relative phase of the interference terms between the N and N+1-particle states would depend on m and n in a random way, so that the net effect of the interference is nearly zero. Secondly, unless n is very close to the groundstate of the N+1-particle system, it will be reasonably easy for the electron to make an inelastic transition to a lower state, e.g., by emission of a phonon. Since the two states in question are now correlated to orthogonal states of the phonon field, such a process is effectively phase-destroying and allows us to neglect interference from then on.

Neither of these arguments is quite as strong as it looks, and much work over the last five years or so has been based on the idea that under certain, albeit stringent, conditions it may be necessary to keep track of the phase coherence between states with different numbers of electrons on the grain. Without going at present into the relevant conditions, let me define a "coherent" effect as one which results (in the calculation) only when we keep track of the phase coherence, and an "incoherent" one as one which we get even when we make the standard Golden-Rule assumption.

If both bulk metal and grain are superconducting the situation is a bit different (and actually simpler). For present purposes I neglect completely the normal electrons in the superconductors, for which considerations similar to the above areas apply. Then we recall that the many-body wave function of a superconductor as described standard by BCS theory has the form (neglecting spin indices so as not to clutter the notation)

$$A\,\Psi(r_1 r_2)\,\Psi(r_3 r_4)\ldots\Psi(r_{N-1} r_N)\,, \tag{19}$$

where A is the antisymmetrization operator and the "pair wave function" $\Psi(r,r')$ is *the same* for all pairs. Violation of this latter condition is equivalent to breaking up a pair, a process which requires an energy of at least 2Δ (Δ = BCS energy gap); for the present we ignore such processes. Moreover, in a simple s-wave superconductor the dependence on relative coordinate $r - r'$ is fixed by the energetics, with a large characteristic locking energy (though cf. below) so the only variable of interest is the center-of-mass coordinate $R = (r + r')/2$, which from now on I shall write simply as r. The quantity $\Psi(r)$ then at first sight plays much the same role as the single-electron wave function in the above discussion, except that it corresponds to two electrons rather than one. However, there is now a crucial difference: considering its behavior within the grain, the "single-particle-like" wave function $\Psi(r)$ can take various forms $\Psi_n(r)$ (for example, a simple s-wave state with radial quantum number zero, a p-like state, a radially excited state, etc.) If we were indeed dealing with the wave function of a single particle, the splitting between the corresponding energies would be of order h^2/mL^2 for the first few (L = characteristic dimension of grain). However, there are $\sim N_0/2$ pairs occupying the state in question, so the energy differences are of order N_0h^2/mL^2. Moreover, the pairs effectively obey Bose, not Fermi, statistics so there is no exclusion principle. Consequently, the probability of finding even the first excited state Ψ_1 thermally occupied at temperature T is of order

$$P_1 \sim \exp\left[-\frac{N_0 h^2}{mL^2 kT}\right], \tag{20}$$

which is vanishingly small even for a 10Å grain of high-temperature superconductor. Moreover, if the system starts with m=0 it is clear that to push it into the state m=1 would require, as an intermediate stage, going through states in which $\sim N_0$ pairs are broken. This would cost an energy $\sim N_0\Delta$ and the probability of such processes is therefore again vanishingly small. Thus, for the grain only the Cooper-pair "groundstate" needs to be considered. For the bulk metal the argument needs to be a little more sophisticated, since there may be sufficient energy in external fields, etc., to take the pairs into an excited state (or a linear superposition of the groundstate and excited states) however, the point is that all Cooper pairs are always in *the same* superposition, so that the phase of this wavefunction close to the tunnelling barrier is a relevant quantity and must be kept in the calculation. Thus, if both grain and bulk metal are superconducting we must *always* keep track of the phase coherence between the states with N_p and N_{p+1} pairs on the grain. (If only one of the two is superconducting, the situation more clearly resembles that in a normal metal (above)).

Thus, we can classify the various effects which have been considered according to two variables (Fig. 5): whether all components involved are superconducting or not, and whether or not it is necessary to keep track of the phase coherence. While the notation is not completely standard, I shall use the term "Coulomb blockade" for normal incoherent effects and "single-electron-tunnelling" (SET) for normal coherent affects (Mullen *et al.*, 1988). I will call those coherent effects occurring in superconductors which vanish in the limit of infinite grain capacitance (as distinct from the well-known Josephson effect, etc.) "superconducting phase fluctuation" (SPF) effects; one often hears in this context the words "Bloch oscillations" but for reasons which will become clear I wish to avoid this terminology. The fourth class (incoherent superconducting effects) is empty, as I have argued above (Cleland *et al.*, 1990). Note carefully that unlike some authors I do *not* use a classification in terms of "constant voltage" as against "constant current"; again, the reasons for this will become clear.

In my opinion the simplest of the three types of effect conceptually, and the best understood (and arguably the best verified experimentally) is the "Coulomb blockade" class. Experiments on such effects go back to Giaever and Zeller (1969), and a good discussion of most of the important theoretical aspects was given in Kulik and Shekhter (1975). The basic principle is very simple. Imagine that initially our two bulk metals are in equilibrium with one another and with the grain, and denote by N the number of electrons on the grain over and above that needed to maintain charge neutrality; the original value of N is denoted N_0. By our earlier argument, the energy E(N) has a term linear in N, which is of no particular interest to us, and a term of the form $N^2e^2/2C$ (note that it doesn't matter whether or not we subtract a term $Ne^2/2C$ from this). There will actually be another term proportional to N^2, corresponding to the familiar finite compressibility of a neutral Fermi gas, but this is of order $N_0^{-1/3}h^2/mL^2$ (where N_0 is the *total* number of electrons originally present) and thus typically very much smaller than the Coulomb term; we will from now on neglect it. If then we define the bulk electrochemical potentials in the usual way, and set as above $\mu(N) \equiv E(N+1) - E(N)$, then since in equilibrium it cannot be energetically advantageous either to add or to subtract an electron from the grain, we have (Fig. 6)

$$\mu(N_0) \geq \mu_1 - \mu_2 \geq \mu(N_0 - 1) = \mu(N_0) - \frac{e^2}{C}\,. \tag{21}$$

Now imagine that we apply a finite electrostatic potential difference V between the two bulk metals: $V_1 = V_2 + V$. In general this implies a partial electrostatic potential difference between each metal and the grain, so that $V_g = V_2 + \lambda V$, $0 \leq \lambda \leq 1$. As we will see, the actual value of λ is of no great interest, but it can if necessary be calculated (see Kulik and Shekhter, 1975). Consider now whether a current will flow between the two metals via the grain (we assume for present purposes that there is no *direct* contact between the metals). At zero temperature, there are two possibilities: either (1) an electron makes an energetically possible transition from 1 to G and then a second energetically possible transition from G to 2 (possibly, but not necessarily, emitting one or more phonons, etc., along the way) or (2) it makes a virtual (energetically forbidden) transition from 1 to G and then goes from G to 2. If we assume for simplicity that the tunnelling matrix elements T are of the same order for the two barriers, then the probability of process is proportional to $T^4/(\Delta E)^2$ where ΔE is the energy deficit, while that of process 1, if it goes at all, is of order $T^4\rho^2(E)$, where $\rho(E)$ is the density

of states in the grain (see below). Thus, process 2 can be ignored unless ΔE is only of the order of the spacing of the single-particle levels, which is so only for very narrow ranges of voltage. At finite temperature a third process is available, namely thermal activation to an "energetically forbidden" states. The probability of this process is proportional to exp[- $\Delta E/kT$] and can be neglected for sufficiently small T; for the moment we make this assumption.

It is clear, now (see Fig. 6) that for process 1 to go we need to satisfy at least one of the conditions

$$\mu_1 \geq \mu(N_0), \quad \mu_2 \leq \mu(N_0 - 1). \tag{22}$$

It is clear from the diagram (Fig. 6) that this will happens when either $e\lambda V = \Delta\mu_0$ or $(1-\lambda)eV = e^2/C - \Delta\mu_0$, whichever happen first. It is clear after a moments thought that at least one of these conditions must be satisfied at some value of the applied voltage V less than e/C: $V = \alpha e/C$, $0 < \alpha < 1$. In general we may not know the value of α a priori; however, it is clear that if we reverse the sense of the voltage then the threshold is $(1-\alpha)e/C$. Once we are beyond the threshold, charge can flow freely into and out of the grain just as if the capacitance were infinite, *provided* that only one extra electron is present at any one time. The actual magnitude of the current will be proportional to the excess of voltage above the threshold (since this determines the width of the "band" of electrons which are free to tunnel). Thus the net effect is to offset the current-voltage characteristic by a step of width e/C along the voltage axis, not necessarily centered at zero, as indicted in Fig. 7. This is the fundamental "Coulomb blockade" effect.

The above analysis implicitly assumes that, when an electron is transferred from 1 to G, it hops off again on to 2 before there is an appreciable possibility of a second electron hopping on to C. In cases where this is not so the situation is more complicated and we get the so-called "Coulomb staircase" with a series of steps spaced by e/C along the voltage axis. Observation of this effect requires that the rate of hopping across one barrier should be considerably greater than that across the other, i.e that the two asymptotic resistances differ by a factor >> 1.

One can go on and investigate the dependence of the Coulomb-blockade and Coulomb-staircase effects on quantities such as temperature the effect of any non-junction-like ("ohmic") resistance shunting the junction, and circuit elements which may lie between the outside (so-called "bulk") metal and the actual voltage source. Qualitatively, it is clear that the effect of temperature will be to smear out the blockade and staircase effects, since it will mean that the effective value of E(N+1)-E(N) fluctuates by an amount of order kT; and that the order of magnitude of the temperature at which the washing-out is effectively complete will be of order e^2/C [which can still be as high as room temperature for the smallest grains investigated to date, cf. Mullen *et al.* (1988)]. Again, it is qualitatively plausible that any ohmic resistance shunting either of the junctions will tend to wash out the effect, since it gives an opportunity of varying $\mu(N)$ by a displacement of the charge in the resistor, and such a variation does not in general corresponds to integral steps of e^2/C. (Note that this does *not* mean that charge is somehow transferred to the grain (as defined by a unique boundary) in units less than e, merely that the effect on the electrochemical potential is the same as if it had been). Finally, the effect of external circuit elements can be thought of as giving fluctuations of the voltage actually presented to the junction around the "mean" value which is externally applied. Indeed, most of the qualitative physics is contained in a simple calculation by Cleland *et al.* (1990) which represents the system (or to be more precise, 1+G) by the circuit shown in Fig. 8, calculates from the fluctuation-dissipation theorem the probability of a fluctuation ΔV of the voltage presented to the tunnelling barrier and takes the actual I-V characteristic to be an average of the characteristic $I(V+\Delta V)$ weighted with $P(\Delta V)$. This is not quite the whole story, because it neglects the time correlation of the fluctuations, and more complete calculations have been recently done by at least three groups (Devoret *et al.*, 1990; Girvin *et al.*, 1990; Nazarov, 1990).

I now turn to the question of superconducting phase fluctuations. (The question of single-electron-tunnelling is, I believe, considerably more complicated and will be treated in the lecture by Dr. Likharev.) Suppose now that both metals and the grain are superconducting, and that the oxide barriers form Josephson tunnel junctions, i.e. that there is a matrix element for the transfer of a *Cooper pair* across the barrier. (From now on I completely neglect any single-electron processes which may be going on in parallel.) As is well known, this leads to a term in the Hamiltonian which is of the form

$$E_{junct} = - E_J \cos\phi, \tag{23}$$

where ϕ is the phase difference of the Cooper-pair wave function in the grain and in the bulk. Now, this phase difference is the variable canonically conjugate to the number of pairs transferred across the junction. To see this, let us go back for a moment to the situation shown in fig. 6, that is, two otherwise isolated chunks of metal connected by a junction. Neglecting any variation of the phase within each bulk metal separately, we write the many-body wave function schematically in the form

$$\Psi_N = (a|R> + b|L>)^N, \tag{24}$$

where we omit the indication of antisymmetrization to avoid cluttering up the notation. Here |R> and |L> are normalized and mutually orthogonal components of the Cooper-pair wave function Ψ corresponding to being on the right (left) of the barrier. If we define the operator n_{LR}^i for each Cooper pair to have eigenstates |R> and |L> with eigenvalue +1 and -1 respectively, then the operator

$$\Delta N = \sum_i n_{LR}^i$$

has the significance of the total excess number of electrons transferred from L to R. Evidently we have for any arbitrary pair of states Ψ, Ψ' of the form (24)

$$(\Psi'_N|\Delta\hat{N}|\Psi_N) = (a'^*a - b'^*b)(\Psi'_{N-1}, \Psi_{N-1}). \tag{25}$$

Now write $a \equiv |a|\exp[-i\Delta\varphi/2]$, $b \equiv |b|\exp[-i\Delta\varphi/2]$, so that $\Delta\varphi$ is the phase difference of the Cooper-pair wave function across the junction, and evaluate the quantity $(\Psi'_N|-i\,\partial/\partial\varphi|\Psi_N)$. Evidently it gives exactly the same expression as (26). Thus, at least within the manifold of states of the type (24) the operator $-i\,\partial/\partial\varphi$ is identical to the operator $\Delta\hat{N}$. [For further details, see e.g. Leggett (1966).]

If the two chunks of metal in question are reasonably macroscopic, then the energy of the system may contain a term linear in ΔN, e.g. due to an external applied potential difference, but there will be only a vanishingly small term quadratic in ΔN. Then it is consistent to treat the phase as a perfectly well-defined (classical) variable and we get the standard Josephson theory of the junction; any linear terms in ΔN will simply produce an a.c. voltage across the junction in the well-known way, with the Josephson frequency $\omega = 2eV/\hbar$. On the other hand, if one of the chunks is a small grain we have a Coulomb term in the energy $E_{Coul} = (\Delta N)^2 e^2/2C$ as discussed above. This term causes the relative phase to fluctuate, and in fact where $e^2/C \geq E_J$ the fluctuations are of order 2π and so the phase is in some sense hardly defined at all. It is effects connected with this feature that I call "superconducting phase fluctuation" effects.

In the limit that the fluctuations of the phase, while not zero, are small compared to 2π these effects have been studied intensively over the last decade or so, both theoretically and experimentally. Most of these studies have concentrated on one of two geometries: the so-called rf SQUID ring and the current-biassed junction. Neither of these are identical to the geometry we have used as an illustration up to now, so let me very briefly remind you what they are. The SQUID ring is simply a bulk superconducting ring (of thickness >> London penetration depth) interrupted by a single Josephson junction; the current-biassed junction is what its name implies, namely a single Josephson junction in series with a large impedance, so

as to produce (or so one hopes!) a constant current into the region of the junction (more on that below). The SQUID ring has the special property that because of the London relation, the phase difference across the junction is, up to a physically irrelevant additive factor of $2n\pi$, simply the total flux ϕ through the ring (external and self-induced) measured in units of the flux quantum $h/2e$; hence it is conventional to take the flux as the relevant coordinate. However, it must be emphasized that the variable p_φ canonically conjugate to the flux is a "distributed" charge, namely $e\Sigma_i\theta_i$ where θ_i is the angle of displacement of the inth electron around the ring; for a detailed discussion of this point, see Leggett (1986, 1987a). (This is clearly just a special case of the generalized "charge" A[f] discussed above.) This variable therefore has a continuous spectrum of eigenvalues, which is necessary for consistency since the variable to which it is conjugate, the flux, is certainly "nonperiodic" (i.e. states differing by are certainly physically distinct.) Moreover, the capacitance associated with p_φ is a "distributed" capacitance, defined by the statement that the minimum energy necessary to produce a given value of p_φ is $p_\varphi^2/2C$. Since in practice, for not too small junctions and not too large rings, it is likely to be dominated by the capacitance of the junction itself, it is often referred to loosely as "the junction capacitance" but it should be remembered that this is strictly speaking an approximation. It should be emphasized that the dimensions of the SQUID rings used in some recent experiments have been quite macroscopic (~1 cm) and therefore must correspond to capacitances ~0.1 pF or greater, i.e. much larger than those of the grains discussed earlier (with corresponding values of $e^2/2C \sim 10$ mK); the only reason that one has any chance of seeing any effects of phase fluctuations at all for such large values is that one can arrange to play off the other large energies in the problem against one another in various ways.

It is useful to note that the standard Hamiltonian of the SQUID ring used in most of the theoretical analysis is

$$H_{SQ} = \frac{p_\varphi^2}{2C} + V(\Phi:\Phi_{ext}) , \tag{26}$$

where $V(\Phi:\Phi_{ext})$, is a sum of two terms corresponding respectively to the Josephson coupling and the inductive energy of the ring:

$$V(\Phi) = -E_J \cos(\Phi) + \frac{(\Phi - \Phi_{ext})^2}{2L} , \tag{27}$$

where the external flux Φ_{ext}, and hence V, may be time-dependent. The general appearance of the curve $V(\Phi)$ depends on the relative importance of the two terms, i.e. on the ratio $2\pi LI_c/\varphi_0$ (sometimes called β_L). If the latter is ≤ 1, the curve is essentially a parabola with a weak periodic ripple superimposed, and has no metastable minimum. On the other hand for $\beta_L >> 1$, the curve has the characteristic shape shown in fig. 9, with a large number ($\sim\beta_L$) of metastable minima separated by barriers which are typically of the order E_J in height.

In the case of the current-biased junction the usual convention is to take the "coordinate" variable as the phase difference φ of the Cooper pairs across the junction, and the Hamiltonian as

$$H_{CBJ} = \frac{p_\varphi^2}{2C} + V(\varphi:I_{ext}) , \tag{28}$$

where $V(\varphi:I_{ext})$ is again, the sum of the Josephson term and a term of the form $-I_{ext}\varphi$, where I_{ext} is the external current fed in through some large impedance (usually not explicitly specified) into the junction region. The resulting form of V, for $I_{ext}<I_c$, the critical current of the junction ($\equiv 2\pi E_J/\varphi_0$) is that shown in Fig. 10 and often called, for obvious reasons the "washboard potential". The additional term $-I_{ext}\varphi$ has the right form to generate the classical equation of motion, namely

$$C\frac{\partial^2\varphi}{\partial t^2} + I_J \sin\varphi = I_{ext} , \tag{29}$$

but is a source of some problems in the quantum case: see below.

Most of the theoretical work on "small" ($\Delta\varphi << 2\pi$) superconducting phase fluctuations over the last decade has been on the phenomena known as (1) "macroscopic quantum tunnelling" (MQT), that is, the escape from a metastable well into what is effectively a continuum, and (2) "macroscopic quantum coherence" (MQC), that is, coherence and consequent oscillations between values of flux corresponding to nearly degenerate potential energy (e.g. the lowest two wells in Fig. 9) (Devoret *et al.*, 1990). MQT is predicted to occur in both SQUID rings and Josephson junctions, while MQC occurs (at least prima facie!) only in SQUIDs. Much of the recent theoretical work has been devoted to the effect of dissipation (due e.g. to any normal resistance shunting the junction) on these phenomena, which is only quantitative for MQT but can easily change things qualitatively for MQC. [For a review see e.g. Leggett (1987b)]. On the experimental side, the existence of the phenomenon of MQT is well confirmed, both in rings and current-biased junctions, and the observed rates generally agree well with the theoretical predictions; with a view to later remarks I note that the most impressive quantitative agreement is obtained in experiments on current-biased junctions. Regarding MQC, some experimental work has been done, but for the most part only in the regime where the spectacular oscillations are expected to be washed out by dissipation; nevertheless the most recent results (15) seem to be quantitatively consistent with the theoretical predictions.

There is however a quite different class of phenomena which have been predicted theoretically, which require phase fluctuations of the order of 2π or greater. The fundamental effect is usually called "Bloch oscillations" and is well reviewed in several recent reviews (Lapointe *et al.*, 1990; Likharev, 1988). (The so-called SET effects are to some extent the one-particle analog of Bloch oscillations, but are more complicated because of the possibility of incoherent as well as coherent behavior). The fundamental experimental prediction is this: If a Josephson junction is biased with a fixed d.c. external current I_{ext}, there the voltage across the junction oscillates *coherently* with a frequency $\omega = 2\pi I/2e$, i.e. the frequency with which on average the Cooper pairs cross the junction. If this prediction is correct, then as emphasized by Likharev it has fundamental implications for metrology and much else.

Once one accepts the Hamiltonian (27) for the current-biased junction, and moreover allows oneself to treat φ as an extended variable (so that $\varphi + 2n\pi$ and φ represent physically distinct states, then the existence of Bloch oscillations follows inexorably. In fact, with these assumptions the situation is identical to that of an electron in a periodic potential in an electric field, which was the situation originally considered by Bloch (hence the name). The argument runs as follows in that case: The energy eigenstates in the absence of the field are Bloch waves with momentum k. In the presence of a weak electric field, and in the semiclassical approximation we have

$$h\frac{dk}{dt} = eE . \tag{30}$$

However, quantities such as the velocity are *periodic* functions of k, with period $2\pi/a$ (a = lattice period). Thus for a d.c. field E they will shown oscillations at a frequency $\omega = 2\pi eE/ha$ (the original "Bloch oscillations"). An exactly analogous argument for the current-biased junction, with φ as the analog of position x and with the voltage across the junction given by the Josephson relation $\varphi = 2eV/\hbar$, produces the above result.

The crucial question is whether the starting Hamiltonian (27) is a legitimate description in cases where fluctuations of the phase over a distance of order 2π, and coherence between the corresponding amplitudes, are crucial. I do not want to get into this argument in the present context except to examine one consideration which is sometimes cited as relevant, namely that the current-biased junction can be regarded as the limit of the SQUID where $\varphi_{ext} >> \varphi$, with

$I_{ext} \rightarrow \varphi_{ext}/L$. In this connection I just want to make two two perhaps trivial remarks. First, there is of course no question of actually obtaining Bloch oscillations as defined above, that is as a d.c. (or ultra-low-frequency) effect. in a SQUID: if we sweep the external flux very slowly, then irrespective of the parameters all quantities, including the current flowing into the junction, simply oscillate at the Josephson frequency $\omega = (2e/\hbar)(d\Phi_x/dt)$. Secondly, we can consider a state in which the ring is highly excited and the wave function is approximately semiclassical and thereby well localized relative to the amplitude of motion. In such a time-dependent state, if we consider a time interval much less than the period of oscillatory motion, we will indeed get an oscillatory voltage due to the residual effect of the Josephson term, but it emerges already at the classical level and its frequency bears no simple relation to I(t). Thus, if there is a residual effect of the Bloch-oscillation phenomenon, it is well hidden.

One might think that the impossibility of producing Bloch oscillations in a SQUID ring had to do with the fact that the wave function is coherent all around the ring, and that one could produce an appropriate configuration by breaking the ring apart and driving the resultant circuit from a voltage source. However, it is very easy to show that the equivalent circuit simply reduces again in effect to the SQUID ring and nothing new is thereby produced. In fact, it is easy to see that one will never produce Bloch oscillations with a purely inductive lead, for the simple reason that such a lead will never give a truly current-biased source where it is needed; to have such a source we need L to be much greater, for *all* φ, than the effective impedance of the junction itself, but the latter is proportional to $(\partial^2 E/\partial\varphi^2)^{-1}$ and therefore becomes infinite at $\varphi = \pi/2$. (The shunting capacitance cannot be relevant in the d.c. limit.)

One might at this point stop to ask why, in that case, the MQT experiments agree with the simple theory, since the relevant values of φ in these experiments are precisely around $\pi/2$? The answer is of course that the leads in these experiments are *resistive*, and thus the competition is between impedances proportional to R and to ωL_{eff}; since the sweep rate ω is typically of the order of a few kHz, we have $R >> \omega L_{eff}(\varphi)$ except for a very tiny interval of φ immediately around $\pi/2$. Since this interval is much smaller than the width of the barrier which is tunnelled through it has negligible effect.

The fundamental conclusion is that, if the Bloch oscillation phenomenon is real, it will *only* be obtained when the leads to the junction are resistive. It is therefore of crucial importance to learn how to formulate a complete quantum-mechanical description of an ohmic resistor. I have high hopes that the approach outlined at the end of my first lecture will enable us to do this, and with luck I may even have succeeded by the time this lecture is delivered!

REFERENCES

Bloch, F., 1970, *Phys. Rev. B*, 2:109.

Chen, Y. C., Fisher, M. P. A., and Leggett, A. J., 1988, *J. Appl. Phys.*, 64:3199.

Cleland, A. N., Schmidt, J. M., and Clarke, J., 1990, unpublished.

Devoret, M. H., Esteve, D., Grabert, H., Ingold, G-L., Pothier, H., and Urbina, C., 1990, unpublished.

Giaever, I., and Zeller, H. R., 1968, *Phys. Rev. Lett.*, 20:1504.

Girvin, S. M., Glazman, L. I., Jonson, M., Penn, D. R., and Stiles, M. D., 1990, unpublished.

Kulik, I. O., and Shekhter, R. S., 1975, *Zh. Eksp. Teor. Fiz.*, 68:623; (*Sov. Phys. JETP*, 41:308).

Lapointe, J., Han, S., and Lukens, J., 1990, unpublished.

Leggett, A. J., 1966, *Prog. Theor. Phys.*, 36:901.

Leggett, A. J., 1986a, *in* Directions in Condensed Matter Physics, ed. C. Grinstein and G. Mazenko, World Scientific, Singapore.

Leggett, A. J., 1987a, *in* Chance and Matter, ed. J. Souletie, J. Vannimenus and R. Stora, North-Holland, Amsterdam.

Leggett, A. J., 1987b, Proc. 18th International Conference on Low Temperature Physics, Kyoto, p. 1986.

Likharev, K. K., 1988, *IBM J. Res. Dev.*, 32:144.

Martinis, J. M., Devoret, M. H., and Clarke, J., 1985, *Phys. Rev. Lett.*, 55:1543.

Mullen, K., Ben-Jacob, E., and Ruggiero, S., 1988, *Phys. Rev. B*, 30:5150.

Nazarov, Yu.V., 1990, unpublished.

Schon, G., and Zaikin, A. D., 1990, unpublished.

SINGLE-ELECTRONICS: CORRELATED TRANSFER OF SINGLE ELECTRONS IN ULTRASMALL JUNCTIONS, ARRAYS, AND SYSTEMS

K.K. Likharev

Department of Physics
Moscow State University
Moscow 119899 GSP, U.S.S.R.

INTRODUCTION

Several recent years have been marked by creation and rapid development of a novel field which was nicknamed the single-electronics (see earlier reviews by Likharev, 1986; Likharev, 1988; Mullen *et al.*, 1988; and Averin and Likharev, 1990). This field deals with physics and applications of a new effect, the correlated transfer of single electrons and/or Cooper pairs in ultrasmall tunnel junctions, as well as in arrays and other systems of such junctions.

The basic physics of this effect is extremely simple, and can be well described by example of the simplest system, an insulated tunnel junction between two normal-metal electrodes (Fig. 1). If the tunnel conductance G of the junction is small enough, the system can be considered as a leaking capacitor, and characterized (except for G) by the capacitance C, roughly proportional to the junction area S. Let a single electron pass through the junction. This event changes the initial electric charge Q of the junction by e, and hence the initial voltage V across the barrier, by $V=e/C$.

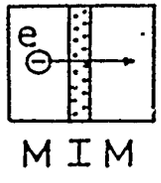


Fig. 1. The simplest system of the single-electronics, an isolated tunnel junction with small capacitance C

Now let us wonder: whether this change affects the following tunneling events? For the typical tunnel junctions used in most experiments (say, of area $S=0.1\times0.1$ mm^2 at temperature T=4.2 K), the answer is *negative*: The voltage change is completely masked by the thermal fluctuations of the voltage, because the relevant elementary energy of the single-electron recharging is

$$E_c = \frac{e^2}{2C} , \qquad (1)$$

and is much less than k_BT. Hence, the tunneling events can be safely considered as uncorrelated.

However, modern nanolithographic techniques allow one to fabricate junctions with S as small as 20×20 nm^2 (see, e.g., Dolan and Dunsmuir, 1988). For these ultrasmall junctions the elementary energy E_c can be as large as 100 K, so that at helium (and lower) temperatures, the thermal fluctuations cannot mask V. Here the quantum fluctuations can become substantial. Elementary theory shows that the energy scale of these fluctuations in our relaxation-type system is hG/C, so that they are also negligible if the junction resistance R=1/G is much larger than the well-known quantum unit $R_Q=h/4e=6.5$ kilohm.

As a result of the above considerations, for ultrasmall junctions with $GR_Q<<1$, the answer to the above question is *positive*: under certain conditions: a single electron tunneling event can lead to a considerable change of the tunneling rate for other electrons, in spite of the fact that the metallic electrodes of the junction can contain quite a large number (millions and even billions) of free electrons! This change results in a *correlation* of the tunneling events either in time, or in space, or both. The type of the correlation is strongly dependent of the system under consideration.

The purpose of this paper is to give a short summary of the resulting effects and their possible applications. I will emphasize very recent results which could be not included in the recent comprehensive review (Averin and Likharev, 1990). I will also skip a discussion of the correlated tunneling of Cooper pairs, which can take place in ultrasmall Josephson junctions, because this part of the field should be covered here by Professor A. Leggett, and thus restrict myself to the single-electron tunneling in normal-metal and semiconductor systems.

THEORETICAL BACKGROUND

In the development of single-electronics, experiments have mostly followed theoretical predictions, so that the role of the theory was (and remains) quite high. This is why before passing to concrete results, I should comment about the theoretical techniques which have proved to be so successful.

Most results were obtained using the so-called orthodox theory of the single-electron tunneling (Averin and Likharev, 1990; Schön and Zaikin, 1990). Considering an arbitrary system of L metallic resistors and N tunnel junctions connecting M metallic electrodes, the theory starts from the following Hamiltonian:

$$H = G + \sum_{i=1}^{N} H_{e,i} + \sum_{j=1}^{M} H_{T,j} + \sum_{k=1}^{L} H_{s,k} . \qquad (2)$$

Here, G is the free (Gibbs) energy of the system, which takes into account the electrical charging of the electrodes and/or the junctions and their interaction with the electrodynamic environment. For example, for the insulated junction (Fig. 1), G equals just Q/2C, but if the junction is connected to a source of fixed current I(t), the Gibbs term -I(t)V(t)dt should be added to G. Tunneling in each junction is described by the standard Hamiltonian (Bardeen, 1961)

$$H_T = H_+ + H_- \ , \ H_+ = \sum_{p,q} T_{pq} c_p^+ c_q \ , \ H_- = H_+^+ \ . \qquad (3)$$

where c^+ and c are the electron creation and annihilation operators, p and q denote electron states on opposite sides of the tunnel barrier, and T_{pq} are the corresponding matrix elements. The electrode Hamiltonians H_e are assumed to have a continuous spectrum, and to commute with the operators of the electric charge Q (the last assumption is only valid when each electrode contains a large number of electrons). On the contrary, H_T generally does *not* commute with the operators of the electric charge of the junctions and operator functions F of these charges; for example, in the simplest system (Fig. 1)

$$H_\pm F(Q) = F(Q \pm e) H_\pm \ . \qquad (4)$$

Lastly, the operators H_s of the metallic resistors can be presented in the standard form suggested by Caldeira and Leggett (1983).

One can see that the orthodox theory neglects several factors including: (a) finite dimensions of the junctions, electrodes, and resistors; (b) finite durations of the electron tunneling process and electric charge redistribution within the electrodes just after that event; (c) energy quantization in electrodes, *etc.*

In spite of these omissions, this theory yields quite a reasonable description of the experimentally observed properties of various systems comprising metallic electrodes with linear sizes from crudely 10 nm to at least 300 nm. Another important advantage of the orthodox theory is its reduction to a very simple "quasi-classical" form for the case when all the junction conductances are small enough ($G_i << 1/R_Q$). In this case, one can forget the Hamiltonians and restrict oneself to calculation of the classical Gibbs energy G of the system as a function of a set of integer numbers n_i of electrons passed through the junctions of the system. The theory predicts that the time rate (i.e. probability per time unit) of tunneling from a state $\{n\}=\{n_1, n_2,...,n_i,...,n_N\}$ to the state $\{n'\}=\{n_1, n_2,...,n_i+1,...n_N\}$ is given by the expression

$$\Gamma_i^\pm = \frac{1}{e} I(\frac{\Delta G^\pm}{e}) \{ 1 - \exp(-\frac{\Delta G^\pm}{k_B T}) \} \ , \qquad (5)$$

where $-G=G\{n'\}-G\{n\}$ is the free energy change due to this tunneling event, and $I_i(U)$ is the dc I-V curve of the i-th junction, for the case when the dc voltage U across it is fixed (and hence its electric recharging by the tunneling electron is excluded). For the normal-metal junctions and relatively small voltages $I_i(U)=G_i U$, but (5) gives also an adequate description of more complex situations (e.g., superconducting electrodes, tunnel barrier suppression, *etc.*) where I(U) is essentially nonlinear.

One can see that even in the quasi-classical limit the theory gives a typically probabilistic prediction rather than a dynamic equation; i.e., the tunneling process is generally random. Nevertheless, in some situations, this *statistical* description shows that the process of tunneling is quite *ordered*, i.e. that the single-electron tunneling can be remarkably correlated.

THE SIMPLEST SYSTEMS: PREDICTIONS AND OBSERVATIONS

Single junction

The simplest system where such a correlation can take place is the single tunnel junction biased by a fixed external current (Fig. 2). Theory says that the dc I-V curve of such a system should consist of two branches (Fig. 3): a horizontal branch with

$$I = 0 \quad \text{for} \quad -\frac{e}{2C} < V < \frac{e}{2C} \quad \left(\text{i.e.} -\frac{e}{2} < Q < \frac{e}{2} \right) , \qquad (6)$$

and a "conductive" branch which approachs linear conductance, but has offset asymptotes

$$I(V) \rightarrow G(V - V_{of} \operatorname{sign} V) \ , \quad \text{for } |V| >> \frac{e}{C} \qquad (7)$$

at large voltages. In the present case, $V_{of}=e/2C$.

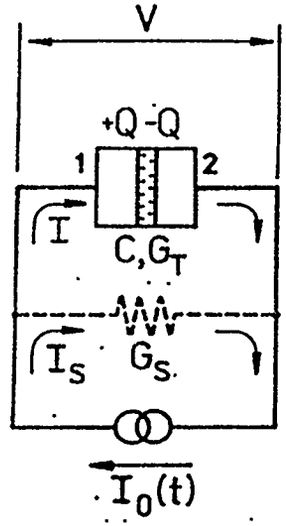


Fig. 2. Current-biased single tunnel junction. Dashed line shows a possible ohmic conductance of the junction environment

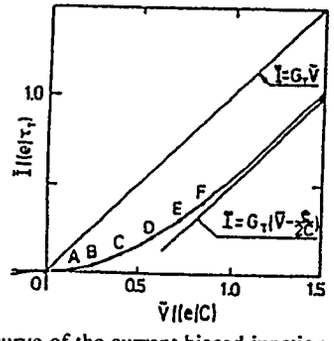


Fig. 3. DC I-V curve of the current biased junction for T=0 and G_S=0 (after Averin and Likharev, 1985, 1986)

In the horizontal branch, corresponding to (6), the tunneling is completely suppressed (at low temperatures); this state was nicknamed the *Coulomb blockade of tunneling* (Averin and Likharev, 1985, 1986). The physics of this effect is extremely simple: within the Coulomb blockade range (6) any tunneling event would lead to an increase of the Gibbs energy $G=Q^2/2C$

(Fig. 4), so that according to (5) the probability of such an event is vanishing small at $k_BT<<E_c$.

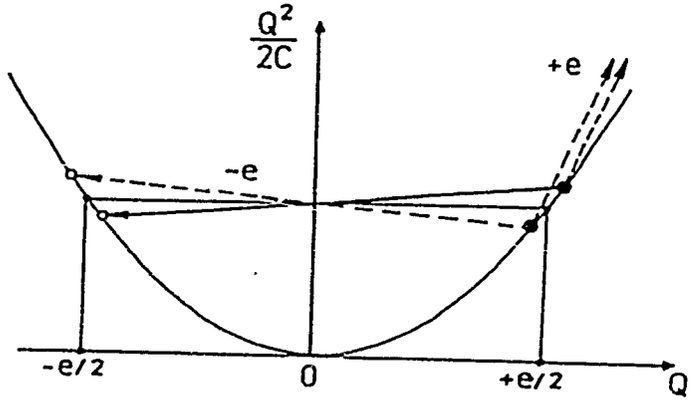


Fig. 4. Energy diagram illustrating the physical origin of the Coulomb blockade of tunneling. Dashed lines show (virtual) energy-unfavorable tunneling events which are suppressed at low temperatures, while the solid line shows the energy-favorable event

What is less evident is the possibility to have the junction charge to a value Q=ne. This possibility (now confirmed by numerous experiments, see below) arises from the fact that the charge passed through metallic conductors is generally *not* quantized on the scale of the elementary charge e, because transfer of this charge results from a collective shift of the free electron gas with respect to the atomic lattice. Thus the charge of a tunnel junction can be generally changed in two different ways: *discretely* due to the discrete single-electron tunneling events, and (quasi)*continuously* due to the metallic conduction of an environment (e.g., the current source in our present system); I should refer the reader to the review (Averin and Likharev 1990) for a more detailed discussion of this point.

The interplay of the two mechanisms of charge transfer results in quite specific dynamics of the junction if it is current-biased to the conductive branch of the I-V curve (Fig. 3): the voltage V and hence the electric charge Q of the junction oscillate in time with the frequency (Averin and Likharev, 1985, 1986)

$$\frac{\omega_s}{2\pi} = \frac{\bar{I}}{e} . \qquad (8)$$

The physics of these so-called *Single-Electron-Tunneling* (SET) oscillations is also very simple: while Q is within the limits (6), the tunneling is Coulomb-blocked, and the junction charge changes in time as Q=I(t). When Q reaches an edge of the range, the tunneling becomes energy-favorable (G>0), it really takes place and brings the system to the opposite edge of the range (see the solid arrow in Fig. 4). After that, the whole process is repeated periodically; the exact e-quantization of the *tunneling* charge immediately yields (8) for the frequency of this process.

Presently this relation between the frequency of the SET oscillations and the dc current seems at least as fundamental as the famous relation $f_J=2eV/h$ between the frequency of the Josephson oscillations and the dc voltage. (Generally there exists a deep though incomplete analogy between the correlated tunneling and the Josephson effect; the reader is referred to the review (Averin and Likharev, 1990) for a discussion of this analogy).

Theory says (Averin and Likharev, 1986, 1987) that, in contrast to the frequency of the SET oscillations, their amplitude is vulnerable, and is suppressed by larger dc current (I>eG/C), temperature ($k_BT>E_c$), and environment conductance ($G_s>G$). Nevertheless, estimates show that the effect could be observed using present-day fabrication and experimental techniques if the junction was really current-biased. Unfortunately, in a typical experimental situation the stray capacitance of the current/voltage leads (necessarily attached to the junction) is much larger than that ($C\sim10^{-16}$ F) of the ultrasmall junction itself. As a result, the junction virtually becomes voltage-biased and the single-electron charging effects vanish.

A possible way (Likharev and Zorin, 1985) to circumvent this problem is to insert nominally large, but physically small, resistors into the leads very close to the junction and thus cut-off most of the stray capacitance. Preparations to such experiments are in progress in several laboratories, and some preliminary encouraging results have already been obtained (Martinis and Kautz ,1989; Cleland *et al.*, 1990; Claeson *et al.*, 1990), but some quantitative improvements are still to be made before an observation of the SET oscillations (and of the closely related "Bloch" oscillations, see the contribution in this volume by A. Leggett) in a single junction becomes possible.

Double junction

The experimental situation is much better for slightly more complex systems of ultrasmall junctions, because there the parasitic effect of the interlead capacitance is not so disastrous. For example, in the double-junction system shown in Fig. 5, the net capacitance $C=C_1+C_2$ between the middle electrode and its environment is not affected by the stray capacitance C_L, so that *some* correlation of the tunneling events can take place even if no care is administered to the lead configuration and C_L is very large.

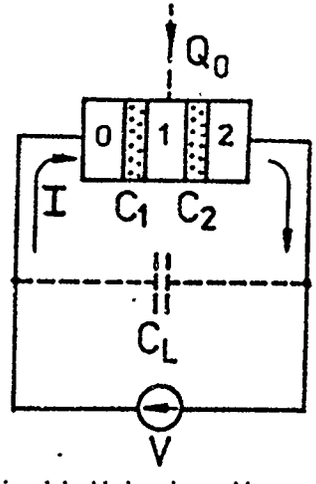


Fig. 5. A voltage-biased double junction and its capacitances.

Theory (Likharev, 1987) shows that it is not the *time* correlation between the events in a particular junction (i.e. the SET oscillations) discussed above, but rather the "*space*" (mutual) correlation of the events in the counterpart junctions that is important. These arise due to the large charging energy Q/2C of the middle electrode, so that an electron entering through either junction leaves it rapidly through the other junction. A direct measurement of this correlation presents rather a problem, but fortunately this effect leaves several signatures on the dc I-V curve of the system (see e.g. Fig. 6).

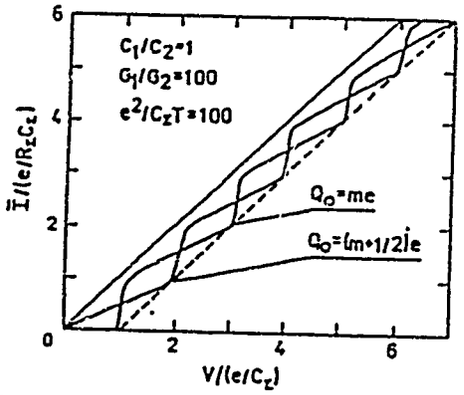


Fig. 6. Typical dc I-V curves of the double-junction system for two values of the background charge Q_0 of its middle electrode

First of all, one can see a clear horizontal part of the curve near the origin, corresponding to the Coulomb blockade of tunneling. Second, the I-V curve asymptotes are again offset, in agreement with (7) with $V_{0f}=e/C_\Sigma$. Third, the I-V curve exhibits a voltage-periodic pattern (the so-called *Coulomb staircase*); theory says that each period of the staircase corresponds to recharging of the middle electrode by one electron. Lastly, if an additional continuous charge Q_0 is injected from outside into the middle electrode of the system (Fig. 5), the staircase pattern is continuously and proportionally shifted with respect to the origin. Shift by one period corresponds to injection of exactly one fundamental charge e, while injection of $Q_0=e/2$ changes the pattern phase by π (Fig. 6).

The first observations of the first two features were made as early as the 1960s (Zeller and Giaever, 1969). Lambe and Jaklevic (1969) also obtained indirect evidence of the voltage-periodic behavior of a similar system. In these works, the observations were correctly identified with the single-electron charging effects, and their semi-quantitative theory was developed (this theory was improved considerably by Kulik and Shekhter, 1975). Unfortunately, the authors of these pioneering works had to deal with experimentally available granular structures consisting of a large number of double junctions, with a considerable random scattering of their parameters. Probably this is why the decisive role of the continuous charge has not been noticed and the very idea of the correlation of the tunneling events had not been put forward before the mid-1980s (Averin and Likharev, 1985, 1986).

To my knowledge, the first observations of the I-V curves of an individual double junction were made in Moscow by L. Kuzmin in early January 1987 (Kuzmin and Likharev, 1987a,b). He used a system consisting of two continuous metallic (lead-indium-gold alloy) thin films separated by the In_2O_3 oxide, with tiny (100-nm) grains of In imbedded into the insulating layer (Fig. 7a). In contrast to the works of the 1960s, measures were taken to insulate completely all grains of the sample except one which played a role of the middle electrode of the double junction. This care allowed reliable observation of all the features listed above and a quantitative confirmation of all predictions of the theory (see, e.g., Fig. 7b).

Subsequently, some of these observations were reported also by Barner and Ruggiero (1987). Moreover, these authors (Ruggiero and Barner 1987) as well as Kuzmin and Safronov (1988), pointed out that numerous observations of the periodic singularities in the dc I-V curves of point contacts with the copper-oxide high-T_c superconductors can be interpreted as a consequence of their double-junction structure, with the role of the middle electrode played by high-conducting grains within the naturally formed semi-insulating layer on the surface of the materials.

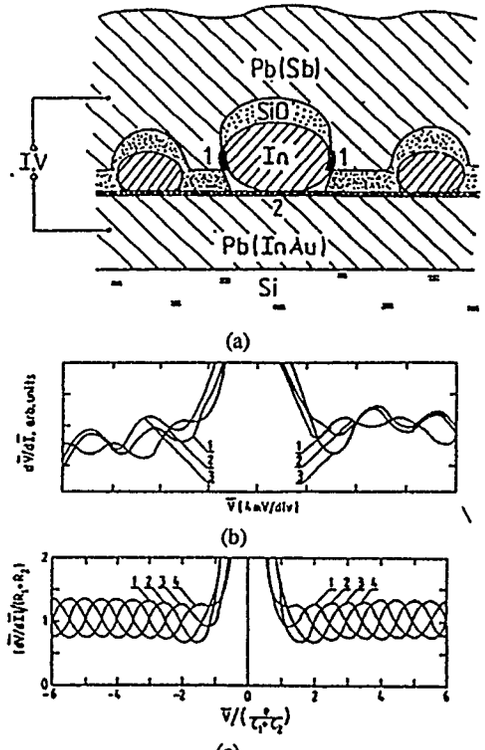
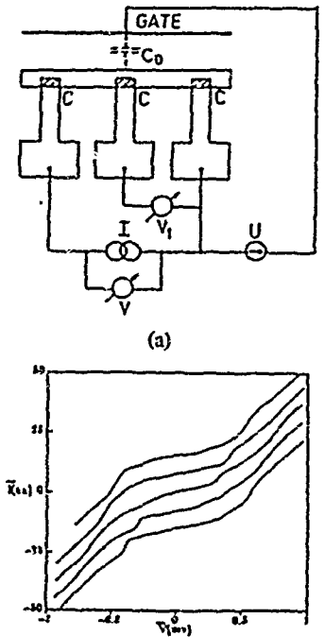


Fig. 7. (a) Scheme of the granular tunnel structure used in the experiments (Kuzmin and Likharev, 1987a,b), (b) the derivative dV/dI as a function of the dc voltage V for several experiments with such a structure, and (c) results of calculations using the orthodox theory for various values of Q_0.

Almost simultaneously, a more advanced version of the double-junction experiment was done by Fulton and Dolan (1987). In this version (Fig. 8a), the central electrode was an aluminum thin film strip connected to external thin film electrodes through the ultrasmall (30x30 nm^2) tunnel junctions, which were formed by a "shadow mask" technique. Advantages of this version include not only a possibility of its reproducible fabrication but also an openness of the middle electrode to external electric fields, in particular to that produced by a special "gate" electrode. It is straightforward to prove (Likharev, 1987) that a change of the gate voltage by U is equivalent to a change of the (continuous) charge of the middle electrode by $Q=C_0U$. In this way, Fulton and Dolan have been able to demonstrate directly the e-periodic dependence of the system I-V curve on Q_0 (Fig. 8b).

Another possible version of the double-junction experiment employs the scanning tunnel microscope (STM) technique (van Bentum *et al.*, 1988). First, separate metallic grains are deposited on a preliminarily oxidized metallic surface. Then the STM is used to select a desired grain and to fix the STM tip over it (Fig. 9a). After that, the STM is switched into the

"spectrometric" mode, which allows the observer to record the dc I-V curves of the resulting double-junction system. Figure 9b shows some results of the recent experiment of this kind (Wilkins *et al.*, 1989). They imply in particular that the orthodox theory of the single-electron tunneling remains *quantitatively* correct for the middle electrode (grain) sizes as small as 5 nm.

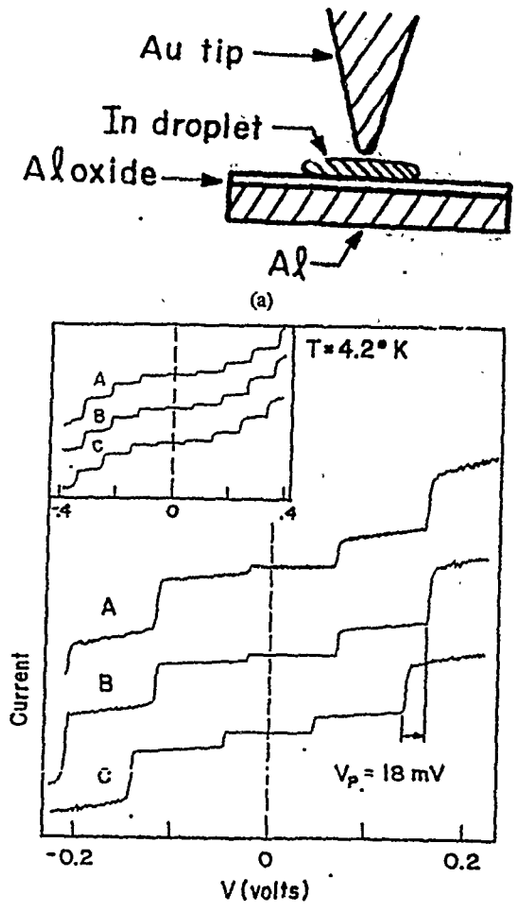


(a)

(b)

Fig. 8. (a) Scheme of the experiment by Fulton and Dolan (1989) and (b) dc I-V curves of this structure for several values of the gate voltage U.

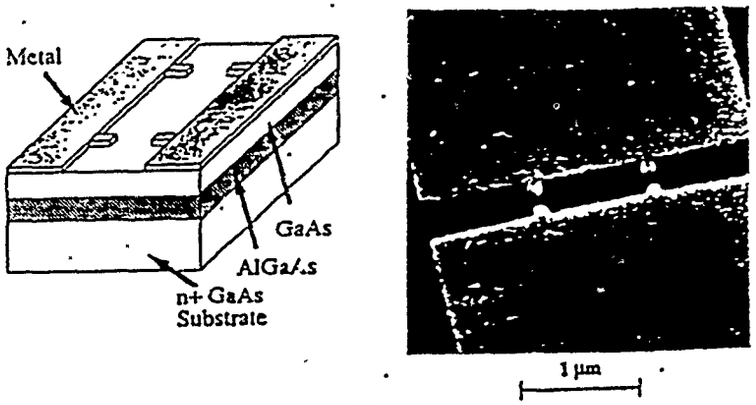
Finally, the typical double-junction behavior was recently observed in one more system: GaAs heterojunctions with a two-dimensional (2-D) electron gas (Meirav *et al.*, 1990). The heterojunction was supplied with a split gate, of the geometry shown in Fig. 10a, so that, within some range of the gate voltage, the electron gas was presumably confined to a long and narrow domain separated from the main 2-D regions by tunnel barriers. An additional insulated gate (not shown in Fig. 10a) was deposited over the split gate, and its voltage could change the effective value of the background charge Q_0 of the "middle electrode" (the confined electron gas). When the zero-field dynamic conductance of the system was recorded as a function of the dc voltage applied to this electrode, a quasiperiodic pattern was clearly observed (Fig. 10b). Calculations have confirmed that each period of the pattern corresponded to Q_0=e for each of several lengths of the confinement area. A detailed comparison of the data with predictions of the orthodox theory has not yet been reported.



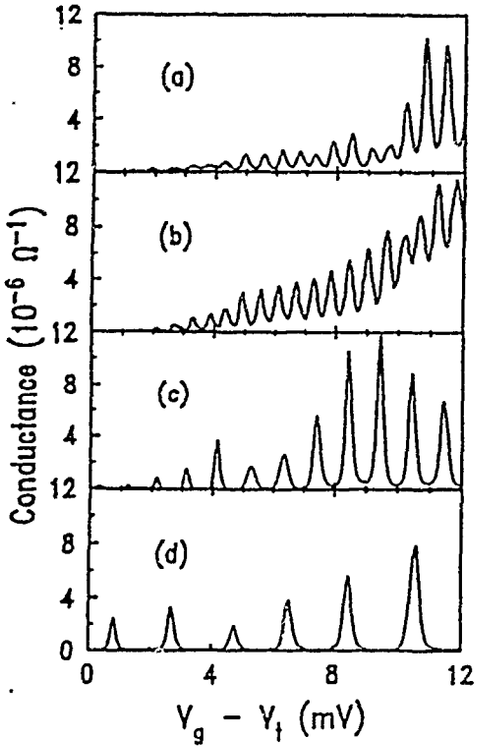
(a)

(b)

Fig. 9. (a) Scheme of the STM double-junction experiment and (b) some results of the experiments by Wilkins *et al.* (1989). Curves A and B are experimental, while curve B was obtained by numerical simulations of the junction dynamics (for the parameters corresponding to curve A)



(a)



(b)

Fig. 10. (a) Scheme of the GaAs heterojunction structure used by Meriav *et al.* (1990) and (b) zero-field dynamic conductance of several structures with various distances between the gate gap constrictions as a function of voltage V_g applied to the additional gate

Multi-junction arrays

A natural question to ask after the first successful double-junction experiments was: whether some other system allows to combine the space correlation of the tunneling events with their time correlation, preferably at the most practical condition of the voltage bias? This question was answered positively by Likharev (1988) (see also Likharev *et al.*, 1989): it is sufficient to use a (quasi)uniform one-dimensional (1-D) array of the ultrasmall junctions (Fig. 11a).

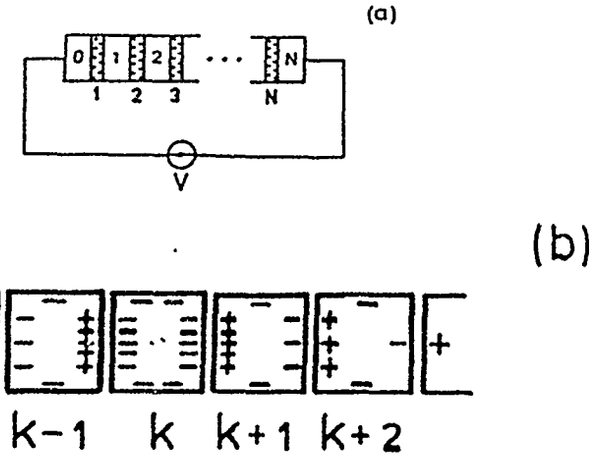
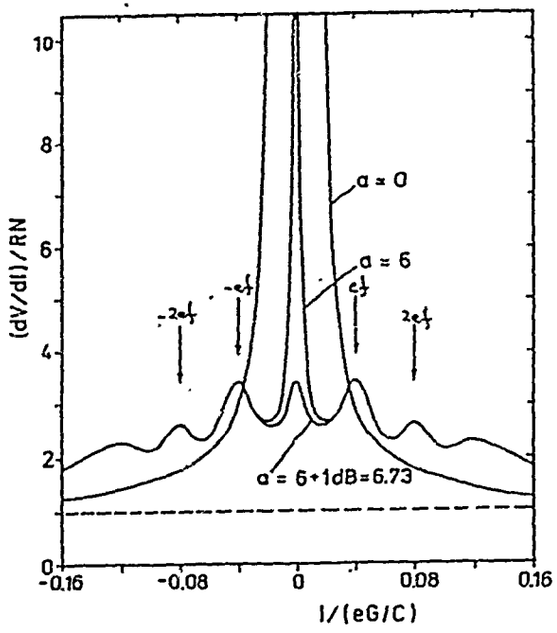


Fig. 11. Schemes of (a) the 1-D array of the tunnel junctions and (b) of the single-electron soliton in such an array (after Likharev, 1988, Likharev *et al.*, 1989)

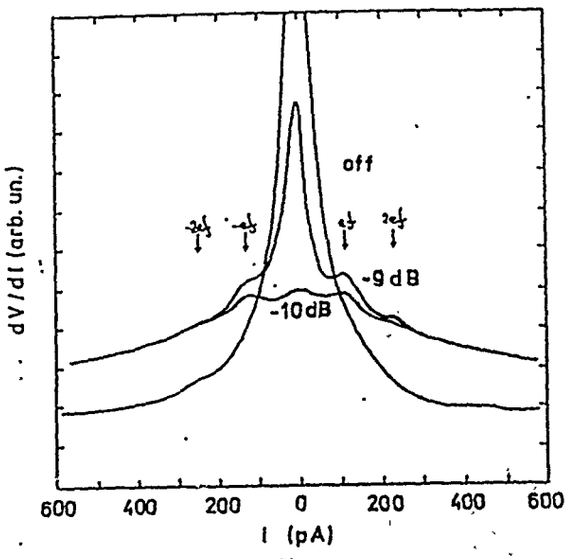
Dynamics of this system can be most naturally explained in terms of *topological single-electron solitons*. Such a soliton can be considered as a result of injection of a single electron/hole into an electrode of a previously discharged long array. In order to provide electroneutrality of the electrode interior, the injected charge comes up to the surface. Those parts of the surface charge which face the neighboring electrodes polarize them, *etc.* (Fig. 11b). This polarization chain would be infinite if a part of the charge was not lost on the lateral sides of the structure, forming a charge pattern (the soliton) of a finite length. (Within a simple but reasonable model (Likharev, 1988), which restricts the general matrix of the electrostatic coupling of the electrodes to the tunnel junction capacitance C and the electrode stray capacitances C_0, the soliton field decays exponentially with the characteristic number of junctions $M=(C/C_0)$ in the usual condition when $C_0<<C$).

In order to drive a soliton into the array, it is sufficient to apply a voltage above a certain threshold V_t (for $V<V_t$ the array remains in its Coulomb-blockade state with a vanishing current). After its entrance to the array edge, the soliton drifts along this edge. The solitons of similar charge repulse each other, so that if the voltage threshold is only slightly overcome the new soliton enters the array only after the first one covers a certain distance from the edge. As a result, a quasi-periodic soliton lattice sliding along the junction array is formed, so that the electron tunneling events become correlated both in time and space. At sufficiently large N and M the correlation can be virtually complete, which means in particular that the system generates narrow-band SET oscillations with frequency given by (8).

The 1-D arrays of ultrasmall tunnel junctions Al/Al-oxide/Al have been fabricated by shadow mask techniques and studied at millikelvin temperatures by the Swedish-Soviet collaboration (Kuzmin *et al.*, 1989; Delsing *et al.*, 1989a,b) and by the Delft group (Geerligs *et al.*, 1989). After the fabrication technology had been improved enough to reduce the random scattering of the junction parameters, the SET oscillations were reliably observed in July 1989 (Delsing *et al.*, 1989b). For this observation, the arrays were irradiated by microwaves with a frequency f in the range 0.7-5 GHz. Theory predicts that a microwave field applied to the array, together with the dc field, should lead to at least partial phase locking of the SET oscillations at the dc current values I=nef (n= 1, 2,...), and, as a result, to peaks of the dynamic resistance of the array at these values (see Fig. 12a). Figure 12b shows an experimental result for this set of parameters; the peaks are clearly visible. Figure 13 shows positions of the peaks as the function of the microwave frequency, one can see that within the experimental (few-per-cent) precision they confirm the fundamental relation (8).



(a)



(b)

Fig. 12. Dynamic reristance of a 19-junction array as a function of the dc current for several values of the microwave amplitude: (a) theory and (b) experiment (after Delsing *et al.*, 1989b)

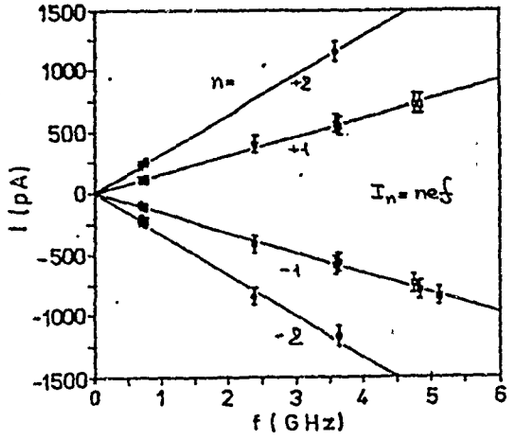


Fig. 13. DC current positions of the dynamic resistance peaks as the functions of the microwave frequency for 15-junction and 19-junction arrays (after Delsing *et al.*1989b)

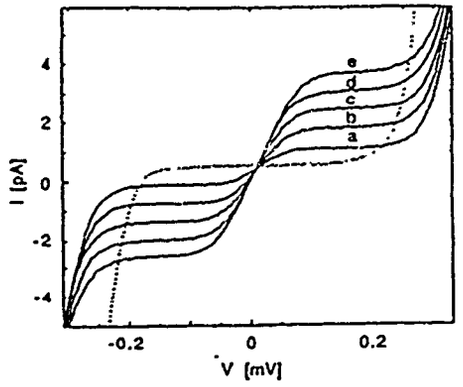


Fig. 14. DC I-V curves of the "turnstile device" by Geerligs *et al.* (1990) at microwase drive frequencies from (a) 4 MHz to (e) 20 MHz

Another remarkable experiment with the 1-D arrays was carried out recently in a joint work of the Delft and Saclay groups (Geerligs *et al.*, 1990). They have noticed that, even if the array is as short as N=4, one can provide transfer of exactly one electron along the array per one period of the microwave field, if the field is applied across the array rather than along it and the system parameters are chosen properly. Figure 14 shows dc I-V curves of this "turnstile" device for several values of f; one can see long dc voltage steps at the dc current I=ef. (Deviations of the step position from this value did not exceed few parts per thousand). This device can be quite useful for applications (see below) despite the fact that it *cannot* generate the SET oscillations in the absence of the microwaves (so that the dc voltage steps can be interpreted as a result of induction of the SET oscillations by the rf drive rather than their phase locking).

A behavior qualitatively similar to that of the 1-D arrays can be expected from similar 2-D arrays (Geigenmüller and Schön, 1989; Bakhvalov *et al.*, 1990). In particular, dynamics of arrays of a large size can be understood again in terms of motion of the single-electron solitons (now two-dimensional ones). At certain conditions, the 2-D lattice formed by these solitons can be driven along the junction array without melting and thus induce the SET oscillations. Numerical simulations of the dynamics show, however, that the parameter window for the narrowband SET oscillation in the uniform 2-D arrays is generally narrower than that for their 1-D counterparts. Experimentally, only the simplest feature of the single-electron charging effects, the Coulomb blockade, was registered reliably for the 2-D arrays (Geerligs and Mooij 1988).

POSSIBLE APPLICATIONS

The quantitative agreement of the experimental data obtained for the metallic-electrode ultrasmall tunnel junctions with the orthodox theory of the single-electron tunneling allows one use this theory to estimate possible applications of these new effects.

DC current standards

An apparent duality between the dc voltage steps at the quantized dc current values I=nef for correlated tunneling and the celebrated dc current ("Shapiro") steps at the quantized values V=nhf/2e in the Josephson effect makes it evident that the former effect can be used for development of the fundamental standard of the dc current, similar in structure to those of the dc voltage.

A detailed analysis of the dc current precision achievable with the standard is still to be carried out. Nevertheless, general results concerning stability of the Coulomb blockade to thermal and quantum fluctuations (Averin, 1986; Khrómov, 1986; Golub, 1987, Zaikin and Panyukov, 1988; Odintsov, 1988, Averin and Odintsov, 1989) do allow one to predict that in principle this precision can be of the same order as that for the Josephson standards, and hence much better than the present-day standards of the dc current (including the projected fundamental standard of current based on combination of the Josephson-effect standard of voltage and the quantized-Hall-effect standard of resistance).

Another major problem is the range of the currents which can be maintained by the standard. For a single 1-D array with the SET oscillation phase locking, estimates give figures from 10^{-11} to 10^{-9} A. This range can be extended to higher currents by some two orders of magnitude to higher currents (i.e. to ~10^{-7} A) using the phase locking of the Bloch oscillations (Zorin *et al.*, 1990), and possibly by two or three more orders (i.e. to 10^{-5} A) using single-chip multi-array circuits. A further increase of the current is apparently possible using the superconducting comparators (Dziuba and Sullivan, 1975).

Extension of the range to smaller currents (down to 10^{-20} A) seems feasible with the turnstile device (Sec. 3.3) modified to suppress a parasitic effect of the macroscopic quantum tunneling of the electric charge (Averin and Odintsov, 1989), which is essential for few-junction systems. Another possibility in nearly the same range (10^{-21} to 10^{-15} A) is to calibrate dc current from some stable source by its injection into the middle electrode of the double-junction system (Fig. 5) and counting oscillations of its I-V curve in time (one oscillation means one injected electron). Advantages and drawbacks of these two methods are still to be compared, both theoretically and experimentally.

Supersensitive Electrometry

The last device discussed in the previous section (the double junction with injection of external charge Q_0 into its middle electrode) is in fact a supersensitive electrometer. Estimates show (Likharev, 1987; Vasenko *et al.*, 1990) that its charge sensitivity in the white-noise frequency range can be as high as 10^{-5} e/Hz$^{1/2}$ for the present-day junction parameters. Experimentally, the sensitivity of a few 10^{-4} e/Hz$^{1/2}$ has been registered (Kuzmin *et al.*, 1989) at 10 Hz where the 1/f noise can be essential.

These figures are to be compared with the charge sensitivity of the order of 10^{2} e/Hz$^{1/2}$ of the best present-day electrometers available commercially. Thus, one can see that the improvement is really very substantial, so that the single-electron devices can really revolutionize the electrometry, just as the celebrated SQUIDs did the magnetometry. Practical implementation of these devices can be hindered by the absence of electrostatic analogs of the superconducting dc transformers which are vital components of the SQUID magnetometers.

Digital Circuits

The most important field of the applied single-electronics is doubtlessly digital devices and circuits. One possible way to design such devices is to employ double junctions as the sub-single-electron-controlled transistors (Likharev, 1987). However, presently another way, which presumably makes full use of the unique ability of the SET devices to control motion of single electrons (Likharev and Semenov, 1987), seems preferable. In the latter "*Single-Electron-Logic*" (SEL) devices, the binary unity/zero is coded by presence/absence of a single electron in a certain metallic electrode during the given clock period. It is important that the time boundaries of the clock period are marked also by single electrons passed through special clock lines. (Most ideas of SEL circuits are borrowed from the Josephson-junction RSFQ logic

circuits (see, e.g., Likharev and Semenov, 1990), although design of practical SEL devices requires many specific techniques).

The SEL circuits are quite simple. For example, Figure 15 shows a universal logic/memory cell (Likharev and Semenov, 1987; Averin and Likharev, 1990), which can perform either OR or OR-NOT function, depending on which of its inputs (A,B) is fed by the signal electrons (the complementary input is fed by the clock electrons). On the other hand, the circuits have quite large (a-few-ten-per-cent) parameter margins.

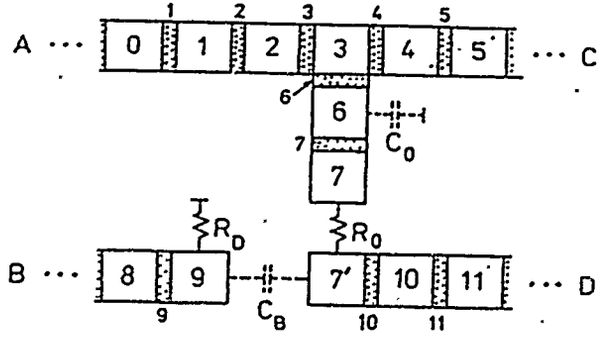


Fig. 15. A universal logic/memory cell of the SEL family (after Likharev and Semenov, 1987). The figure does not show dc drive sources inducing drift of the signal and clock electrons in the general direction from left to right

However, as one can see from Table 1 (Likharev, 1987; Averin and Likharev, 1990), the main advantage of the single-electron digital devices is their possible incredible integration scale. A rather conservative estimate of the scale yields a figure like 10^9 logic/memory cells (not just junctions!) per 1-cm chip even for the present-day fabrication technology. (Of course, in order to fabricate such chips, the technology should be extended to VLSI circuits, which is apparently a hard task). This figure is to be compared with estimates of 10^7 gates per chip for the semiconductor-transistor and Josephson-junction devices.

Note also that the SEL circuits can combine this integration scale with quite a reasonable operation speed and a power dissipation low enough to avoid self-heating. Moreover, the low power dissipation makes 3-D integration physically feasible. Probably, the only possible way to these 3-D circuits is a use of the organic macromolecules which would combine an acceptable internal structure (say, a metallic-cluster nuclei surrounded by a protein shell) with an ability to self-assemble into predetermined structures (cf. Fig. 15).

Of course, practical implementation of this program can take much time and effort, but apparently possible 3-D integration scale of these "circuits" (up to 10 cells per cm) is so incredible that the program deserves at least a careful preliminary study. To the best of my knowledge the single-electronics presents the only realistic physical basis for processing the digital information on the molecular level.

CONCLUSION: PROBLEMS AND PROSPECTS

In the field of practical single-electronics, the most urgent problem is an improvement in technologies for fabrication of thin-film structures comprising ultrasmall tunnel junctions and small-size high-ohmic resistors. Fundamental problems of the field are more numerous.

Table 1
Estimates of the basic parameters of the single-electron tunnel junction devices (after Likharev 1987, Averin and Likharev 1990)

Fabrication technology level	Junction area S (nm)	Temperature limit T_{max} (K)	Power scale (W)	Time scale (ps)	Integration scale (gates/cm)
State-of-the-art junctions	30x30	30	10^{-11}	3	10^9
Apparent limit of the nano-lithograghy	10x10	300	10^{-9}	0.3	10^{10}
Hypothetic molecular structures	3x3	3000	10^{-7}	0.03	10^{11}

Energy quantization

All basic assumptions of the orthodox single-electronics become invalid when the size of the system electrodes is reduced beyond certain limits. Probably, the first assumption to fail is that of the continuous energy spectrum: in smaller electrodes the free electron energy becomes quantized. Influence of this quantization on the correlated tunneling in the double junctions has been analyzed recently by Averin and Korotkov (1990). They have concluded that, for realistic values of the energy relaxation rate inside the electrodes, the only effect of the quantization is some fine structure of the dc I-V curves (superimposed on the Coulomb staircase). Two experimental techniques seem most suitable for a verification of these predictions, as well as for more general studies of the energy quantization effects: STM spectroscopy applied to ultrasmall metallic particles, atomic clusters, and certain macromolecules, and electric measurements of the ultrasmall-area quantum wells (Averin *et al.* 1990). For more complex structures (for example, long 1-D arrays) the situation may be more complex, so that the problem deserves much attention.

Tunneling and relaxation time effects

With a decrease of the junction size, both classical (C/G) and quantum (h/E_c) time scales of the single-electron recharging may become comparable with the "traversal" time of tunneling τ_t (Büttiker and Landauer, 1982) and the reciprocal plasma frequency ω_p (which seems to be a relevant time scale of the surface charge pattern formation). If this is the case, all conclusions of the orthodox theory (which ignores two latter times) should be revised.

Presently I know next to nothing about the possible results of such a revision. One can guess, however, that, for example, in the 1-D arrays, there should exist a crossover between the orthodox single-electronic behavior (i.e., a consequent tunneling of the electron through the junctions with a complete relaxation of the acquired energy and loss of the wave-function phase memory after each tunneling event) and the usual resonant tunneling (implying a well-determined wave-function phase). Experimentally, this crossover can be presumably most easily approached in semiconductor (notably GaAs) superlattices of an ultrasmall cross-section (say, 100x100 nm^2), because these structures can provide relatively large values of τ_t and ω_p.

Of course, beyond this crossover, the *quantitative* picture of the single-electron tunneling should be quite different from those discussed above. One can ask, however, whether the *qualitative* conclusion concerning a substantial correlation of the electron motion should be necessarily wrong there. At this stage of research, I would not exclude a chance that some well-known models describing the 1-D resonant tunneling (say, the Hubbard model) do really imply some kind of the correlation within a certain parameter window. I would very much like to know the real answer to this question.

Correlation without tunneling

One more tempting question concerning the correlated single-electron transport is whether the correlation can be achieved *without tunneling at all*. Strange as it is, answer to this question may also be not necessarily negative. Consider a metallic microshort in the tunnel barrier of the ultrasmall tunnel junction. If an electron can pass through such a point contact, one can repeat all arguments above and arrive again at all the basic conclusions of the orthodox theory, provided that the contact conductance G satisfies the usual requirement $GR_Q<<1$. For usual 3-D metals, this condition is very hard to satisfy: in order to ensure it, the microshort should be so small that the electron transfer should be considered more as a diffraction than the usual ballistic passage (Zorin and Likharev, 1978). In the 2-D electron gas formed in GaAs heterojunctions, such a microshort can be apparently formed using lateral confinement; one cannot exclude a possibility that exactly this regime occurs in the above-cited experiments (Meriav *et al.*, 1990). I believe that this question deserves a careful experimental study, which should also answer the following question: is it generally possible to tell tunneling from metallic transfer in such structures? (Mind the discrete-vs-continuous-charge dilemma discussed as a possible criterion of the difference).

The problem becomes even more interesting when we proceed from the virtually-0-D configuration of the point contact to a 1-D system, say a narrow channel formed of the 2-D electron gas by its lateral confinement. These systems are now well known due to the recently discovered effect of quantization of their conductance: $G=n/2R_Q$ (Wharam *et al.*, 1988; van Wees *et al.*, 1988). One cannot exclude a possibility that at somewhat stronger confinement ($G<<1/R_Q$), the system would allow formation of the 1-D Wigner crystal of electrons inside the channel. If now the external electric field could drive the crystal along the channel without melting, we would again arrive at a periodic process with the fundamental frequency (8), i.e. to the SET oscillations! Theoretical analysis of this problem is in progress.

To summarize, single-electronics has eventually become a multi-problem scientific and technical field where each new result provides more questions than answers and which gives nice prospects for creating revolutionary new electronic devices. I believe that it is a very nice field, in particular for young physicists and electronic engineers with their noble desire to learn more about nature and to use this knowledge for speeding the technological advance of the mankind.

ACKNOWLEDGEMENTS

Multiple fruitful discussions of problems of the single-electronics with many colleagues, in particular with members of the Laboratory of Cryoelectronics of the Moscow State University, are gratefully acknowledged. I would also like to thank the University of Wuppertal (where this paper was partly written) and personally Professor Helmut Piel for kind hospitality, and the authors of the cited unpublished papers for sending their preprints prior to publication.

REFERENCES

Averin, D.V., 1986, Zh. Eksp. Teor. Fiz. [Sov. Phys. - JETP] 90:2226.

Averin, D.V., and K.K. Likharev, 1985, Probable Coherent Oscillations at the Single-Electron Tunneling, in: SQUID'85, ed. by H.-D. Hahlbohm and H. Lübiig (W. de Gruyter, Berlin), p. 197.

Averin, D.V., and K.K. Likharev, 1986, J. Low Temp. Phys. 62:345.

Averin, D.V., and K.K. Likharev, 1987, IEEE Trans. Magn. 23:1138.

Averin, D.V., and K.K. Likharev, 1990, Single-Electronics, in: Quantum Effects in Small Disordered Systems, ed. by B. Al'tshuler *et al.* (Elsevier, Amsterdam), to be published.

Averin, D.V., and A.N. Korotkov, 1990, Correlated Single-Electron Tunneling via Mesoscopic Metal Particle: Effects of the Energy Quantization, submitted to J. Low Temp. Phys.

Averin, D.V., and A.A. Odintsov, 1989, Phys. Lett. 140:A251.

Bakhvalov, N.S., G.S. Kazacha, K.K. Likharev, and S.I. Serdyukova, 1990, Single-Electron Solitons in 2-D Arrays of Ultrasmall Tunnel Junctions, submitted to LT-19 (Brighton, August 1990); Physica B, to be published.

Bardeen, J., 1961, Phys. Rev. Lett. 6:57.

Barner, J.B., and S.T. Ruggiero, 1987, Phys. Rev. Lett. 59:807.

Büttiker, M., and R. Landauer, 1982, Phys. Rev Lett. 49:1739.

Caldeira, A.O., and A.J. Leggett, 1983, Ann. Phys. 149:374.

Claeson, T., P. Delsing, and L.S. Kuzmin, 1990, private communication.

Cleland, A.N., J.M. Schmidt, and J. Clarke, 1990, Charge Fluctuations in Small Capacitance Junctions, preprint.

Delsing, P., K.K. Likharev, L.S. Kuzmin, and T. Claeson, 1989a, Phys. Rev. Lett. 63:1180.

Delsing, P., K.K. Likharev, L.S. Kuzmin, and T. Claeson, 1989b, Phys. Rev. Lett. 63:1861.

Dolan, G.J., and J.H. Dunsmuir, 1988, Physica 152:B7.

Dziuba, R.F., and D.B. Sullivan, 1975, IEEE Trans. Magn. 11:716.

Fulton, T.A., and D.J. Dolan, 1987, Phys. Rev. Lett. 59:109.

Geerligs, L.J., and J.E. Mooij, 1988, Physica 152:B212.

Geerligs, L.J., M. Peters, L.E.M. de Groot, A. Verbruggen, and J.E. Mooij, 1989, Phys. Rev. Lett. 63:326.

Geerligs, L.J., V.A. Andregg, P.A.M. Holweg, J.E. Mooij, H. Pothier, D. Esteve, C. Urbina, and M.H. Devoret, 1990, Frequency-locked Turnstile Device for Single Electrons, submitted to Phys. Rev. Lett.

Geigenmüller, U., and G. Schön, 1989, Europhys. Lett. 10:765.

Golub, A.A., 1987, Pis'ma Zh. Eksp. Teor. Fiz. [JETP Lett.] 45:184.

Khromov, I.E., 1986, Solution of an Equation Describing the Single Electron Oscillations, Preprint No. 231, Lebedev Phys. Inst., Moscow (in Russian).

Korotkov, A.N., D.V. Averin, and K.K. Likharev, 1990, Single-Electron Charging of Quantum Wells and Dots, submitted to LT-19 (Brighton, August 1990).

Kulik, I.O., and R.I. Shekhter, 1975, Zh. Eksp. Teor. Fiz. [Sov. Phys. - JETP] 68:623.

Kuzmin, L.S., and K.K. Likharev, 1987a, Pis'ma Zh. Eksp. Teor. Fiz. [JETP Lett.] 45:289.

Kuzmin, L.S., and K.K. Likharev, 1987b, Jpn. J. Appl. Phys. 26 (suppl. 3):1387.

Kuzmin, L.S., and M.A. Safronov, 1988, Pis'ma Zh. Eksp. Teor. Fiz. [JETP Lett.] 48:250.

Kuzmin, L.S., P. Delsing, T. Claeson, and K.K. Likharev, 1989, Phys. Rev. Lett. 62:2539.

Lambe, J., and R.C. Jaklevic, 1969, Phys. Rev. Lett. 22:1371.

Likharev, K.K., 1986, Dynamics of Josephson Junctions and Circuits (Gordon and Breach, New York), Chapter 16.
Likharev, K.K., 1987, IEEE Trans. Magn. 23:1142.
Likharev, K.K., 1988, IBM J. Res. Devel. 32:144.
Likharev, K.K., and V.K. Semenov, 1987, Possible Logic Circuits Based on the Correlated Single-Electron Tunneling in Ultrasmall Junctions, in: Ext. Abstr. of ISEC'87 (Tokyo), p. 182.
Likharev, K.K., and V.K. Semenov, 1990, RSFQ Logic Family, in preparation.
Likharev, K.K., and A.B. Zorin, 1985, J. Low Temp. Phys. 59:347.
Likharev, K.K., N.S. Bakhvalov, G.S. Kazacha, and S.I. Serdyukova, 1989, IEEE Trans. Magn. 25:1436.
Martinis, J.M., and R.L. Kautz, 1989, Phys. Rev. Lett. 63:1507.
Meirav, U., M.A. Kastner, and S.J. Wind, 1990, Single Electron Charging and Periodic Conductance Resonances in GaAs Nanostructures, submitted to Phys. Rev. Lett.
Mullen, K., Y. Gefen, and E. Ben-Jacob, 1988, Physica 152:B172.
Odintsov, A.A., 1988, Zh. Eksp. Teor. Fiz. [Sov. Phys. - JETP] 94:312.
Schön, G., and A.D. Zaikin, 1990, submitted to Phys. Repts.
Van Bentum, P.J.M., R.T.M. Smokers, and H. van Kampen, 1988, Phys. Rev. Lett. 60:2543.
Vasenko, S.A., A.N. Korotkov, and K.K. Likharev, 1990, Noise Properties of the Single-Electron Transistors, in preparation.
Wilkins, R., E. Ben-Jacob, and R.C. Jaklevic, 1989, Phys. Rev. Lett. 63:801.
Wharam, D.A., T.J. Thornton, R. Newbury, M. Pepper, H. Ahmed, J.E.F. Frost, D.G. Hasko, D.C. Peacock, D.A. Ritchie, and G.A. Jones, 1988, J. Phys. 21:CL209.
Zeller, H.R., and I. Giaever, 1969, Phys. Rev. 181:789.
Zorin, A.B., and K.K. Likharev, 1978, J. de Phys. 39 (suppl.):C6-573.
Zorin, A.B., L.S. Kuzmin, and K.K. Likharev, 1990, Two Ways Toward Experimental Observation of the Bloch oscillations in Ultrasmall Josephson Junctions, submitted to LT-19 (Brighton, August 1990).

25

CHARGING EFFECTS AND 'TURNSTILE' CLOCKING OF SINGLE ELECTRONS IN SMALL TUNNEL JUNCTIONS

L.J. Geerligs and J.E. Mooij
Department of Applied Physics, Delft University of Technology
P.O. Box 5046, 2600 GA Delft, The Netherlands

INTRODUCTION

In materials consisting of small metal grains, coupled by tunnel barriers, at low temperatures the electrical properties are strongly influenced by charging effects resulting from the small capacitance of the grains. Because charge is transferred in discrete units (e for normal metal junctions, $2e$ or e for superconducting tunnel junctions), the energy change of the system during tunneling can be significant. If the energy of the system would increase, the tunneling is forbidden at zero temperature. This phenomenon is called Coulomb blockade of (electron) tunneling. Typical energy changes are of order $E_C \equiv e^2/2C$, so that the temperature must be below E_C/k_B to observe charging effects. Already in 1968 this effect of small grain capacitance was appreciated as well as observed experimentally (Giaever and Zeller, 1968, 1969; Lambe and Jaklevic, 1969).

With the advance of submicron lithography it has become possible to artificially produce planar tunnel junctions with capacitance as small as 10^{-16} F, for which charging effects can be observed at liquid helium temperatures (although for most experiments the lower temperatures attainable in a dilution refrigerator are still useful). Many experiments have confirmed the basic theoretical description of these charging effects. We will present this basic theory and the related experiments on artificial tunnel junctions in the second section of this paper. We will not discuss the experiments on granular systems (Kuzmin and Likharev, 1987; Barner and Ruggiero, 1987, van Bentum *et al.*, 1988a, 1988b; Kuzmin and Safronov, 1988; Wilkins *et al.*, 1989) that have also provided convincing confirmation of the basic theory. Especially the recent possibility of using a scanning tunneling microscope on granular material allowed for observation of charging effects at much higher temperatures ($C \approx 10^{-18}$ F, or E_C/k_B > 100 K). However, this configuration is less flexible in device design and control of parameters.

In the third section we will discuss the applicability of small tunnel junctions for practical purposes. As an example, the single electron turnstile that has recently been developed together with the CEN Saclay (Geerligs *et al.*, 1990) will be discussed in more detail, since it shows the possibility of controlling charge transfer at the single electron level.

Coulomb blockade of electron tunneling is not absolute. Passing of an electron through several junctions in one process may be energetically favorable, even if the intermediate states where the electron resides on the electrodes between the junctions, have a high energy. This process is predicted to occur at a rate inversely proportional to the product of the junction resistances. It has been named macroscopic quantum tunneling of the charge, since this tunneling of a single electron corresponds to tunneling of the charge state of the total system through an energy barrier. This extension of the basic theory, which is valid only for high junction tunnel resistance (compared to the resistance quantum h/e^2), will be discussed in the fourth section, together with recent experiments. These results are important for practical applications based on high-resistance junctions, especially single-electronic logic circuits like the turnstile. Finally, in the fifth section we discuss the role of highly dissipative environments, that cause the breakdown of Coulomb blockade in single tunnel junctions or junctions of low tunnel resistance.

In this paper, we will not consider the very interesting physics that occurs in tunnel junctions with superconducting electrodes. There the interaction between Josephson coupling and charging energy, and the resulting competition between charge and phase fluctuations of the junction, gives rise to macroscopic quantum effects. Two reviews, by Averin and Likharev (1990), and Schön and Zaikin (1990) provide recent experimental results as well as a thorough background. We also only mention here that Coulomb blockade effects have probably been observed recently in split-gate confined GaAs-AlGaAs heterostructures (Scott-Thomas *et al.*, 1989; van Houten and Beenakker, 1989; Meirav *et al.*, 1989; Field *et al.*, 1990; Kouwenhoven, 1990; Brown *et al.*, 1990). However, in these systems the description is necessarily more complicated due to e.g. the discreteness of single-particle levels.

The experiments that will be presented have all been performed on aluminum tunnel junctions. These junctions have been brought in the normal state by applying a high magnetic field (typically 2 T). We have found no reason to suspect that the field affects the physics of the Coulomb blockade in a measurable way.

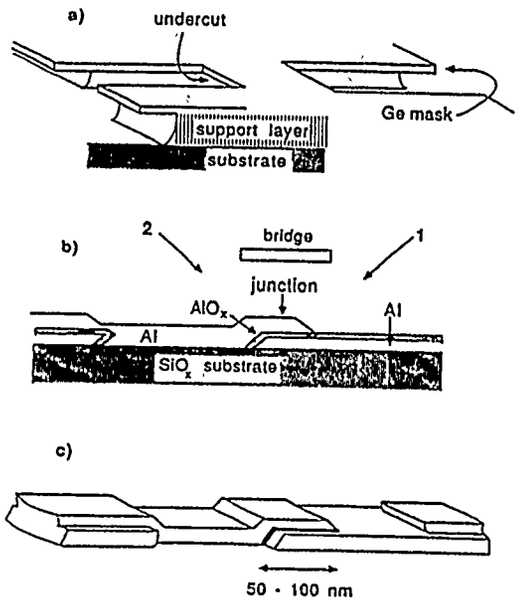


Fig. 1 Processing steps for shadow evaporation of a tunnel junction. (a) Suspended mask. (b) Obligue angle evaporation. (c) The resulting planar junction.

A junction area of $(100\ \text{nm})^2$ yields a capacitance of about 10^{-15} F, depending on the barrier thickness. The smallest planar junctions that have been produced so far (Fulton and Dolan, 1987; Kuzmin *et al.*, 1989; Geerligs *et al.*, 1989) were all fabricated from aluminum. For such a small junction area, useful tunnel resistances (of around 100 kΩ) are obtained if the

aluminum is thermally oxidized at room temperature in oxygen at a pressure of about 1 mbar to create the tunnel barrier. Together with the requirement of high purity metal electrodes this low oxidation pressure means that the total junction be preferably fabricated in one vacuum cycle. This is conventionally done by shadow evaporation (Fig. 1). A mask is suspended at around 200 nm above the (oxidized silicon) substrate. The mask is patterned by conventional submicron lithography. The supporting layer for the mask is an organic material (e.g. resist) that can be undercut by isotropic etching, either wet or with reactive ion etching. The two electrodes of a junction are evaporated from two angles. A mask patterned with a small channel interrupted by a bridge, thus results in a junction because of the interruption of the aluminum strips by the bridge shadow. On both sides of the junction the leads are actually also composed of a double aluminum layer with oxide barrier in between, i.e. the leads are large junctions. This creation of large junctions in series with the small ones can be partly avoided by using a slightly different geometry (Fulton and Dolan, 1987; Kuzmin *et al.*, 1989). A photograph of a two-dimensional array of junctions fabricated in this way is given in Fig. 2. This fabrication procedure has proven to be sufficient for creating junctions with area down to $(30\ nm)^2$. For significantly smaller dimensions, probably new methods have to be developed.

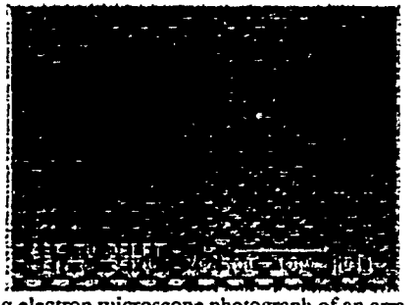


Fig. 2 Scanning electron microscope photograph of an array of small tunnel junctions produced by shadow evaporation.

CLASSICAL THEORY FOR COULOMB BLOCKADE

In this section we consider tunnel junctions with high tunnel resistance, $R_t >> h/e^2$. The charge transport through the junction can then be calculated by treating the charge Q on the junction as a classical variable. The charge can change in a *continuous* way by applying a polarizing voltage to the junction. Trapped charges in the oxide barrier of the junction or in the substrate close to the junction likewise provide the possibility of the junction having any non-integer charge. The junction charge can change stochastically due to tunneling events during which *discrete* charge units are transferred across the barrier.

The rate for a tunneling process is determined by the energy change $\Delta E = E_f - E_i$ during tunneling (Averin and Likharev, 1986):

$$\Gamma(\Delta E,T) = \frac{\Delta E}{e^2 R_t}\left[\exp(\Delta E/k_B T) - 1\right]^{-1} \qquad (1)$$

For $|\Delta E| >> k_B T$:

$$\Gamma(\Delta E) \approx \begin{cases} -\dfrac{\Delta E}{e^2 R_t} & \text{for } \Delta E < 0 \\ 0 & \text{for } \Delta E > 0 \end{cases} \qquad (1a)$$

The relevant energy change is the change in free energy, the sum of the capacitive energies in the system and the work performed by the voltage sources (Likharev, 1988; Bakhvalov *et al.*, 1989):

$$E = \sum_i \frac{Q_i^2}{2C_i} - \sum_j Q_{ij} V_j \qquad (2)$$

The index i denotes summation over tunnel junctions as well as true capacitors, the summation in j is over all voltages sources in the system. Q_{ij} denotes the charge transferred through voltage source V_j. Note that a large stray capacitor on a chip can act as a voltage source and change an experimentally applied current bias for high frequencies into a voltage bias. This is often the case in experiments.

For a circuit consisting only of capacitors and voltage sources, (2) can be reduced to a simpler form for each individual junction (Esteve, 1990). Using Thévenin's rule the circuit to which the junction is coupled is reduced to an equivalent capacitor C_e in series with a voltage source V_e (Fig. 3). In the expression for ΔE, V_e and the charge on C_e cancel so that the energy change during tunneling depends only on junction charge Q and a critical charge Q_c (to be calculated for each junction individually):

$$\Delta E = -\frac{e}{C}(Q - Q_c) \qquad (3)$$

with

$$Q_c = \frac{e}{2}\left(1 + C_e/C\right) \qquad (4)$$

At low temperature, an electron can tunnel only if $|Q| > Q_c$: the junction will show a Coulomb gap (threshold voltage for conduction) of Q_c/C. This concept of a critical charge is useful to calculate the tunneling characteristics of complicated systems subject to charging effects. Here we will use it to consider several simple cases. A single junction biased via a very small capacitor (Büttiker, 1986) will show a Coulomb gap e/2C. Two equal junctions in series each have $Q_c = e/4$. A double junction will therefore show a total Coulomb gap e/2C. Likewise, n equal junctions in series have a Coulomb gap (n-1)e/2C. This Coulomb gap can be influenced by charging the metal islands between the junctions, a possibility that is discussed below.

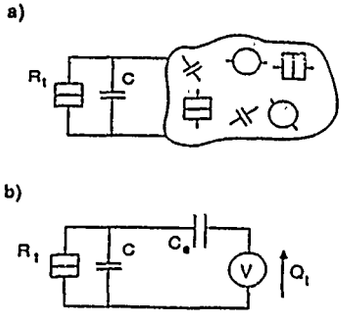


Fig. 3 The reduction of a junction environment consisting of voltage sources and capacitors (a) to an equivalent circuit with one external capacitor C_e and one voltage source. The charge Q_t transferred through the voltage source is relevant for the energy change during tunneling.

First we consider a special case, the current biased single junction (Averin and Likharev, 1986). Since the current bias implies a very good decoupling from the environment, we can put $C_e=0$ so that $Q_e=e/2$. However, in this case the externally applied current I_x induces a smooth time evolution of the charge:

$$\frac{dQ}{dt} = I_x + \left.\frac{dQ}{dt}\right|_{tunneling} \qquad (5)$$

If the current is small compared to e/R_tC a tunneling event will occur at a charge only slightly larger than e/2, changing the charge on the junction to about -e/2. Then it takes a time period e/I_x to recharge the junction for a new tunneling event. At T=0 and small current the resulting dc I-V curve has a parabolic shape:

$$\langle V \rangle = \sqrt{\frac{\pi I_x R_t e}{2C}} \qquad (6)$$

At larger currents the I-V curve approaches a linear form with voltage offset e/2C and slope $1/R_t$. At low currents the tunneling events are correlated in time. The voltage noise spectrum will peak at the Single Electron Tunneling frequency $f_{SET} = I_x/e$ and harmonics. By applying a high-frequency alternating current (frequency f) in addition to the dc current, resonances should occur in the I-V curve at currents I=(h/m)e·f. This has not yet been observed, but a similar phenomenon has been observed in long one-dimensional arrays of tunnel junctions (Delsing *et al.*, 1989b), where for a different reason time correlation of tunneling events also occurs (Likharev *et al.*, 1989; Bakhvalov *et al.*, 1989). In a chain of junctions the current is carried by mutually repulsing charge solitons. A charge soliton consists of a charged metal island between two junctions, together with the associated polarization of the neighboring junctions. Due to the repulsion the charge is transferred in a train of regularly spaced solitons. On a given junction, a tunneling event occurs each time a soliton passes. Therefore the tunneling events are again correlated in time. Delsing *et al.* (1989b) have observed that under high frequency irradiation the I-V curve of such an array shows resonances in the differential resistance at I=e·f and I=2e·f.

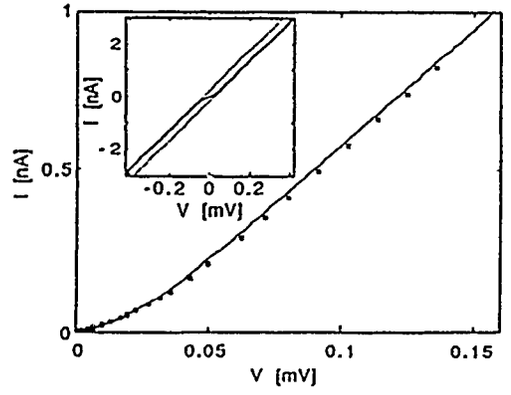


Fig. 4 I-V curve of a small current biased tunnel junction at a temperature of 55 mK. The current bias is possible because the junction is decoupled from the environment by 2-D junction arrays (90 junctions long, 9 wide) in the leads. The junction resistance and capacitance as determined from the I-V curve asymptote (inset) are R_t=132 kΩ and C=2.9 fF. These parameters yield a theoretical prediction for the small-signal I-V curve (open boxes) in good agreement with the measurement.

The ratio of the junction capacitance to the self-capacitance C_0 of the islands between the junctions determines the size of a soliton. The junction charge in a soliton decays as

$$Q = \frac{e}{\sqrt{1 + 4C_0/C}} [1 - \exp(-1/\Lambda)] \exp(-d/\Lambda) \qquad (7)$$

where d is the distance, in number of junctions, from the soliton center (the charged island) and the decay length is given by $\Lambda^{-1}=\text{arccosh}(1 + C_0/2C)$. On a given junction the charge increases in small steps if the soliton approaches, decreases by e if the soliton passes, and after tunneling again increases smoothly in time if one soliton moves away and a new approaches. Therefore chains of tunnel junctions, but also 2-D arrays of junctions (Mooij *et al.*, 1990) can be used to provide a current bias in a single junction. Figure 4 shows the I-V curve of a single junction in a 4-wire measurement. In each lead close to the junction a 90x9 junction array was incorporated to ensure current bias or (for the voltage leads) decouple the junction from the environment. The I-V curve shows the asymptotic linear behavior (inset) from which junction resistance and capacitance can be determined. With these two parameters the experimental I-V curve can be compared to the theory without fitting. The agreement is very good, showing that normal metal junction arrays can indeed provide a good current bias and decoupling from the environment.

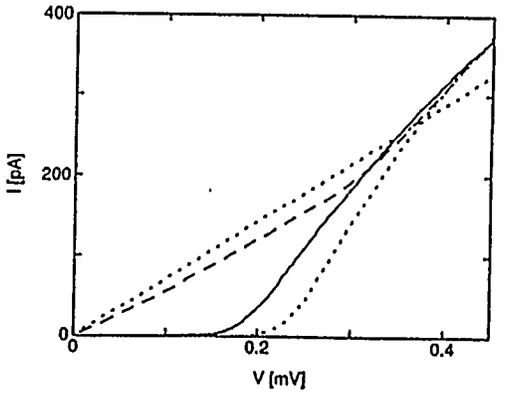


Fig. 5 I-V curve for a double tunnel junction with nominal R_t=345 kΩ and C=0.32 fF at two gate voltages, corresponding to a gate charge 0 (solid curve) and e/2 (dashed curve). The temperature is about 15 mK. Also plotted (dotted curves) are the two corresponding theoretical predictions for the I-V curves at 60 mK. The discrepancy between theory and experiment may be due to several factors, such as imperfect symmetry of the voltage bias or difference between the two junctions.

For the rest of the paper we restrict ourselves to voltage biased systems of two or more junctions (Fulton and Dolan, 1987; Mullen *et al.*, 1988; Ben-Jacob *et al.*, 1988; Likharev, 1988). These are configurations that are easily realized experimentally. In addition they provide po bility for extra control of electron motion. The metal islands between the junctions always have a self-capacitance, i.e. a capacitance to ground. They can also be purposely coupled capacitively to a gate electrode. This provides an extra possibility to charge the junctions, apart from a bias voltage directly applied to the junctions. In a double junction the central metal island can be polarized by a gate voltage, shifting charge from junction capacitance to the gate capacitor. For example, a 'gate charge' C_gV_g=e/2 on the gate capacitor results in a charge ±e/4 for each of the junctions, in addition to the charge CV/2 provided by the bias voltage. Since for these junctions Q_e=e/4, the Coulomb gap is completely suppressed. Figure 5 gives the measured I-V curve for a double junction for the two gate charges where the

Coulomb gap is maximum and minimum. In the case of the maximum Coulomb gap (solid curve) the conduction below the threshold voltage is very low, although not completely zero. We will consider the charge MQT that causes the leakage in this device below. With gate charge the Coulomb gap can be completely suppressed to an almost Ohmic curve (dashed curve). At high voltages the same voltage offset e/2C is recovered. As a function of gate voltage the I-V curve evolves continuously between the two extremes shown. With the average current through the device fixed at a low level, the voltage versus gate voltage can be recorded. An example is given for a similar double junction in Fig. 6. The curve is periodic because gate charges V_g and V_g' are equivalent if $C_g(V_g - V_g') = e$. This is a clear proof of the discrete nature of electron tunneling. If two gate voltages differ by an amount e/C_g one electron tunneling is sufficient to produce the same junction polarization and thus the same I-V curve. At the same time the continuous evolution of the I-V curve as a function of gate voltage proves the possibility of a continuous charging of a tunnel junction.

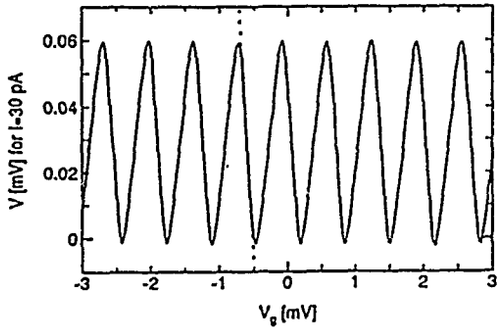


Fig. 6 V-V_g curve for a double tunnel junction with dc current fixed at 30 pA, T≈15 mK. The maximum voltage gain (dV/dV_g) is about 0.35 (slope of dotted curve).

In Fig. 7 we show the current through linear arrays of 2,3 or 5 junctions for a fixed bias voltage, again as a function of gate voltage. Of course this shows the same periodic behavior as the previous figure. For 3 junctions the gate voltage is applied via two gate capacitors to the two islands between the junctions, for 5 junctions via 4 capacitors to 4 islands. Within the main period of e/C_g, a total of n-1 dips can be observed for n junctions. Figure 7 illustrates an important aspect of experiments with gate voltages. Most curves show a minimum in the current which does not occur for the expected zero gate voltage but instead for a seemingly random value. Curve b shows telegraph noise: the current jumps between two positions corresponding to two I-V_g curves which are slightly offset in V_g-direction. The curves for arrays of 5 junctions all differ in their fine structure, whereas theory predicts one pattern for any device of 5 equal junctions. All these results show that the junctions have a random offset charge, probably caused by trapped charges near the junctions. The impossibility to predict even approximately the gate voltage that is necessary to maximize or minimize the Coulomb gap, seriously limits the usefulness of these junctions in large scale integrated applications.

PRACTICAL USE OF COULOMB BLOCKADE OF ELECTRON TUNNELING

Various applications of the Coulomb blockade in small junctions have been proposed (Likharev, 1987, 1988; Yoshikawa *et al.*, 1989). Some possible advantages of these circuits are the extreme integration level, the high speed (the typical operating frequency should be measured in $(R_tC)^{-1}$) and the low dissipation. These aspects will be treated more thoroughly by Likharev in these proceedings. Here we will present some experimental results that give a feeling of the possibilities and problems. As mentioned above one serious problem seems to be the presence of offset charging of junctions by trapped charges. In many applications this may in the future be circumvented by using a resistive gate instead of a gate capacitance (Likharev, 1987). The low temperature necessary to work with junctions with the presently attainable capacitance also forms a limitation. All experiments presented here have been performed in a dilution refrigerator, with the devices at temperatures down to 10 mK. We have found that low pass filtering of the leads to the devices is important. The filters need to be cooled to low temperatures in order to suppress their own thermal noise. The filtering and attenuation of the gate voltage line also turned out to be crucial, especially in the experiments on the turnstile device to be discussed below.

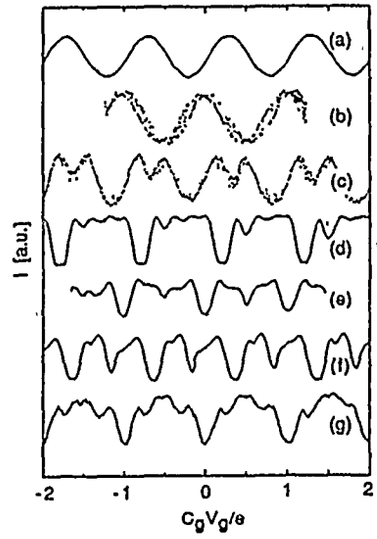


Fig. 7 I-V_g curves for linear arrays of 2 (a,b), 3 (c) and 5 (d-g) junctions at fixed bias voltage, T≈15 mK. (d) and (e) are for the same device with V_g offset by 10 periods. It shows beating due to differing gate capacitances.

Obviously, a double junction is a sensitive detector for charge on the gate electrode. It can be used to count electrons, just as a DC SQUID is used to count flux quanta. Like the SQUID the sensitivity is higher than the electron charge. In preliminary measurements we have found that the gate charge fluctuations corresponding to the measured current noise in curves like Fig. 7(a), is about 10^{-4} $e/\sqrt{Hz}$ between 10 and 200 Hz. Compared to the SQUID a severe problem is the application of the charge to the gate. The input line needs to have a small capacitance compared to C_g. Otherwise much of the charge that should polarize the gate capacitor is lost to the parasitic lead capacitance.

A double junction can also be used as a high quality switch. The difference in resistance of the two states of the device of Fig. 5 is for low voltages almost infinite. Apart from conventional applications in digital circuits, it would be interesting to evaluate the use of such a double junction for experiments on mesoscopic circuits. As an example, with this switch these circuits could perhapse at will be coupled and decoupled from a part of the environment, in one experiment. Similarly, it could be used as a very high impedance voltmeter (by using the gate electrode as the voltage probe) very close to a mesoscopic circuit.

The maximum slope of the V-V_g curve in Fig. 6 is 0.35, corresponding approximately to the ratio of C_g to C. By increasing the gate capacitance to a value larger than the junction capacitance, an amplifying element would be obtained, be it with a very small input voltage.

Other, more complicated circuits have been proposed. Some of those belong to the class of single-electronic devices, such as the memory cell of Yoshikawa *et al.* (1989). In these devices the information is stored not as a voltage but as an extra charge (e.g. one electron). The operation of such devices requires the control of motion of single electrons at high frequencies. That this is indeed possible has recently been shown by the succesful operation of single electron turnstile devices in Delft and Saclay (Geerligs *et al.*, 1990). In these devices, two or more junctions on each side of a central gate capacitor are used to block electron tunneling during part of a clock cycle. The clocking signal consists of a high frequency alternating voltage (added to a dc voltage) applied to the central gate capacitor. Only once per cycle can an electron tunnel across one arm and only once per cycle can it tunnel across the other arm. Coulomb blockade is used to ensure that precisely one electron tunnels. The turnstile creates a very accurate current or charge source.

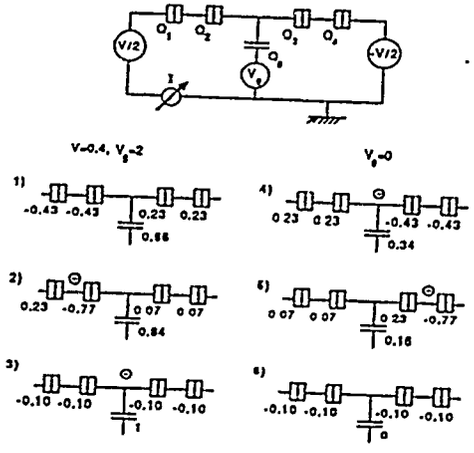


Fig. 8 Working principle of a turnstile for single electrons. An AC plus DC voltage is applied to the central gate between the 4 junctions. The numbers denote consecutive moments in one ac cycle. Junctions are denoted by double boxes. C_g=C/2, hence Q_c=e/3 for all junctions. The momentanous charges are indicated in units of e.

The working principle can be conveniently illustrated using the concept of the critical charge. It is shown in Fig. 8 for a device with four junctions. For simplicity we consider a square-wave gate voltage modulation. The gate capacitance is close to C/2 so that each junction has the same critical charge of e/3. In the first part of the cycle the critical charge is exceeded for the junctions in the left arm but not for those in the right arm. If a tunneling has occurred in one junction, the second will follow almost immediately. If the electron has reached the central island, it will mainly polarize the relatively large gate capacitor and all junction charges will be lower than the critical charge. No other tunneling events occur until the gate voltage is decreased in the second half of the cycle. Then the critical charge is exceeded for the junctions in the right arm but not for those in the left arm (for the gate voltage amplitude within a certain window). Consequently, the electron leaves on the other side of the device. After this no tunneling can happen until the start of the next cycle. In absence of the ac component of the gate voltage no tunneling can happen, i.e. the conduction is zero. Thus, after switching on the ac gate voltage, at each moment in time the passed charge is known up to at most a single electron. As Fig. 9 shows the physical layout is very close to the schematic. An important difference is the addition of small additional gate electrodes to tune out offset charges on the two metal islands in the left and right arm of the device.

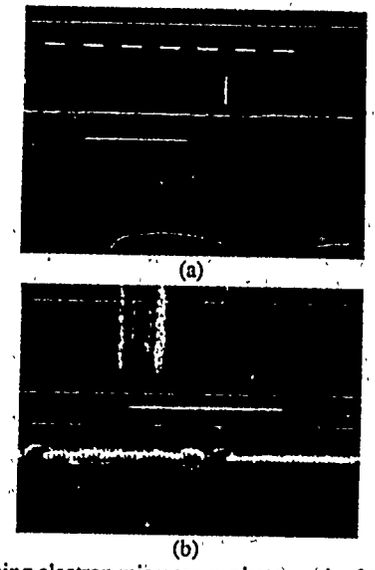


Fig. 9 (a) Scanning electron microscope photograph of the turnstile device as realized with aluminum junctions. The main difference from Fig. 8 is the addition of auxiliary gate electrodes. (b) Enlargement of one arm with two junctions and an auxiliary gate electrode.

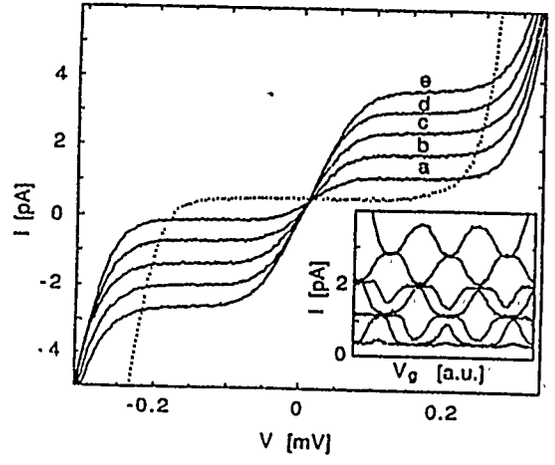


Fig. 10 I-V curves of the turnstile device of Fig. 9 without ac gate voltage (dotted) and with ac gate voltage of frequency 4 to 20 MHz in steps of 4 MHz (a-e). The inset shows the I-V_g curves for an ac gate voltage (5 MHz) of increasing amplitude (top to bottom), taken at a bias voltage of about 0.15 mV. R_t=340 kΩ, C=0.5 fF, C_g=0.3 fF.

Figure 10 shows I-V curves of the device without ac gate voltage (dotted) and with ac gate voltage at a set of frequencies f between 4 and 20 MHz. The zero current Coulomb gap in the absence of the ac voltage (dotted) is lifted to a plateau I=e·f if the ac voltage is applied. The width of the plateaus is dependent on the amplitude of the ac signal but the height is not. To obtain wide flat plateaus it was necessary to tune the auxiliary gate electrodes. However, qualitatively similar I-V curves have been obtained without these gates. The inset of Fig. 10 shows I-V_g curves in the presence of ac signals of various amplitudes (one frequency f=5 MHz) for a bias voltage in the middle of the plateaus. The curves tend to be oscillate between consecutive multiples of e·f. This shows that the device can also pass several electrons per cycle in a controlled way. At higher ac amplitude it is possible to fill the central island with more than one electron. On decreasing the gate voltage these trapped electrons are released one by one through the other arm. It has been possible to obtain quantization (although less accurate) at levels as high as 8e·f (Pothier *et al.*, 1990). Figure 11 gives both the measured and the calculated dependence of the I-V curves on the gate voltage amplitude, with calculations based on (1)-(4). Since the junction capacitance and resistance can be measured from the I-V curve asymptote, and the gate capacitance from the period of gate voltage modulation, no parameters need to be fitted. For the calculations shown, only 1 dB attenuation of the ac signal in the transmission lines and a sample temperature slightly higher than the mixing chamber temperature were assumed. The correspondence of theory and experiment gives convincing evidence that the behavior of circuits of small high resistance tunnel junctions can indeed be very well described by simple theory.

In the experiment, deviations of the current quantization form the relation I=ef were smaller than the accuracy of the current measurement. This amounts to a few fA's or about 0.3 % for the plateaus at frequencies below 20 MHz. For higher frequencies the plateaus do not show a part that is flat within the current noise. The determination of the level of the current plateau as the current at the inflexion point then causes larger errors. Experimental results on the accuracy of the current quantization are given in Table 1. It is easy to estimate the expected intrinsic accuracy of the current quantization in this device from the simple theory. To obtain a high accuracy of the relation I=ef, the ac cycle should last long enough to let tunneling to and from the central island happen with high probability, i.e. f must be much smaller than $(R_tC)^{-1}$ to avoid cycles being lost. On the other hand an electron trapped on the central electrode should have a negligible probability to escape by a thermally assisted transfer. At finite temperature there is a trade-off between the two requirements: a thermally assisted escape will be more probable for lower frequencies. For an electron transfer in the situation shown in Fig. 8 the first tunnel event of each half of the cycle ($\Delta E=0.1e^2/C$) can occur in two junctions with a rate $\Gamma=(10R_tC)^{-1}$. For a square wave modulation this yields a probability to miss a cycle of about $\exp(-\Gamma/f)=\exp(-1/10fR_tC)$. For the device used in the experiments, $(R_tC)^{-1}\approx 5$ GHz, so at 5 MHz this probability is $\exp(-100)\approx 10^{-44}$, while at 50 MHz it is already about 10^{-5}. Obviously, the required accuracy puts an upper limit to the allowed frequency. To estimate the effect of thermal fluctuations, we compare the rate for unwanted tunneling events, $\tilde{\Gamma}$, with the one for favorable events, Γ. From (1) we find that the ratio is of order $\exp(-\Delta E/k_BT)$. For an accuracy of e.g. 10^{-8}, it is necessary to have $\tilde{\Gamma}/f\approx 10^{-8}$, which combined with the requirement $\Gamma/f=10^3$ yields $\exp(-\Delta E/k_BT)=10^{-11}$, or $k_BT\approx\Delta E/25$. Since typically ΔE is on the order of $0.1e^2/C$, for the present device this corresponds to temperatures of about 15 mK. Comparable problems with unwanted transitions could arise from insufficient screening from noise and interference in the experiments. The simulations in Fig. 11 suggest that in the present experiment these disturbances seem to be described well by a temperature of not more than 50 mK, which is already close to the temperature requirement derived above. More careful screening is possible. These limitations are relaxed by the use of smaller junctions. For junctions of 0.1 fF with the same resistance, the requirement that $f<10^{-3}/R_tC$ corresponds to f<30 MHz and $k_BT<0.1e^2/C$ to T<75 mK.

A third cause for accuracy degradation is the already mentioned macroscopic quantum tunneling (MQT) of the charge. This amounts to the escape (at zero temperature) of a trapped charge on the central electrode, through both junctions in one event. The rate is proportional to the product of the junction conductances. The addition of junctions to each arm of the device would decrease the rate of this process. This addition would not significantly increase the chance of cycles being lost, since the tunneling of an electron through the wings is an avalanche process, where the first tunnel event takes most of the time. Below we will consider charge MQT more detailed and present experiments that confirm the higher order perturbative description. With this description, if we denote the rate for unwanted transitions again by $\tilde{\Gamma}$, it turns out that $\tilde{\Gamma}/f=10^{-8}$ together with $\Gamma/f=10^3$ corresponds to the approximate condition $(R_t/R_Q)^{n-1} < 10^{13-n}$ where n is the number of junctions in each arm and $R_Q=h/4e^2\approx 6.5$ kΩ. This is e.g. fulfilled for wings of 5 junctions of $R_t\approx 650$ kΩ.

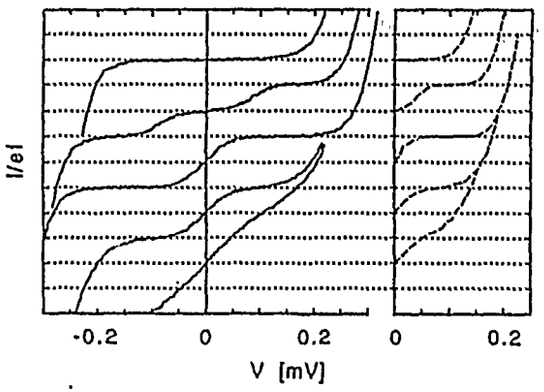


Fig. 11 I-V curves at f=5 MHz for different amplitudes of ac gate voltage. The dotted horizontal lines are at intervals ef=0.80 pA. From top to bottom the calculated ac voltage amplitudes at the sample are 0, 0.60, 0.95, 1.50 and 1.89, in units of e/C=0.30 mV. On the right, the corresponding simulated I-V curves (at 50 and 75 mK) are shown as dashed lines, with 1 dB extra attenuation assumed.

Table 1. Accuracy of the current quantization in the turnstile device of Fig. 10. The measured current plateau I_S is compared with the relation I_S=ef. σ_m is the standard deviation of I_S, as determined from averaging about 50 data points, well inside the current plateau.

f (fA)	I_S (MHz)	σ_m (fA)	ef-I_S (fA)
4.012	635	2	8
6.011	967	2	-4
8.031	1287	2	0
10.040	1610	2	-1
12.029	1930	2	-3
14.028	2243	2	5
16.026	2560	3	7
18.063	2890	3	4
20.011	3196	3	10
30.036	4856	3	-44

Obviously, for single electron logic applications, an error probability of 10^{-8} per time period of e.g. $(10\text{ MHz})^{-1}$ is not sufficient to avoid errors. Error correction logic seems in that case necessary for practical application.

PERTURBATIVE CORRECTION TO THE CLASSICAL THEORY FOR COULOMB BLOCKADE

In an array of junctions with low bias voltage, electrons residing on the central metal islands increase the energy of the total system. This produces a barrier for electron transport across the system. Thermal fluctuations of the charge on the junctions can cause passage of this barrier. At low temperatures electron transport is exponentially (in E_C/k_BT) suppressed, giving rise to the Coulomb gap and the possibility of trapping an electron in the turnstile. However, quantum fluctuations of the charge can cause the system to change the charge distribution to a state where one electron charge has passed through the complete array, a macroscopic quantum tunneling process. Effectively virtual tunneling events have occurred to the intermediate forbidden states. It need not be the same electron that crosses the various junctions. Indeed it is assumed that there is no contribution from interference terms of the two transition amplitudes. Averin and Odintsov (1989) have shown that for high tunnel resistances the rate for this process is of higher order in the junction conductances:

$$\Gamma \propto \frac{(2\pi)^2}{h} \prod_i \frac{\alpha_i}{\pi^2} \qquad (8)$$

where $\alpha=R_q/R_t$, $R_q=h/4e^2$, and i indexes the junctions. For a double junction,

$$\Gamma = \frac{h}{(2\pi)^2 e^4 R_{t1} R_{t2}} \left\{ \left(1 + \frac{2}{eV} \frac{E_1 E_2}{E_1+E_2+eV}\right)\left[\sum_{i=1,2} \ln\left(1+\frac{eV}{E_i}\right)\right] - 2 \right\} eV \qquad (9)$$

E_1 and E_2 are the energies of the (virtual) intermediate state if the first tunnel event occurs in the left junction and the right junction, respectively;

$$E_1 = \frac{e}{2C_\Sigma}\left(\frac{e}{2} + Q_0 - (C_2 + C_g)V\right) \qquad (10a)$$

$$E_2 = \frac{e}{2C_\Sigma}\left(\frac{e}{2} - Q_0 - (C_1 + C_g)V\right) \qquad (10b)$$

$$C_\Sigma = C_1 + C_2 + C_g \qquad (10)$$

Q_0 is the trapped or induced charge on the central electrode; $Q_0 \approx C_g V_g$ in the absence of offset charges. For small voltages the I-V curve of a Coulomb gap subject to this quantum leakage is predicted to be cubic: $I \propto V^3$. This should experimentally be well distinguishable from the exponential current decay if only thermal fluctuations were important.

In Fig. 12 we compare measurements of the I-V curves of 4 double junctions (with maximum Coulomb gap, so $Q_0 \approx 0$) with the theoretical prediction from the classical theory and charge MQT respectively. The measurements have been scaled to dimensionless voltage VC/e and current IR_tC/e to allow for easy comparison. To perform this scaling the asymptotes of the measured I-V curves have been used to determine R_t and C. The classical theory predicts scaled I-V curves equal for all devices, the charge MQT theory predicts more current in devices with lower R_t. It is clear that the measurements follow this prediction from charge MQT. Moreover, for the two highest resistance devices, the measured I-V curve is in good agreement with the prediction from charge MQT. Only for the lowest two resistance devices significantly more current is measured than is predicted. Perhaps for these tunnel resistances smaller than 100 kΩ the perturbative approach yielding (9) already breaks down. To obtain rough agreement with the predictions from thermal fluctuations in (1)-(4), it is necessary to introduce some *ad hoc* corrections. A high temperature of 125 mK is used to obtain a curve that is at least in the range of the measurements. In addition it is necessary to assume a misadjustment of the gate charge, that is systematically larger for low resistance samples.

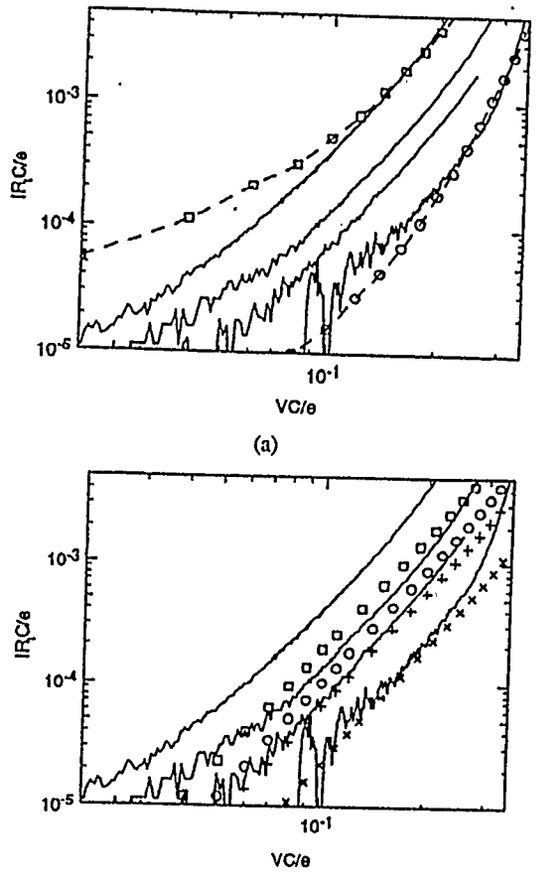


Fig. 12 I-V curves of 4 double junction devices, with C≈0.35 fF, C_g≈0.03 fF and R_t=42, 77, 119 and 345 kΩ (top left curve to bottom right curve). The gate charge was adjusted to 0. (a) Comparison with the classical theory [tunneling rate according (1)] for $k_BT=0.02e^2/C \approx 125$ mK. Open circles: no gate charge; open boxes: gate charge of 0.2e. (b) Comparison with charge MQT theory [tunneling rate according to (9)]. The prediction is dependent on R_t; boxes: R_t=42 kΩ, circles: 77 kΩ, plusses: 119 kΩ and crosses 345 kΩ.

In addition to this quantitative support for the charge MQT model, we have found that I-V curves for linear arrays of e.g. five junctions are considerably sharper (i.e. they display less leakage current) than the I-V curves of the corresponding double junctions. Quantum leakage is therefore a relevant factor in the description of devices based on tunnel junctions with realistic values of R_t (below or of order of 1 MΩ). As already mentioned by Averin and Odintsov (1989), for single-electron logic circuits like the turnstile it is therefore advisable to use more than two junctions to block electron tunneling reliably.

QUANTUM CHARGE FLUCTUATIONS IN A NON-PERTURBATIVE APPROACH

A systematic approach to the description of tunneling in small junctions has been developed on the basis of microscopic theory (Ambegoakar *et al.*, 1982; Ben-Jacob *et al.*, 1983; Eckern *et al.*, 1984). With this technique high tunnel conductances and strong coupling to a dissipative environment, both giving rise to quantum charge fluctuations on the junctions, can be treated in principle. An effective action can be obtained in which all microscopic degrees of freedom have been traced out and only the macroscopic degrees of freedom, the junction charge Q and a generalized phase difference $\varphi = \int(e/h)V dt$, remain. In this section we will consider the effect of quantum charge fluctuations on the I-V curves, especially the conductance in the linear response regime of junctions with low R_t. The effect of the environment, causing a strong suppression of the Coulomb gap in single junctions, will also be shortly considered.

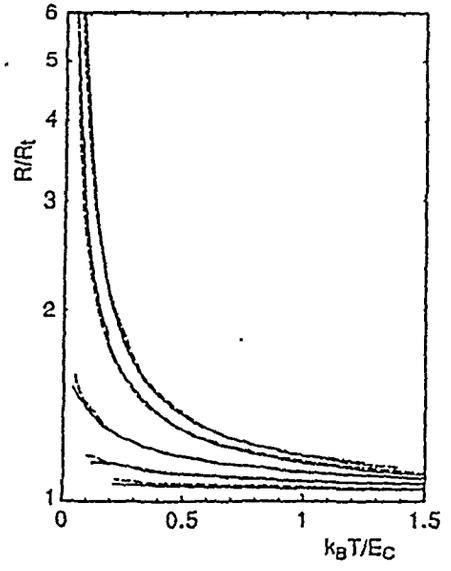


Fig. 13 Resistance versus temperature in the linear response regime for linear junction arrays. The increase of resistance due to Coulomb blockade is suppressed for low R_t. The solid curves are the measurements, the dashed curves give the theoretical predictions from Brown and Simánek (1986) for the respective junction resistances. *From top to bottom* the devices are characterized by: 2 junctions, R_t=82 kΩ, C=1.1 fF; 5 junctions, 5.4 kΩ, 1.0 fF; 10 junctions, 1.3 kΩ, 0.8 fF; 10 junctions, 0.52 kΩ, 2.3 fF; 10 junctions, 0.24 kΩ, 3.7 fF.

For low R_t the charge on a junction is no longer a well defined quantity. Qualitatively one might say that the electronic wavefunctions leak too strongly through the barrier. Brown and Simánek (1986) could obtain a closed-form expression for the conductance of a tunnel junction for arbitrary R_t. In a variational approach they replaced the effective action with a new one with effective ohmic dissipation. The current-current time correlation function was used to calculate the junction resistance with the Kubo formula, for arbitrary temperature. Also Odintsov (1988) replaced quasiparticle dissipation by an effective Ohmic one to calculate the I-V curve for a very low R_t junction. Here we compare measurements of the linear response of junctions arrays with the Brown and Simánek theory. The use of arrays is necessary to fix the junction capacitance and exclude the parasitic capacitance of the leads. In Fig. 13 we give for several arrays the resistance normalized to the tunnel resistance as a function of temperature. R_t varies between 0.24 and 82 kΩ. The sharp (actually exponential) resistance increase at low temperature for the high resistance sample is strongly suppressed in the low resistance samples. The agreement with the theory from Brown and Simánek is good if the capacitance is allowed to be used as a fit parameter. In principle it should be equal to the capacitance determined from the asymptote of the I-V curve. For the devices with R_t>1 kΩ the fitting adjustment is within a factor 2. At low temperatures the measured R(T) curves for high R_t samples are dependent on the gate voltage. In this case the fit with theory is less satisfactory.

Much attention is being given at this moment to the influence of the environment on Coulomb blockade. Generally, the capacitance between leads to a single junction is very large compared to the junction parallel plate capacitance. This results in the absence of a Coulomb gap in a single junction without special precautions (Delsing *et al.* 1989a, Geerligs *et al.* 1989). By using high impedance leads the effect of parasitic lead capacitance can be effectively avoided. One possibility to realize high impedance leads is to use arrays of junctions (Fig. 4, see also Delsing *et al.*, 1989a). Cleland *et al.* (1990) showed that leads in the form of narrow strips of high sheet resistance material also cause a clear Coulomb gap to appear in single junctions. Although these may seem trivial results, until recently there was some controversy about this subject. It was argued by Büttiker and Landauer (1986) and supported by van Bentum *et al.* (1988a) that due to the short tunneling time a Coulomb gap should be observed also in a single junction. The reasoning is that only the capacitance within a small radius given by the product of speed of electromagnetic field and electron barrier traversal time (approximately $(10^8\ ms^{-1})\cdot(10^{-15}\ s)\approx 100$ nm) can contribute to the capacitance for charging effects—a 'relativistic horizon' argument (Geigenmüller and Schön, 1989). Recently this problem has been treated by various authors (Nazarov, 1989a, 1989b; Devoret *et al.*, 1990; Cleland *et al.*, 1990; Averin and Schön, 1990). The results have been put in different words but amount to similar physics. The models consider the influence of an arbitrary frequency dependent shunt impedance on the tunneling process in a junction. The electromagnetic field in the shunt geometry is influenced by but also has a backinfluence on the tunneling process. Devoret *et al.* showed that a Coulomb gap is due to inelastic tunneling; i.e. it arises if during tunneling low frequency modes can be excited in the environment. If only elastic tunneling is possible (due to prevalence of high energy modes in a low impedance environment) no Coulomb gap arises. These authors as well as Cleland *et al.* showed that the Coulomb gap in a single junction can also be understood to be washed out by quantum fluctuations in the environment. Calculating these fluctuations from the fluctuation-dissipation theorem for a well controlled experimental impedance Cleland *et al.* found agreement between theory and observed suppression of the Coulomb gap.

The solution to the relativistic horizon paradox is that two timescales are important in this problem. In addition to the Büttiker-Landauer traversal time the inverse of the energy change during tunneling is important (following from the energy-time uncertainty). The important one is the longer of the two, which is for not too high voltages in these experiments the latter one. Thus for high voltages (where this time is shortened or, in the alternative Devoret formalism, inelastic tunneling is possible), it is expected that the Coulomb gap appears. This has not been conclusively observed. An alternative way to cut off the effective tunneling time is to increase the current to a level where the mean time between tunneling events - e/I - becomes short. We have observed a Coulomb gap in single junctions (Fig. 14) with the current rather than the voltage apparently being the relevant quantity.

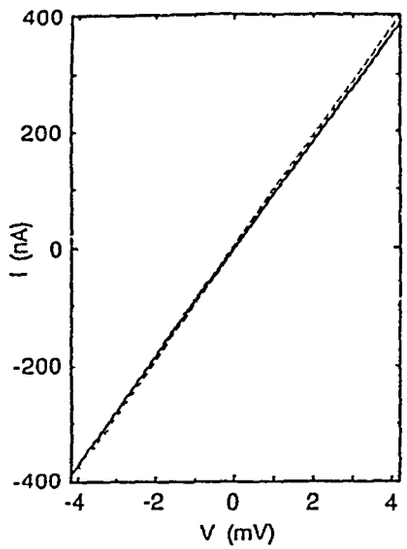


Fig. 14 I-V curve of a single junction at high current. The asymptotes (dashed lines) yield R_t=10.7 kΩ and C=1.3 fF, the capacitance being in agreement with the 0.01 μm^2 junction area.

ACKNOWLEDGEMENTS

In Delft, the following physicists have contributed to experiments and discussions: V. Anderegg, M. Peters, H. van der Zant, B. van Wees and L. Kouwenhoven from the experimental group and U. Geigenmüller, D. Averin (during his visit) and G. Schön from the theoretical group. It is a pleasure to acknowledge the cooperation and discussions with the Groupe Quantronique of the CEN in Gif-Sur-Yvette: H. Pothier, C. Urbina, M. Devoret, and especially D. Esteve with his brilliant idea for the turnstile. Finally, discussions with K. Likharev were always extremely educational. This work was supported by the Dutch Foundation for Fundamental Research on Matter (FOM). The CEA is acknowledged for hospitality during several visits. Part of the lithography was performed at the Delft Institue of Microelectronics and Submicron Technology (DIMES).

REFERENCES

Ambegaokar, V., Eckern, U., and Schön, G., 1982, *Phys. Rev. Lett.*, 48:1745.

Averin, D. V., and Likharev, K. K., 1986, *J. Low Temp. Phys.*, 62:345.

Averin, D. V., and Odintsov, A.A., 1989, *Phys. Lett.*, 140:A251.

Averin, D. V., and Likharev, K. K., 1990, "Single-Electronics", to be published in "Quantum Effects in Small Disordered Systems", eds. B.L. Al'tshuler, P.A. Lee and R.A. Webb (Elsevier, Amsterdam).

Averin, D. V., and Schön, G., 1990, to be published in the proceedings of the NATO Advanced Study Institute on Quantum Coherence in Mesoscopic Systems, Les Arcs, April 1990.

Bakhvalov, N. S., Kazacha, G. S., Likharev, K. K., and Serdyukova, S. I., 1989, *Zh. Eksp. Teor. Fiz.*, 95:1010 [*Sov. Phys. JETP*, 68:581].

Barner, J. B., and Ruggiero, S. T., 1987, *Phys. Rev. Lett.*, 59:807.

Ben-Jacob, E., Mottola, E., and Schön, G., 1983, *Phys. Rev. Lett.*, 51:2064.

Ben-Jacob, E., Mullen, K., Gefen, Y., and Schuss, Z., 1988, *Phys. Rev.B*, 37:7400.

Brown, R., and Simánek, E., 1986, *Phys. Rev. B*, 34:2957.

Brown, R. J., Pepper, M., Ahmed, H., Hasko, D. G., Ritchie, D. A., Frost, J. E. F., Peacock, D. C., and Jones, G. A. C., 1990, *J. Phys. C*, 2:2105.

Büttiker, M., 1986, *Physica Scripta* T14:82; *Phys. Rev. B* 36:3548.

Büttiker, M., and Landauer, R., 1986, *IBM J. Res. Develop.* 30:451.

Cleland, A. N., Schmidt, J. M., and Clarke, J., 1990, *Phys. Rev. Lett.* 64:1565.

Delsing, P., Likharev, K. K., Kuzmin, L. S., and Claeson, T., 1989a, *Phys. Rev. Lett.* 63:1180.

Delsing, P., Likharev, K. K., Kuzmin, L. S., and Claeson, T., 1989b, *Phys. Rev. Lett.* 63:1861.

Devoret, M. H., Esteve, D., Grabert, H., Ingold, G.-L., Pothier, H., and Urbina, C., 1990, *Phys. Rev. Lett.* 64:1824.

Eckern, U., Schön, G., and Ambegaokar, V., 1984, *Phys. Rev. B* 30:6419.

Esteve, D., 1990, private communication.

Field, S. B., Kastner, M. A., Meirav, U., Scott-Thomas, J. H. F., Antoniadis, D. A., Smith, H. I., and Wind, S. J., 1990, Conductance oscillations periodic in the density of one-dimensional electron gases, preprint.

Fulton, T. A., and Dolan, G. J., 1987, *Phys. Rev. Lett.* 59:109.

Geerligs, L. J., Anderegg, V. F., van der Jeugd, C. A., Romijn, J., and Mooij, J. E., 1989, *Europhys. Lett.* 10:79.

Geerligs, L. J., Anderegg, V. F., Holweg, P. A. M., Mooij, J. E., Pothier, H., Esteve, D., Urbina, C., and Devoret, M. H., 1990, Frequency-locked turnstile device for single electrons, to be published in *Phys. Rev. Lett.*

Geigenmüller, U., and Schön, G., 1989, *Europhys. Lett.* 10:765.

Giaever, I., and Zeller, H. R., 1968, *Phys. Rev. Lett.* 20:1504.

Kouwenhoven, L., 1990, unpublished, observed a Coulomb staircase and field-independent conductance oscillations in a chain of quantum dots.

Kuzmin, L. S., and Likharev, K. K., 1987, *Pis'ma Zh. Teor. Eksp. Fiz.* 45:389 [*JETP Lett.* 45:495].

Kuzmin, L. S., and Safronov, M. V., 1988, *Pis'ma Zh. Teor. Eksp. Fiz.* 48:250 [*JETP Lett.* 48:272]

Kuzmin, L. S., Delsing, P., Claeson, T., and Likharev, K. K., 1989, *Phys. Rev. Lett.* 62:2539.

Lambe, J., and Jaklevic, R. C., 1969, *Phys. Rev. Lett.* 22:1371.

Likharev, K. K., 1987, *IEEE Trans. Magn.* 23:1142.

Likharev, K. K., 1988, *IBM J. Res. Develop.* 32:144.

Likharev, K. K., Bakhvalov, N. S., Kazacha, G. S., and Serdyukova, S. I., 1989, *IEEE Trans. Magn.* 25:1436.

Meirav, U., Kastner, M. A., Heiblum, M., and Wind, S. J., 1989, *Phys. Rev. B* 40:5871.

Mooij, J. E., van Wees, B. J., Geerligs, L. J., Peters, M., Fazio, R., and Schön, G., 1990, Charge-anticharge unbinding in two-dimensional arrays of tunnel junctions, submitted to *Phys. Rev. Lett.*

Mullen, K., Ben-Jacob, E., Jaklevic, R. C., and Schuss, Z., 1988, *Phys. Rev. B* 37:98.

Nazarov, Yu. V., 1989a, *Pis'ma Zh. Eskp. Teor. Fiz.* 49:105 [*JETP Lett.* 49:126].

Nazarov, Yu. V., 1989b, *Zh. Eskp. Teor. Fiz.* 95:975 [*Sov. Phys. JETP* 68:561].

Odintsov, A. A., 1988, *Zh. Eksp. Teor. Fiz.* 94:312 [*Sov. Phys. JETP* 67:1265].

Pothier, H., Esteve, D., Urbina, C., Orfila, P. F., Devoret, M., Geerligs, L. J., Anderegg, V. F., Holweg, P. A. M., and Mooij, J. E., 1990, internal report CEN Saclay. They also consider more detailed the explanation of the V-V_g curves.

Schön, G., and Zaikin, A. D., 1990, to be published in *Physics Reports*.

Scott-Thomas, J. H. F., Field, S. B., Kastner, M. A., Smith, H. I., and Antoniadis, D. A.,

1989, *Phys. Rev. Lett.* 62:583.
van Bentum, P. J . M., van Kempen, H., van de Leemput, L. E. C., and Teunissen, P. A. A., 1988a, *Phys. Rev. Lett.* 60:369.
van Bentum, P. J. M., Smokers, R. T. M., and van Kempen, H., 1988b, *Phys. Rev. Lett.* 60:2543.
van Houten, H., and Beenakker, C. W. J., 1989, *Phys. Rev. Lett.* 63:1893.
Wilkins, R., Ben-Jacob, E., and Jaklevic, R. C., 1989, *Phys. Rev. Lett.* 63:801.
Yoshihiro, K., Kinoshita, J., Inagaki, K., Yamanouchi, C., Kobayashi, S., and Karasawa, T., 1987, *Jpn. J. Appl. Phys.* 26:1379.
Yoshikawa, N., Sugahara, M., and Murakami, T., 1989, *IEEE Trans. Magn.* 25:1286.
Zeller, H.R., and Giaever, I., 1969, *Phys. Rev.* 181:789.

27

MONTE CARLO ALGORITHMS FOR QUANTUM TRANSPORT

Lino Reggiani, Patrizia Poli and Lucio Rota

Dipartimento di Fisica e Centro Interuniversitario di Struttura della Materia
Universita' di Modena, Via Campi 213/A, 41100 Modena, Italy

INTRODUCTION

The fast development in the field of submicron devices has provided a renewed interest in the theory of electron transport beyond the free-particle approach based on the semi-classical (SC) Boltzmann equation. Indeed, quantum theory has indicated that a proper treatment of high-field transport in semiconductors should include the intra-collisional field effect (ICFE) as well as collisional broadening (CB) (for a review see Reggiani, 1985).

The intra-collisional field effect accounts for the presence of the electric field in the collision operator of the kinetic equation. In other words, a scattering event does not occur between states described by the plane waves of a free electron, but between those of an electron in the field. Introduced by Levinson and Yasevichyute (1972), ICFE has subsequently been investigated by many researchers (Barker, 1978; Thornber, 1978; Herbert *et al.*, 1982; Pottier *et al.*, 1982; Seminozhenko, 1982; Marsh *et al.*, 1984; Lowe, 1985; Sarker *et al.*, 1986; Khan *et al.*, 1987).

Collisional broadening accounts for the finite lifetime of the interacting (impurities, phonons, etc) carriers. Thus, there is no unique relation between the carrier energy $\hbar\omega$ and the wave vector k in contrast to the effective mass model where $\hbar\omega = \hbar^2k^2/2m$ (m is the effective mass in units of the freee electron mass). Instead, $\hbar\omega$ and k should be taken as independent variables and a spectral function $A(\mathbf{k}, \omega)$, which describes the relationship between them (Mahan, 1972, 1981), is introduced. The effective mass model is recovered when $A(\mathbf{k}, \omega) = 2\pi\delta(\hbar\omega - \hbar^2k^2/2m)$. In the context of high-field transport in semiconductors CB was described in detail by Barker (1972) and its importance within a Monte Carlo (MC) simulation was first pointed out by Capasso *et al.* (1981).

Since then, several research groups (Tang *et al.*, 1981; Chang *et al.*, 1983; Tang *et al.*, 1983; Brennan *et al.*, 1984a, 1984b; Brorson *et al.*, 1985; Artaki *et al.*, 1985; Porod *et al.*, 1985; Lugli *et al.*, 1986; Reggiani *et al.*, 1987, 1988; Kim *et al.*, 1987; Brunetti *et al.*, 1989; Abdolsalami *et al.*, 1990) have attempted to estimate the importance of these quantum features by suitable generalizations of standard MC algorithms using a full k space representation.

The main objective of this paper is to present an alternative description of ICFE and CB which uses a total energy scheme. Accordingly, the total carrier energy along the field is represented as the sum of its kinetic and potential contribution. This enables us to construct a spectral density which simultaneously accounts for ICFE and CB. Then, for a simple semiconductor model, we devise an original MC procedure which gives a quantitative evaluation of the above effects on the carrier energy distribution function under stationary and space homogeneous conditions. We remark that, at low electric fields, the spectral density reduces to the well known delta function behavior when ICFE or CB are neglected. Therefore, the present scheme yields the correct SC Boltzmann limit.

In the following section, we describe the general features of the spectral density to be used. Then, we present an original MC procedure which, when applied to a simple semiconductor model, is able to include both ICFE and CB. The fourth section displays the numerical results obtained, while the last section draws the main conclusions of the present approach.

THEORY

Following the recent results of Bertoncini *et al.* (1989; 1990) we consider a spectral density which accounts simultaneously for ICFE and CB. To this end, we work in a mixed representation defined by the Airy transform space-variable s along the direction z of the electric field E and the Fourier transform wavevector-variable k_t along the transverse direction. We remark that the variable $s = \varepsilon_z/eE$ has a physical interpretation as the turning point in z of a non-coupled electron with kinetic energy $\varepsilon(k_t, s) = \hbar^2k_t^2/2m + eEs$. In the limit of vanishing fields we have $E \to 0$, $s \to \infty$, and $\varepsilon_z = \hbar^2k_z^2/2m$. Then, by considering non-polar optical emission processes as the only scattering mechanism at zero temperature (spontaneuous emission), the spectral density $A(x_t, x_s; x_\omega)$ and the scattering rate $P(\zeta)$ can be given, respectively, the following expressions (Bertoncini *et al.*, 1989):

$$A(x_t, x_s; x_\omega) = -\frac{2}{\gamma^2}\frac{Im\{S\}}{(x_\omega - x_t - x_s - Re\{S\})^2 + (Im\{S\})^2}, \tag{1}$$

$$P(\zeta) = -\frac{2}{\hbar}\gamma^2 Im\{S\}, \tag{2}$$

with

$$Re\{S(\zeta)\} = -x_\Theta^{1/2}\left[Ai'(\zeta)Bi'(\zeta) - \zeta Ai(\zeta)Bi(\zeta) + \frac{\zeta^{1/2}}{\pi}\theta(\zeta)\right], \tag{3}$$

$$Im\{S(\zeta)\} = -x_\Theta^{1/2}\left[Ai'^2(\zeta) - \zeta Ai^2(\zeta)\right], \tag{4}$$

where

$$x_\omega = \frac{\hbar\omega}{\gamma^2}, \qquad x_t = \frac{\varepsilon(k_t)}{\gamma^2}, \qquad x_s = \frac{eEs}{\gamma^2},$$

$$x_o = \frac{\hbar\omega_o}{\gamma^2}, \qquad x_\Theta = \frac{\Theta}{\gamma^2}, \qquad \zeta = \frac{x_s - x_\omega + x_o}{x_\Theta},$$

$$\Theta = \left[\frac{3(e\hbar E)^2}{2m}\right]^{1/3}, \qquad \gamma^2 = \left[\frac{(D_tK)^2m^{3/2}}{2^{5/2}\pi\rho\hbar^2\omega_o}\right]$$

are the total energy, transverse energy, longitudinal energy, optical phonon energy, auxiliary energy from the electric field, reduced parameter for the Airy functions, the field length, and scattering strength, respectively, in normalized units, and $\theta(\zeta)$ is the Heaviside step function (Reggiani *et al.*, 1988). In this framework the spectral density, and thus the transport

properties, for a given field depend on 3 parameters, namely: the interaction energy γ^2, the optical phonon energy $\hbar\omega_0$, and the effective mass m.

An interesting limit of (1) is the case when ICFE is neglected, i.e. when only CB is considered, and the spectral density A^{CB} assumes the expression (Reggiani *et al.*, 1988):

$$A^{CB}(x_t; x_\omega) = \frac{2}{\gamma^2} \frac{(x_\omega - x_0)^{1/2}}{(x_\omega - x_k)^2 + (x_\omega - x_0)} \tag{5}$$

with $x_k = \hbar^2 k^2 / 2m\gamma^2$. By letting $\gamma \to 0$ in (5) the free electron case A^{FREE} is recovered:

$$A^{FREE}(x_t; x_\omega) = \frac{2\pi}{\gamma^2} \delta(x_k - x_\omega) . \tag{6}$$

The shape of the spectral density in (1) is shown in Figs. 1 and 2 for different values of the variables. Figure 1 reports the spectral density as a function of $x_\omega - x_0$ for the case when x_t and x_s are given. We notice a typical double-peak structure separating the positive from the negative values of $x_\omega - x_0$. The broadening and the peak in the positive region of $x_\omega - x_0$ are basically due to CB, while the same features in the negative region of $x_\omega - x_0$ are associated with ICFE. The tail in the negative energy region is related to the renormalization of the quasi-particle energy associated with the presence of the electric field, a result well known in electro-optical studies (Franz-Keldish effect). We remark that, by increasing the field, the broadening smooths out, and a single-peak structure due to CB becomes the dominant one. This reflects the result that ICFE is mostly effective in the low energy region of the quasi-particle.

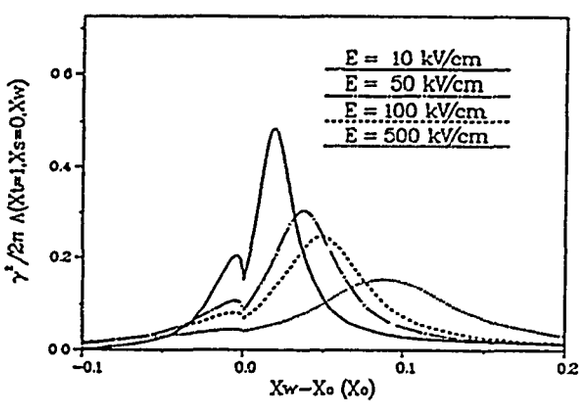


Fig. 1 Spectral density of the quasi-particle as a function of the total energy for given transverse and longitudinal kinetic energies at different electric fields. Energies are reported in dimensionless units.

Figure 2, which presents a 3D graphic of the spectral density as a function of x_t and x_s for a given x_ω, is useful to understand the concept of "broadening in position and transverse kinetic energy." In fact, the spectral density gives a functional relation among x_t, x_s and x_ω. Classically there is the identity $x_\omega = x_t + x_s$, which corresponds in a 3D graphic to a wall of infinite height and zero thickness. In the quantum case, this wall has variable height and thickness from point to point, and the broadening in x_s corresponds to the broadening in the position of the quasi-particle. We remark the existence of negative values of x_s which represent the possibility for the quasi-particle to penetrate the energy region which is classically forbidden (tunnelling effect).

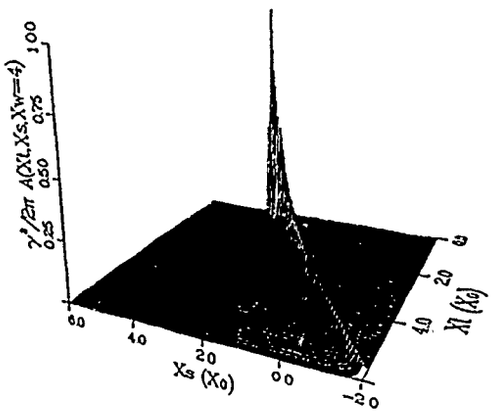


Fig. 2 Spectral density of the quasi-particle as a function of transverse and longitudinal kinetic energies for a given total energy at $E = 500$ kV/cm. Energies are reported in dimensionless units.

MONTE CARLO PROCEDURE

From the scheme presented in the previous section, we suggest the following procedure to account for ICFE and CB within the usual Ensemble MC technique.

(i) In a reference frame where the field is taken along the z direction, without loss of generality, we assume the initial condition of motion at time $t = 0$ as: $z' = 0$, $x_\omega' = 2x_0$ and generate x_t' and x_s' according to (1).

(ii) A carrier free-flight of time duration τ is then determined according to the scattering rate of (2) by making use of the self-scattering procedure.

(iii) The state just before scattering $(x_t, x_s; x_\omega)$ is evaluated as following. Since x_t' is unaffected by the field, $x_t = x_t'$. For x_s we use the SC result

$$x_s = x_s' + \frac{eE}{\gamma^2}(z - z') \tag{7a}$$

where:

$$z = z' + \left(\frac{2\gamma^2 x_s'}{m}\right)^{1/2} \tau + \frac{eE}{2m}\tau^2 . \tag{7b}$$

Then, from the knowledge of x_t and x_s, x_ω is generated according to (1).

(iv) The state after scattering $(x_t', x_s'; x_\omega')$ is finally determined by taking $x_\omega' = x_\omega - x_0$ (which ensures the conservation of the total energy in the scattering) and generating (x_t', x_s') according to (1).

The loop is thus closed and a new free flight is generated as described in (ii). Since we are looking for stationary conditions, the final results of the simulation should not depend on the choice of the initial conditions, as verifiable numerically after a sufficiently long simulation time.

To obtain a faster procedure, without losing the main features above explained, in place of (1) we have introduced the following approximations in the algorithm to be used in the MC simulation. First, we neglect the real part of the self-energy. Second, the imaginary part of the self-energy is taken the same as for the SC scattering rate, that is $Im\ \{S\} = (x_\omega - x_0)^{1/2}$. Third, the low and high energy tails of the spectral density are controlled by an appropriate cut-off. Within these simplifications which, for the very high fields here considered, are quite justifiable, the spectral density in (1) is handled in the following way.

For a given x_ω we generate x_f and x_s through the definition $x = x_f + x_s$ and, from a random number r, x is generated as:

$$x = x_\omega + x_\omega^{1/2} \tan[\frac{\pi r}{B'} + \arctan(-x_\omega^{1/2})] , \tag{8}$$

with

$$B' = \pi\,[\arctan(\frac{\bar{x} - x_\omega}{x_\omega^{1/2}}) - \arctan(-x_\omega^{1/2})] , \tag{9}$$

where

$$\bar{x} = 2x_\omega$$

is the (arbitrary) high energy cut-off. Then, x_f and x_s are generated through a new random number r as:

$$x_f = rx , \qquad x_s = x - x_f . \tag{10}$$

The above procedure implies a spectral density of the form:

$$A(x, \bar{x}; x_\omega) = \frac{2}{\gamma^2} B' \frac{x_\omega^{1/2}}{(x - x_\omega)^2 + x_\omega} . \tag{11}$$

For given x_f and x_s we generate x_ω in a similar way as:

$$x_\omega = x + x^{1/2} \tan[\frac{\pi r}{B} + \arctan(-x^{1/2})] , \tag{12}$$

with

$$B = \pi\,[\arctan(\frac{\bar{x}_\omega - x}{x^{1/2}}) - \arctan(-x^{1/2})]^{-1} \tag{13}$$

and

$$\bar{x}_\omega = 2x$$

is the high energy cut-off. The above procedure implies a spectral density of the form:

$$A(x; x_\omega; \bar{x}_\omega) = \frac{2}{\gamma^2} B \frac{x^{1/2}}{(x_\omega - x)^2 + x} . \tag{14}$$

As numerically proved, the shapes of the spectral densities given by (11) and (14) well compares with that given by (1) for the fields considered here.

APPLICATION TO A MODEL SEMICONDUCTOR

To investigate the main features of ICFE and CB we have performed MC simulations using 10^5 electrons for a simple model semiconductor characterized by the three parameters $m = 0.3m_0$ (m_0 is the free electron mass), $\hbar\omega_0 = 40$ meV and $\gamma^2 = 1.1$ meV. The choice of these values can be considered as typical for several cubic semiconductors.

The main effects in passing from the SC to the quantum case are illustrated in Fig. 3. In the former case, scattering processes conserve both total and kinetic energies and occur vertically in z space. In the latter case, only the total energy in the scattering event is conserved. As a consequence, the kinetic energy is no longer conserved and transitions are broadened in total-energy as well as in z space. Furthermore, penetration beyond the classical turning point is possible.

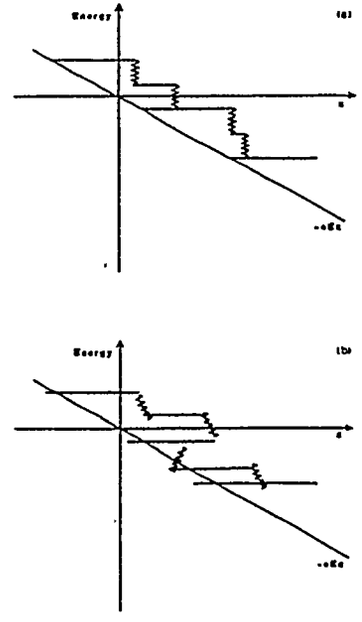


Fig. 3 Quasi-particle trajectories in a total energy scheme under the presence of an electric field. The discontinuous jumps correspond to the emission of optical phonons. (a) Semi-classical case. (b) Quantum case.

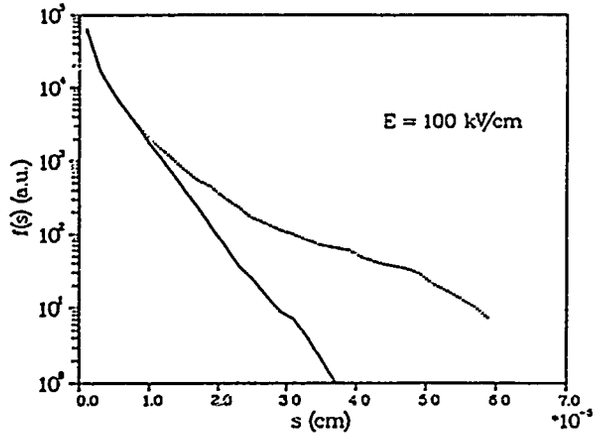


Fig. 4 Distribution function of the longitudinal kinetic energy of the quasi-particle at E = 100 kV/cm. Continuous curve refers to a semi-classical simulation, dotted curve to a simulation which includes simultaneously intra-collisional field effects and collisional broadening.

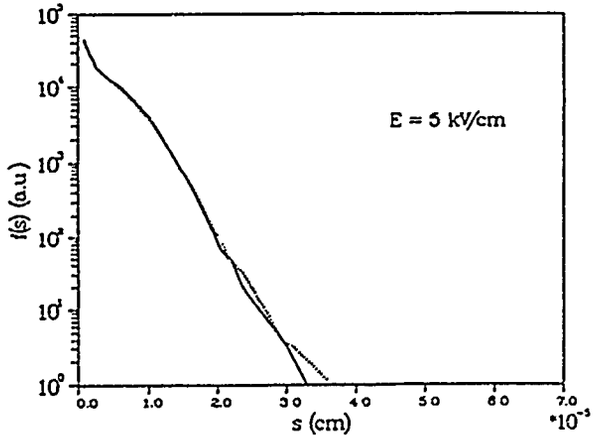


Fig. 5 Distribution function of the longitudinal kinetic energy of the quasi-particle at E = 5 kV/cm. Continuous curve refers to a semi-classical simulation, dotted curve to a simulation which includes simultaneously intra-collisional field effects and collisional broadening.

Figures 4 and 5 display the distribution function of the longitudinal kinetic energy in s space as obtained from the MC simulation at two electric fields. For the sake of comparison, we have also given the results for the SC case obtained by using a delta function for the spectral density. (In order to facilitate a detailed comparison, all curves show the direct MC results.) For the high electric field of Fig. 4, the presence of ICFE and CB is found to strongly increase the number of carriers in the high energy tail of the distribution function with respect to the SC case. An interesting and unexplained feature of such a high-energy tail is that it does not seem to approach a constant energy slope at the highest energy values. At this high field the SC model is characterized by a heated Maxwell-Boltzmann distribution, as expected within the quasi-elastic regime (Reggiani, 1985). For the low electric field of Fig. 5, the calculations show that quantum effects become negligible.

CONCLUSIONS

We have presented an original algorithm, compatible to standard Monte Carlo codes, for estimating quantum effects associated simultaneously to intra-collisional field effects and collisional broadening, in semiconductor high-field transport. The method is based on a spectral density which describes the quasi-particle in terms of the transverse (with respect of the field direction) wavevector, k_t, the Airy transformed space-variable, s, and the total energy $\hbar\omega$. The Monte Carlo simulation performed in such a mixed representation is found to display the following physical effects: (i) Kinetic energy is no longer conserved within a scattering event, processes forbidden in the semiclassical scheme are now allowed. (ii) Scattering processes are no longer local in space, a position broadening along the field direction is evidenced, and its effects are evaluated in terms of a longitudinal kinetic-energy distribution described by a s space variable. (iii) Under semiclassical conditions, the variable s reduces to the classical turning point and the simulation provides an alternative description of transport totally equivalent to the usual full k wave vector representation.

ACKNOWLEDGMENTS

The authors are grateful to Drs. R. Bertoncini, D.K. Ferry, A. Jauho, and A. Kriman for many useful comments on the subject. This research is partially financed by the finalized project "Materiali e Dispositivi per l'Elettronica a Stato Solido (MADESS)" del Consiglio Nazionale delle Ricerche (CNR). The Computer Center of the Modena University is acknowledged for having provided computer facilities.

REFERENCES

Abdolsalami, F., and Khan, F. S., 1990, *Phys. Rev. B*, 30:7394.
Artaki., M, and Hess, K., 1985, *Superlattices and Microstruc.*, 1:489.
Barker, J. R., 1973, *J. Phys. C*, 6:2663.
Barker, J. R., 1978, *Solid-State Electron.*, 21:267.
Bertoncini, R., Kriman, A. M., and Ferry, D. K., 1989, *Phys. Rev. B*, 40:3371.
Bertoncini, R., Kriman, A. M., Ferry, D. K., Reggiani, L., Rota, L., Poli, P., and Jauho, A. P., 1989, *Solid-State Electron.*, 32:1167.
Bertoncini, R., Kriman, A. M., and Ferry, D. K., 1990, *Phys. Rev. B*, 41:1890.
Brennan, K., and Hess, K., 1984, *Solid-State Electron.*, 27:347.
Brennan, K., and Hess, K., 1984, *Phys. Rev. B*, 29:5581.
Brorson, S. D., DiMaria, D. J., Fischetti, M. V., Pesavento, F. L., Solomon, P. M., and Dong, D. W., 1985, *J. Appl. Phys.*, 58:1902.
Brunetti, R., Jacoboni, C., and Rossi, F., 1989, *Phys. Rev. B*, 39:10781.
Capasso, F., Pearsall, T. P., and Thornber, K. K., 1981, *IEEE Electron Device Lett.*, 2:295.
Chang, Y. C., Ting, D. Z. Y., Tang, J. Y., and Hess, K., 1983, *Appl. Phys. Lett.*, 42:76.
Herbert, D. C., and Till, S. J., 1982, *J. Phys. C*, 15:5411.
Khan, F. S., Davies, J. H., and Wilkins, J. W., 1987, *Phys. Rev. B*, 36:2578.
Kim, K., Mason, B. A., and Hess, K., 1987, *Phys. Rev. B*, 36:6547.
Levinson, I. B., and Yasevichyute, Ya., 1972, *Sov. Phys. JETP*, 35:991.
Lowe, D. J., 1985, *J. Phys. C*, 18:L209.

Lugli, P., Reggiani, L., and Jacoboni, C., 1986, *Superlattices and Microstruc.*, 2:143.
Mahan, G. D., 1972, "Polarons in ionic crystals and polar semiconductors," Ed. J.T. Devreese, North-Holland, Amsterdam, p.554.
Mahan, G. D., 1981, "Many-particle Physics," Plenum, New York.
Marsh, A. C., and Inkson, J. C., 1984, *J. Phys. C*, 17:4501.
Porod, W., and Ferry, D. K., 1985, *Physica*, 134B:137.
Pottier, N., and Calecki, D., 1982, *Physica*, 110A:471.
Reggiani, L., 1985, "Hot electron transport in semiconductors," Topics in Applied Physics, Vol. 58, Springer Verlag, Heidelberg.
Reggiani, L., 1985, *Physica*, 134B:123.
Reggiani, L., Lugli, P., and Jauho, A. P., 1987, *Phys. Rev. B*, 36:6602.
Reggiani, L., Lugli, P., and Jauho, A. P. 1988a, *J. Appl. Phys.*, 64, 3072.
Reggiani, L., Lugli, P., and Jauho, A. P. 1988b, *Physica Scripta*, 38:117.
Sarker, S. K., Davies, J. H., Khan, F. S., and Wilkins, J. W., 1986, *Phys. Rev. B*, 33:7263.
Seminozhenko, V. P., 1982, *Phys. Reps.*, 3:103.
Tang, J. Y., and Hess, K., 1983, *J. Appl. Phys.*, 54:5139.
Tang, J. Y., Shichijo, H., Hess, K., and Iafrate, G. J., 1981, *J. Phys. Coll.* (C7), 42:63.
Thornber, K. K., 1978, *Solid-State Electron.*, 21:259.

28

THE FEW-BODY PROBLEM IN NANOELECTRONICS

R. F. O'Connell and G. Y. Hu

Department of Physics and Astronomy
Louisiana State University
Baton Rouge, LA 70803-4001

INTRODUCTION

Few electron systems are achieved by constraining electron motion so that one is no longer dealing with macroscopic bulk matter. (Similar remarks clearly apply to holes, which will be discussed specifically in our remarks on excitons). The constraints are effectively potential barriers which facilitate the creation of low-dimensional (d = 2, 1 or 0) systems. The zero-dimensional systems are the archetype of few electron systems but few electron aspects also play an important role in all low-d-systems. However, since d = 3 (bulk matter) and d = 2 (quantum well, semiconductor-insulator interface) systems have been extensively reviewed in the recent literature our emphasis will be on d = 1 (quantum wires) and d = 0 (quantum dots or boxes). We will concentrate on transport and noise properties but, in the case d = 0, we will also review other topics such as magnetic susceptibility, energy levels etc.

In the following, we will highlight the key features that emerge from reduction of d. The third section will present a general discussion of transport and noise, but will not attempt to survey the variety of approaches that one encounters in the literature. Instead, we will concentrate on the approach which we have initiated, viz. the use of the generalized (quantum) Langevin equation (GLE). Our reasons are two-fold: it focuses the discussion and, above all, we feel it is the preferred method for the treatment of many problems in this area and we will attempt to justify our point of view. In particular, we show that the Kubo formula for conductivity may be derived from the linear-response of our approach. Going beyond linear-response, we are led to an excess (1/f,---) noise. In a related context, we also discuss the observation of discrete resistance fluctuations and their possible explanation in terms of transitions in two-level trapped states. The fourth section will treat transport in semiconductor quantum wires with particular emphasis on the influence of lateral quantization on weak localization in the situation where the width is of the order of the Fermi wavelength. Coulomb interactions are particularly important in this system and we review an analytic approach to the problem which we have recently developed. In Sect. V, we discuss d = 0 systems, with emphasis on the effect of confinement and Coulomb interactions on single-particle and exciton energy levels, finally, magnetic field effects.

FROM BULK MATTER TO QUANTUM DOTS

Electron-beam and focused-ion beam lithography (which result in resolutions the order of 0.1 μm, in contrast to the upper limit of 0.5 μm achieved with ultraviolet light lithography) has led to the fabrication of GaAs semiconductor wires with widths as small as 20 nm and quantum boxes with an average electron number per dot as low as 3±1 (Sikorski and Merkt, 1989). Thus one is presented with the ability to restrict the freedom of motion of electrons in any dimension or, in other words, the ability to fabricate low-d structures.

Our aim in this section is to highlight the key features which emerge from the reduction of d and since these are particularly manifest in transport phenomena we will, first of all, give a broad-brush review of the salient features of charge transport. This, of course, is a problem in non-equilibrium statistical mechanics. In most approaches, one assumes that the system of interest is in thermal equilibrium at a time $t=-\infty$. Next, a c-number force f(t) is applied and eventually, in the distant future, a steady-state situation is achieved. In the calculation of the electrical conductivity, one assumes the so-called thermodynamic limit viz. that the number density of electrons n=(N/V) is finite, while both the total number of particles N and the volume V are assumed to be infinite. Such an assumption ensures that Poincaré recurrences (where the system returns to its initial configuration) are avoided. However, such recurrences are generally not a problem even for finite N since the time interval for such an occurrence is of the order of e^N (Huang, 1987).

In many calculations one is interested in the case of weak impurity scattering and weak E (electric) fields (linear response) and to avoid incorrect results one should take the various limits in a specific order (Van Hove, 1957; Lei *et al.*, 1989): (a) steady state limit ($t\to\infty$), (b) weak impurity limit ($\lambda^2\to 0$, where λ denotes the strength of the impurity potential), and (c) weak E limit.

In order to get an initial feeling for the effects due to dimensionality, we consider a non-interacting gas of electrons near zero temperature with Fermi energy $\varepsilon_F = (mv_F^2/2) = (\hbar^2 k_F^2/2m)$ where $2\pi k_F^{-1} = \lambda_F$, the Fermi wavelength. Then, it follows that $n = a_d k_F^d$ where $a_d = \pi^{-1}, (2\pi)^{-1}, (3\pi^2)^{-1}$ for d = 1, 2, 3, respectively. Also, the density of states (DOS) at the Fermi energy is

$$N(\varepsilon_F) = \frac{dn}{d\varepsilon_F} = \frac{d}{2}\frac{n}{\varepsilon_F} = b_D v_F^{d-2} , \qquad (1)$$

where $b_d = (\pi\hbar)^{-1}, (m/\pi\hbar^2), (m^2/\pi^2\hbar^3)$, for d = 1, 2, 3 respectively. In other words, as an function of increasing ε_F, the DOS increases as $\varepsilon_F^{1/2}$ for d = 3, is constant for d = 2 and decreases as $\varepsilon_F^{-1/2}$ for d = 1 (and for d = 1, 2, where the spectrum consists of a number of discrete subbands, we are referring to the continuous spectrum of each subband). Thus, for quantum wires, there are large discontinuities in the density as one sweeps through the Fermi energy and allows more subbands to be occupied. For d = 0, we have quantization in all directions so that the DOS is zero except for energies corresponding to the eigenenergies.

It is also convenient to write the classical conductivity in the following forms:

$$\sigma_o = (ne^2\tau/m) = \left(\frac{e^2}{\hbar}\right)\left(\frac{n}{k_F}\right) l = a_d \left(\frac{e^2}{\hbar}\right) k_F^{d-1} l = e^2 N(\varepsilon_F) D , \qquad (2)$$

where τ is the elastic scattering mean free time and $l = v_F \tau$ is the elastic scattering mean free path (mfp), which is less than the inelastic mfp at low temperatures. Also $D = (l^2/\tau)d^{-1} = (v_F^2 \tau)d^{-1}$ is the diffusion coefficient. We note that $k_F l >> 1$ for metals. It is also of interest to note that in the case of weak scattering by random impurities we have

$$(\hbar/\tau) = 2\pi n_i N(\varepsilon_F) \int d\Omega \; |U|^2 (1 - \cos\theta), \qquad (3)$$

where U is the Fourier transform of the impurity potential and n_i is the number density of impurities. This result is a useful touchstone, as we will see in the discussion below.

It is now well known that strong localization can occur for small enough values of $k_F \ell$ corresponding to large disorder. However, even for the case $k_F \ell >> 1$ (i.e., when the electron mean free path is much greater than the Fermi wavelength), there exist quantum corrections which reduce the conductivity (weak localization). The origin of weak localization is a quantum interference effect which comes into play if the phase of the electron wave is not changed upon completion of a complete diffusion loop after multiple elastic scatterings. This leads to the introduction of a phase coherent length L_ϕ which is the distance over which the phase of the electron does not change. Since the quantum correction is associated with backscattering and since the latter is enhanced the smaller the dimension, it is clear that low dimension favors weak localization.

Following Pepper (1988) and others, we define the dimensionality of a system with reference to the relative magnitude of the thickness t and width W with respect to L_ϕ:

$$3d \text{ if } L_\phi < W,t\,;\ 2d \text{ if } t < L_\phi < W\,;\ 1d \text{ if } t,W < L_\phi\,. \qquad (4)$$

In our discussion of quantum wires, we shall also distinguish between $W >> \lambda_F$ (effective 1-d, corresponding to many subbands or channels in the transverse direction), $W \approx \lambda_F$ (quasi 1-d) and $W << \lambda_F$ (strictly 1-d).

Another important quantum effect is the universal conductance fluctuations (UCF) $\delta G \sim (e^2/h)$ and it is clear that the smaller the length of the system, L, the larger will be the relative fluctuation. In fact, theory (Al'tshuler and Lee 1988) predicts $(\Delta G/G) = (L_\phi/L)^{d/2}$ if $L >> L_\phi$. Thus, it is clear that the study of quantum interference effects demands

$$\lambda_F << \ell < L_\phi \leq L\,, \qquad (5)$$

the so-called mesoscopic region. Typical values are (Webb *et al.*, 1988; Hiramoto, 1988) $\lambda_F \sim 0.1$nm-0.1μm; $\ell \sim 10$nm-10μm; and $L_\phi \sim 0.1$μm - 1μm. The first inequality ($\ell >> \lambda_F$) ensures that we are in the metallic regime, the second ($L > \ell$) that the motion is diffusive and $L \sim L_\phi$ implies that $(\delta G/G) \sim 1$. The largest L_ϕ values are achieved at low temperatures in semiconductors, as distinct from metals (because D is larger in semiconductors and $L_\phi = (D\tau_\phi)^{1/2}$, where τ_ϕ is the phase breaking time), GaAs being preferable to Si in that respect (because of higher mobilities and higher electron densities).

A characteristic of quantum wires is the small width w which gives rise to quantization of the lateral motion. Denoting the subband energies by ε_i, it is clear that the maximum energy associated with the direction along the wire (x-direction) is $\varepsilon_F - \varepsilon_i \equiv k_{Fi}^2/2m$. Modeling the lateral confinement by an infinitely high square well potential of width W, it follows that $\varepsilon_i = (\hbar^2 k_i^2/2m)$ where $k_i = (\pi i)/W$, where i = 1, 2, ... denotes the number of subbands. Thus, all things else being equal, it is clear that as the width W is increased, the number of occupied subbands increases and the current increases and also that i_{max} can never exceed $(2W/\lambda_F)$.

If $W < \lambda_F$ and also $\ell \geq L$, W then the electron motion is no longer diffusive but, instead, it is ballistic with each subband (channel) carrying the same current. In other words, the wire becomes an electron waveguide. This leads to the conclusion that G is quantized in units of $2e^2/h$ if W is varied (Van Wees et al. 1988; Wharam et al. 1988). However, our present considerations will be confined to the regime defined by (2.5) and the role of subbands will be discussed at length in Sect. IV.

GENERALIZED QUANTUM LANGEVIN EQUATION APPROACH TO QUANTUM TRANSPORT AND FLUCTUATION PHENOMENA

Before getting into technical details, we will summarize why we feel that this approach has many advantages compared to conventional methods:

(i) The GLE separates frictional (dissipative) and random (fluctuating) forces thus giving a natural separation between the conductivity $\sigma(\omega)$ and the noise. Thus, for example, $\sigma(\omega)$ may be calculated without making use of the expression for the random force.

(ii) The GLE is a macroscopic equation corresponding to a reduced description of the original system and it includes dissipative effects in a term which is not invariant under $t \to -t$, where t denotes the time.

(iii) We can make use of the many powerful results developed in the area of stochastic physics.

(iv) We do not need to consider density matrices (and their oft-times accompanying redundant information), Green's functions, nor the correlation functions of Kubo's approach. In fact it is relatively simple to obtain the well-known Kubo formula for the conductivity from our approach but we will argue that our result is much simpler from a computational point of view because it is expressed directly in terms of the susceptibility, a c-number (but which can also incorporate quantum effects).

(iv) Use of a non-linear GLE enables us to go beyond Kubo linear response.

Mori (1965) also considered the GLE, but his approach was to start with the Kubo formula whereas we have used the philosophy of Ford *et al.*, (1988a) to obtain the GLE in a straightforward manner from the many-body Hamiltonian (Hu *et al.*, 1987) using Heisenberg equations of motion. For now we will concentrate on the linear GLE (corresponding to the linear response of Kubo) but below we will also consider the non-linear situation. The GLE is a differential equation for the coordinate of the center of mass of the many-body system. By taking the Fourier transform (F.T.) of this equation we were able to obtain a simple result for $R(\omega)$, the F.T. of R(t). Next, averaging over the heat-bath variables (i.e. over the impurities, phonons and relative electrons) we obtained the ensemble average

$$\langle R_\alpha(\omega)\rangle = N\, \alpha_{\alpha\beta}(\omega)\, f_\beta(\omega) \quad (\alpha, \beta = 1,2,3)\,, \qquad (6)$$

where $\vec{f}(\omega)$ is the F.T. of the external force $\vec{f}(t) = e\vec{E}(t)$, $\vec{E}(t)$ is the electric field and $\alpha_{\alpha\beta}(\omega)$ is the susceptibility tensor. In fact

$$[\alpha^{-1}(\omega)]_{\alpha\beta} = -\,M\,\omega^2\,\delta_{\alpha\beta} - i\,\omega\,\mu_{\alpha\beta}(\omega)\,, \qquad (7)$$

where M = Nm is the total mass and $\mu_{\alpha\beta}(\omega)$ is the F.T. of the memory function $\mu_{\alpha\beta}(t)$ appearing in the GLE. Also, the memory function used by Götze *et al.* (1972) and others, is simply $(i/M)\,\mu(\omega)$.

If $j(\omega)$ denotes the current produced by an electric field of frequency ω, then, using the fact that the F.T. of R'(t) is $-i\omega R(\omega)$, where the prime denotes a time derivative, it follows from (6) that

$$\langle j_\alpha(\omega)\rangle = ne\,(-i\omega)\,\langle R_\alpha(\omega)\rangle \tag{8a}$$

$$= -\,i\,Nn\,e^2\,\omega\,\alpha_{\alpha\beta}(\omega)\,E_\beta(\omega)\,. \tag{8b}$$

But, since

$$\langle j_\alpha(\omega)\rangle = \sigma_{\alpha\beta}(\omega)\,E_\beta(\omega)\,, \tag{9}$$

it follows immediately that

$$\sigma_{\alpha\beta}(\omega) = -\,i\,Nn\,e^2\,\omega\,\alpha_{\alpha\beta}(\omega)\,. \tag{10}$$

Thus, we have obtained a simple expression for the conductivity in terms of the susceptibility, which, in turn, using (7), is obtained simply from the F.T. of $\mu(t)$, the latter being the memory term which appears in the GLE. These results form the basis of many calculations which we have already carried out. However, before proceeding, it will be both instructive and enlightening if we link up with the Kubo formula. To this end we use the fluctuation dissipation theorem (Callen *et al.*, 1951; Ford *et al*, 1988b) and the fact that $j_\alpha = neR'_\alpha(t)$ to obtain

$$S^{(j)}_{\alpha\beta}(t-t') \equiv \frac{1}{2}\langle j_\alpha(t)\,j_\beta(t') + j_\beta(t')\,j_\alpha(t)\rangle$$

$$= \frac{\hbar}{2\pi}(ne)^2\int_{-\infty}^{\infty} d\omega\,\omega^2\coth\left(\frac{\hbar\omega}{2kT}\right)\mathrm{Im}\,\alpha_{\alpha\beta}(\omega)\,e^{-i\omega(t-t')}\,. \tag{11}$$

Thus, taking the F.T. we get

$$\tilde{S}^{(j)}_{\alpha\beta}(\omega) \equiv \frac{1}{2}\int_{-\infty}^{\infty} dt\,\langle j_\alpha(t)\,j_\beta(t') + j_\beta(t')\,j_\alpha(t)\rangle\,e^{i\omega(t-t')}$$

$$= \hbar\,(ne)^2\,\omega^2\coth\left(\frac{\hbar\omega}{2kT}\right)\mathrm{Im}\,\alpha_{\alpha\beta}(\omega) = \frac{\hbar}{V}\,\omega\coth\left(\frac{\hbar\omega}{2kT}\right)\mathrm{Re}\,\sigma_{\alpha\beta}(\omega)\,, \tag{12}$$

where the last equality follows from (10). Hence

$$\mathrm{Re}\,\sigma_{\alpha\beta}(\omega) = \frac{V}{2\hbar\omega}\frac{1}{\coth(\hbar\omega/2kT)}\int_{-\infty}^{\infty} dt\,\langle j_\alpha(t)\,j_\beta(t') + j_\beta(t')\,j_\alpha(t)\rangle\,e^{i\omega(t-t')}$$

$$= \frac{V}{2\hbar\omega}\left(1 - e^{-\hbar\omega/kT}\right)\int_{-\infty}^{\infty} dt\,\langle j_\alpha(t)\,j_\beta(t')\rangle\,e^{i\omega(t-t')}\,, \tag{13}$$

which is a form of the Kubo formula (Mahan, 1981). Thus, our approach is consistent with that of Kubo but it is our contention that our result, given by (10), is a simpler form for the conductivity compared to (13). In particular, (3) follows readily from our approach (Hu *et al.*, 1987), in contrast to the sophistication needed in the Kubo approach, where self-energy and vertex corrections must be included (Mahan,1981). More generally, our approach has a further important advantage in that dissipation is incorporated in a natural way: its presence is guaranteed when $\mathrm{Re}\{\mu(\omega)\}$ as shown by Ford *et al.* (1988a), and the GLE provides what we consider to be the preferred way to include dissipation because it incorporates the breaking of the time-reversal invariance of the original microscopic equations. By contrast, as Landauer as emphaiszed on many occasions (1987, 1989), in the Kubo approach one is dealing with a closed conservative system which "... does not permit dissipative effects and, typically, some supplementary cheating accounts for these" (Landauer, 1987).

We have previously applied the GLE to a variety of problems, but here we mention: (a) the effect of fluctuations of the center of mass (which leads to dissipative effects which are proportional to N^{-1} and thus are accentuated in few electron systems), which in turn led to an explanation of heretofore unexplained results for cyclotron resonance in 2-d systems (Hu *et al.*, 1989a,b); (b) the incorporation of backscattering and thus quantum effects (Hu *et al.*, 1988), which forms the basis of our approach to transport in 1-d systems (see next Sect.); (c) the generalization to the non-linear regime, which was used in our discussion high electric field effects (Hu *et al.*, 1989c) and of 1/f noise ($f = \omega/2\pi$) in low-d systems (Hu *et al.*, 1990a). Here, we dwell only on the latter, because of its intrinsic importance in low d systems and also the link with research on UCF.

The GLE provides a natural way to go beyond ensemble averages and calculate fluctuations. First, we note that in the $\omega \to 0$ limit, $\tilde{S}^{(j)}_{\alpha\beta}(\omega)$ in (12) leads directly to the power spectrum of the voltage autocorrelation function i.e. Nyquist noise (Hu *et al.*, 1990a). In the nonlinear GLE, in contrast to the linear case, the memory function and the random force are functions of the particle coordinates as well as t. A detailed analysis (Hu *et al.*, 1990a) leads, in the low-frequency limit to an excess noise with a frequency dependence $f^{-1/2}$ for d = 3 and f^{-1} for d = 2. Our calculations show that 1/f noise is a universal property, and that it increases with decreasing number of electrons (as N^{-1}). Furthermore, it is proportional to $(\delta G/G)^2$, thus providing a link with UCF research (Feng *et al.*, 1986). In the work of Feng *et al.*, the source of the 1/f noise is considered to be a change in the random configurations due to defect motion (back and forth between two sites) whereas our emphasis is on the non-linear contribution to the time variation of the random force exerted on the electrons due to the impurities which, in turn, appears to have its physical origin, at least partially, in the deviation of the diffusion coefficient from its conventional value, due to low-d effects.

In a related context, we now discuss the observation of discrete resistance fluctuations (DRF) in several small systems and their possible relationship of 1/f noise in large devices. Ralls *et al.* (1984) measured the DRF in small wires (narrow submicrometer MOSFETS) over a range of temperatures T(°K - 100°K) and electron densities, on time scales of the order of 0.01 - 10s. Plots of resistance as a function of time display a random switching, of magnitude up to 1%, between two values, the switching times being strong functions of T and gate voltage. The discrete nature of this "telegraph noise" leads to the conclusion that it is due to the fluctuating occupancy of electron traps. In these devices, the $Si\text{-}SiO_2$ interface has a very low density of fixed charges and defects (which contribute to the conductivity of electrons in semiconductors) at which electrons can be trapped (Howard *et al.*, 1986). In fact, the number of defects can be as low as 1 in 10^5 sites and, since a given trap needs to be close to the interface and have an energy within several times kT, the conclusion of Ralls *et al.* (1984) is that they are probably observing the switching of a single trap. Since the traps which cause the switching have a broad distribution of activation energies, it was also concluded that the superposition of many individual switching events which occur in larger devices is a possible way to obtain 1/f noise. This idea was also discussed by Uren *et al.* (1985) and it was made more general and quantitative by Feng et al. (1986).

Rogers and Buhrman (1985) observed DRF in tunnel junctions and concluded that the switching is due to thermal activation above ~15°K but due to tunneling for lower temperatures. Farmer *et al.* (1987) added a new dimension in their observations of DRF in MOS insulating tunnel diodes: they observed changes in tunnel resistance greater than 10% which they attributed to trap states switching in synchronization - a collective mechanism, resulting from a strongly interacting cluster of localized trap states, which they speculate might be the trigger for electrical breakdown in insulating films. Furthermore, DRF was observed (Ralls *et al.*, 1988;

Ralph *et al.*, 1989) in metallic structures (with constrictions as small as 10 nm), ranging up to 0.2% of the total resistance, lending more credence to the view that the observations result from defects fluctuating between two metastable configurations. Since the latter essentially implies fluctuating forces on the electron, our future plans are to examine quantitatively how it can be incorporated into our GLE framework.

TRANSPORT IN QUANTUM WIRES

Semiconductor quantum wires, with widths w as small as 20 nm, have recently been the subject of intense study (Skocpol *et al.*, 1982, 1986; Howard *et al.*, 1986; Hiramoto, 1988; Hiramoto *et al.*, 1989). The key characteristic of a quantum wire is that W is the order of magnitude of λ_F so that the electron states are quantized laterally (y direction). Thus, a quantum wire is different from a strictly 1-d system (Mott *et al.*, 1961; Landauer, 1970) -- where there is no lateral degree of freedom--and the effective 1-d systems (Thouless 1977, 1980)--for which $W >> \lambda_F$ and there is effectively no lateral quantization. This intermediate region, between strictly 1-d and effective 1-d, in which no more than, say, about 2 to 5 subbands are occupied, we will refer to as quasi 1-d.

Typical numbers for GaAs wires (Hiramoto, 1988; Hiramoto *et al.*, 1989) are: thickness (z-direction) 2 nm, width 100 nm, λ_F~50 nm, t~1 μm, and subband energy spacing ($\hbar\omega_o$ say) ~2 meV (corresponding to an effective temperature of 23°K). Also $\varepsilon_F = (2\pi\hbar)^2/2m\lambda_F^2 \simeq 9$ meV so that, in this case, no more than 5 subbands can be occupied.

Here, we will focus attention on the quasi 1-d regime (Q1d) and investigate the influence of lateral quantization on such topics as weak localization and electron-electron interactions, as well as the influence of the latter on such topics as impurity screening, plasmon excitations, absorption and the classical conductivity.

Weak Localization

Since the Thouless diffusive picture is no longer applicable in this few-bands regime, we introduced another mechanism to explain the quantum effects arising from coherent back scattering (CBS). First, we recall that according to the diffusive picture, the CBS is realized through coherent scattering sequences (fan diagram), where an electron of Fermi momentum k_F moves in a diffusive way such that its momentum gradually changes to $-k_F + q$ (with $q/k_F << 1$). As already noted, the average distance (the phase coherent length L_ϕ) over which the electron diffuses during these sequences, is estimated to be $\sqrt{D\tau_\phi}$, where D is the diffusion constant and τ_ϕ is the phase coherent time, the average time for a CBS process. In our picture (Hu *et al.*, 1989d, 1990b) – which we refer to as the sudden reversal picture – for a quantum wire in the quasi 1-d regime, the electrons are now assumed to be scattered by impurities into only two kinds of states. One is a small momentum transfer forward process which essentially does not change the velocity of electrons, the other is a large momentum transfer (~ $2k_F$) process which makes the electron move essentially in the reversed direction. In addition, the assumption that the system is lightly doped makes the probability of the reversal scattering much less than the forward scattering. (The opposite case, i.e., when the reversal scattering dominates, corresponds to the strictly 1D case). In this way an electron will experience many forward scatterings with little change in its original speed. Eventually it will experience a reversal scattering. Hence, an electron will travel a distance $L_\phi \sim v_F\tau_\phi$ in a CBS process, as distinct from the result $L_\phi \sim \sqrt{D\tau_\phi}$ in the diffusive picture. Thus, in our picture, larger values of L_ϕ are obtained; in fact, much larger than the mean-free-path l.

Physically, due to the relatively large value of the Fermi wave length of the semiconductor, the dilute impurities in the quantum wire can not individually block the way of the moving electrons and hence ensures that the reversal scattering has a small probability (roughly proportional to the ratio of the size of the impurity to the width of the wire), in contrast to the strictly 1-d case. At the same time, the lateral quantization of the sample restricts the motion of the electrons essentially in a 1d fashion and thus makes the other possible way of impurity scattering, the forward scattering, the dominant process. Our method of calculation (Hu *et al.*, 1989d, 1990b) is to use the static limit of our general result for the conductivity, including the CBS contribution and the effects due to a strong electric field (Hu *et al.*, 1988). By keeping the usual form for expressing the weak localization conductivity in terms of L_ϕ (and by taking account of the fact that the fraction of electrons scattered by boundaries lead to a random change in phase, in contrast to the case of impurity scattering), we extracted a theoretical expression for L_ϕ which we compared to the experiments of Hiramoto *et al.* (1988,1989). For small E field values, we find that L_ϕ is larger than typical values obtained in the usual diffusive picture. In addition, we find that L_ϕ decreases with E for large E. Explicitly, we find that when E exceeds a critical value $E_c = \hbar v_F/eL_\phi^2$, it will introduce a new cut-off length $L_s = (E_c/E)^{1/2} L_\phi$; the latter is obtained by equating the energy gained (eEL_s) in travelling a distance L_s to the broadening energy ($\hbar v_F/L_s$) associated with a CBS process. This enables us to calculate a critical current $I_c = 7.5 \times 10^8$ Å, in good agreement with experiment (Hiramoto *et al.*, 1989) and in contrast with the result obtained by use of a cut-off length derived from a diffusive picture (Mott and Kaveh, 1981).

We emphasize that our theory leads to a situation where the quantum wire will remain non-localized (i.e., σ does not approach 0) even though the length of the system is much larger than l. In other words, much larger values of the phase coherent length, L_ϕ, are possible. In particular, this is relevant to the analysis of temperature effects since the quantum correction to the conductivity depends on L when $L_\phi > L$ (which occurs for low temperature since L_ϕ increases with decreasing temperature) but depends on L_ϕ at finite temperatures when $L > L_\phi$. Thus, we predict that the conductivity remains independent of the temperature T for values of T larger than that obtained from conventional theory and our theoretical results are in very good agreement with the experimental results of Hiramoto *et al.* (1989).

We have also gone beyond the static results discussed above and examined the frequency dependence of the conductivity. In particular, we find that, for a weakly localized quantum wire, there exists a critical frequency beyond which the system will be delocalized and will exhibit a ω^2 behavior. The latter behavior has been previously obtained for strongly localized systems (Mott and Twose, 1961; Landauer, 1970), which suggests that the $\sigma(\omega)$~ω^2 behavior should not be used as a criteria for judging whether the system is strongly localized. Instead, one may use the clear different temperature dependence of the weakly localized ($\sigma \sim T^{-n}$, where n depends on the inelastic scattering mechanism) and the strongly localized ($\sigma \sim \exp(aT^{-n})$) systems. We note that existing experimental data (Hiramoto *et al.*, 1989) for the lightly doped semiconductor quantum wires, indicate a power law dependence for the conductance of these systems, in support of the weak localization description. Unfortunately, to our knowledge, no a.c. conductance measurements on those systems have been performed up to now. According to our theory, the $\sigma(\omega)$ measurement will give a clue to the dynamics of the CBS process.

Electron-Electron (e-e) Interactions

For a Q1d system, due to the lateral quantization, the form of the Coulomb interaction is quite complicated and is generally evaluated numerically both for the intra-subband and inter-subband cases (Das Sarma *et al.*, 1985; Li and Das Sarma, 1989) – with the ground state intra-subband Coulomb interaction as the only analytical case appearing in the literature. The large amount of numerical work involved in evaluating the Coulomb interaction not only constitutes an obstacle to setting up a practical transport theory for the Q1d system, but also gives no hint of the answer to some basic physical questions such as the difference between the intra-subband and inter-subband interactions. Therefore it is very desirable to find a simple form for the Q1d Coulomb potential matrix elements, which we have now done, using a harmonic confinement potential in the lateral direction (Hu *et al.*, 1990c). We show that the intra-subband Coulomb interaction keeps the typical logarithmic divergent behavior of the 1d system in the long wave length limit ($q \rightarrow 0$), while the inter-subband Coulomb interaction approaches discrete values depending on the band index of the relevant subbands. This latter intriguing behavior of the inter-subband Coulomb interaction was unnoticed in numerical studies of the same problem appearing in the literature. We also find that the general rule for the inter-subband Coulomb interaction in the $q \rightarrow 0$ limit, makes it possible to study plasmon excitations for a Q1d system with an unlimited number of subbands. In particular, we find that there are M different modes for the collective excitations of a single subband separation of a Q1d system with M occupied subbands. Each of these modes has a plasmon frequency shifted from the single particle value by a factor depending on the related Fermi momentum, the frequency associated with the harmonic potential model, and the intrinsic properties of the system. This is the so-called depolarization effect known from the study of the subband effect in the two-dimensional systems. The difference here is that numerous populated subbands are involved, and we find that multi-modes exist, each mode contributing a different amount to the depolarization effect. Using the data from the experiments of Demel *et al.* (1988), the largest value of the plasmon frequency, ω_p^M say, that can be obtained is 2.4 meV, which is to be compared with the experimental results of Demel et al., who found a plasmon peak around 3.4 meV. This is encouraging especially when it is expected (Hu *et al.*, 1990c) that the inclusion of the plasmon-plasmon coupling will increase the theoretical estimate for ω_p^M. In addition, we note that when the subband separation increases due to the strengthening of the confinement potential, the Fermi momentum of the top populated subbands also increases. Thus, when one decreases the gate voltage to decrease the electron density and thereby to strengthen the confinement potential, as was done in the experiments, we find an increase of the inter-subband plasmon frequency (Hu *et al.*, 1990d) in good agreement with the experimental data (Brinkop *et al.*, 1988). We have also applied the derived analytical form of the Q1d e-e interaction to a study of the Q1d polarizability beyond the RPA (Hu *et al.*, 1990d). This, in turn, is used to study impurity screening.

Next, we focus our attention on subband effects on electron transport in Q1d interacting electron-impurity systems. Since our main interest here is on quantum effects arising from subband structure, as distinct from quantum interference effects, our results are geared toward analysis of a system of multiple parallel quantum wires (since the multiplicity of wires serves to average out interference effects).

From (1), we see that for d=1, the DOS in the electron model is divergent near the bottom of each of the subbands. This feature causes the conductivity, σ_o, to vanish whenever the Fermi energy crosses the bottom of a subband. However, taking into account electron-electron interaction effects and also impurity screening, as well as broadening effects resulting from fluctuations of the center-of-mass of the system (Hu *et al.*, 1989a,b), we find that the divergences in the DOS are eliminated and σ_o does not vanish. Finally, we evaluated σ_o and find that its behavior clearly reflects the role of an increasing number of subbands as the Fermi energy is increased. In particular, we find that σ_o is greatly enhanced when the E_F is such that only the lower half of each subband is filled. Physically, this is because in this region the screening is very effective (which enhances the mobility) and the dissipative intersubband scattering (which tends to reduce the mobility) is not very effective. Such features may be identified in the experimental results of Ismail *et al.* (1989) but quantitative agreement between theory and experiment is lacking and we are presently trying to ascertain whether the inclusion of such effects as electron-phonon interactions might make a difference.

ZERO DIMENSIONAL SYSTEMS

Confinement in all 3 directions leads to a quasi 0-d system. These so-called quantum boxes or dots have been fabricated with electron numbers as low as 3. The DOS is non-zero except for energies corresponding to the eigenenergies. In fact, Reed *et al.* (1988) interpreted their observations of fine-structure in the resonant tunneling through a 3-d confined In GaAs quantum well as being due to the discrete density of states in a 0-d quantum box. Also, the absorption spectrum should consist of a series of discrete lines and much interest has focused on the calculation of energy levels. Other subjects of particular interest are the observations of discrete resistance fluctuations and effects of the confinement on magnetic susceptibilities. Hence, we focus attention on these three topics in the following.

Energy Levels of Impurities and Excitons in Quantum Boxes

Coulomb interaction effects scale as L^{-1} whereas kinetic energy matrix element scale as L^{-2}, where L is a typical confinement dimension. Thus, interaction effects are relatively larger for larger systems and also for lower dimensional systems (Bryant, 1984, 1987). Thus, in long narrow boxes with $L \leq 100$ nm, strong correlations occur and the formation of a Wigner lattice is a possibility. In smaller boxes, $L \leq 10$ nm, confinement effects play a larger role whereas the correlations are weak but not as weak for boxes comprising finite barriers as distinct from infinite barriers (because the higher the barrier the larger the kinetic energies).

Consider next the spectrum of a hydrogenic impurity placed in a quantum-well wire (QWW) mode from a cylinder of GaAs surrounded by GaAlAs (Bryant, 1985): since $m = 6.7 \times 10^{-2}\ m_o$ (where m_o is the electron mass) and the dielectric constant is 13.1, it follows that, if we were dealing with a bulk system, the ground state energy and the corresponding radius would be 5.3 meV and 103 Å, respectively. As the system dimensions are decreased, the binding energy increases, being a factor of 4 larger when the system becomes a very narrow well. When the dimension of the QWW approaches that of the radius of the bound state it might be expected that shape effects would play an important role. However, Bryant (1985) showed that this is not so but instead the important parameter affecting the energy spectrum is the cross-sectional area.

Turning to experiments, Sikorski and Merkt (1989) prepared arrays of 10^8 dots on p-type InSb surfaces. Using far-infrared magnetospectroscopy, they observed intraband transitions between the discrete states of the quantum dots. InSb is a narrow gap semiconductor which has a small effective mass $m = 1.4 \times 10^{-2}\ m_o$ at the conduction-band edge, resulting in energies of 10 meV at widths of 100 nm. The number of electrons per dot were controlled by the gate voltage and electron numbers ranging from as low as 3 to 20 were deduced. This study of few-electron systems was made possible by the application of magnetic fields up to 3T so that magnetic energies are larger than Coulomb binding energies. Thus, we are in the "strong magnetic field regime" which has been the subject of intense study (O'Connell,1982). In fact it would be of interest to apply the results of these latter studies to the present problem instead of treating the binding as being due to a harmonic oscillator potential (Sikorski *et al.*, 1989). In a related experiment, Hansen *et al.* (1989) also considered the splitting of 0-d energy levels by a magnetic field but, in contrast to the case of semiconductor impurities, they considered electrons in microstructured heterojunctions, so that we are dealing with an external confining potential, which is very different from the Coulomb potential. The observations here led to plots of dC/dV_g versus V_g, where C is the capacitance and V_g the gate voltage.

Further evidence for discrete energy levels in quantum dots is provided by the measurements of Smith *et al.* (1988), who performed capacitance spectroscopy; peaks in the

DOS were determined from the first derivative of the capacitance. The samples also exhibited Shubnikov-de Haas-type oscillations of the capacitance as a function of magnetic field.

Que and Kirczenow (1988) have suggested that, even though the quantum dots in an array are electrically insulated from each other, the long-range Coulomb force couples the quantum dots and this could lead to collective excitations, which should be experimentally detectable (the relevant energies being much higher than the energy-level splittings in individual dots).

Similar considerations apply to the spectrum of excitons in quantum boxes (Bryant, 1988): confinement effects become more dominant and correlation effects less so as the size of the quantum box shrinks, particularly for dimensions ≤100 nm. When L≤10 nm, the box and exciton are about the same size and the electron-hole pair is in the state of lowest energy. Also, as the box size decreases, there is an increased electron-hole overlap and the exciton oscillator strength increases, thus increasing the recombination rates of the lowest-energy exciton. Thus, as Bryant (1988) suggests, quantum boxes may possess enhanced optoelectronics properties, which may provide the explanation for the enhanced luminescence efficiency observed by Kash *et al.* (1986) in photoluminescence measurements excitation measurements. Finally, we note that, in the related area of ultrafine particles, attempts have been made to use 0-d superlattices as memory devices (Hayashi, 1987).

Magnetic Susceptibility

It is now well-known that the usual bulk matter results for magnetization and susceptibility have to be modified for small structures because of the increasing importance of surface effects (Khaikan, 1969; Wang *et al.*, 1986; Robnik, 1986). In general, surface effects are small unless the size is as small as about 100Å (Wang *et al.*, 1986). In particular, the latter authors considered a small metallic ball and they took surface effects into account by means of an isotropic-oscillator potential (and an anharmonic perturbation was also considered). In general, they found a reduction of the susceptibility below the Landau result. However, they did not consider oscillatory effects, which come into play at high magnetic field (B) values and low temperatures (T).

For quantum dots, we have already alluded to the fact that quantization energies due to B can be comparable to and larger than the discrete energies in the absence of a magnetic field. Thus, if the temperature is sufficiently low, we are in a regime where de Haas-van alphen (dHvA) oscillations in the magnetization and susceptibility may occur. This regime has recently been studied by Sivan and Imry (1988) and they find that the effect of the surface states is to smoothen the sharp edges and also add a high frequency low amplitude modulation to the curve describing bulk matter. However, for high quality GaAs quantum dots of a few mm diameter, they find the level of flux measurement required is an order of magnitude below the noise limits of present day SQUID systems. On the other hand, the same authors claim that a signal out of 10^4 such dots should be detectable.

ACKNOWLEDGEMENTS

This work was supported in part by the U. S. Office of Naval Research under grant No. N00014-90-J-1124. The authors are very grateful to Professor D. K. Ferry for sharing with them, on several occasions, part of his encyclopedic knowledge of the literature but particularly for steering them to the work of the Ikoma group. We thank Professor T. Ikoma for sending us preprints etc. and especially a copy of the beautiful thesis of his former student, Dr. T. Hiramoto. We also benefited from conversations with Dr. H. U. Baranger, Professor J. Barker, and Dr. D. Browne.

REFERENCES

Al'tshuler, B. L. and Lee, P. A., 1988, Disordered Electronic Systems, Phys. To-Day 41(12):36.

Brinkop, F., Hansen, W., Kotthaus, J. P. and Ploog, K., 1988, One-dimensional Subbands of Narrow Electron Channels in Gated Al_xGa_{1-x} As/GaAs Heterojunctions, Phys. Rev. B 37:6547.

Bryant, G. W., 1984, Hydrogenic Impurity States in Quantum-Well Wires, Phys. Rev. B 29:6632.

Bryant, G. W., 1985, Hydrogenic Impurity States in Quantum-Well Wires: Shape Effects, Phys. Rev. B 31:7812.

Bryant, G. W., 1987, Electronic Structure of Ultrasmall Quantum-Well Boxes, Phys. Rev. Lett. 59:1140.

Bryant, G. W., 1988, Excitons in Quantum Boxes: Correlation Effects and Quantum Confinement, Phys. Rev. B37:8763.

Callen, H. B. and Welton, T. A., 1951, Irreversibility and Generalized Noise, Phys. Rev. 83:34.

Das Sarma, S. and Lai, W. Y., 1985, Screening and Elementary Excitations in Narrow-Channel Semiconductor Microstructures, Phys. Rev. B 32:1401.

Demel, T., Heitmann, D., Grambow, P. and Plogg, K., 1988, Far-infrared Response of One-Dimensional Electronic Systems in Single- and Two-layered Quantum Wires, Phys. Rev. B 38:12732.

Farmer, K. R., Rogers, C. T. and Buhrman, R. A., 1987, Localized State Interactions in Metal-Oxide-Semiconductor Tunnel Diodes, Phys. Rev. Lett. 58:2255.

Feng, S., Lee, P. A. and Stone, A. D., 1986, Sensitivity of the Conductance of a Disordered Metal to the Motion of a Single Atom: Implications for 1/f Noise, Phys. Rev. Lett. 56; 1960, 2772(E).

Ford, G. W., Lewis, J. T. and O'Connell, R. F., 1988a, Quantum Langevin Equation, Phys. Rev. A 37:4419.

Ford, G. W., Lewis, J. T. and O'Connell, R. F., 1988b, Quantum Oscillator in a Blackbody Radiation Field II. Direct Calculation of the Energy Using the Fluctuation-Dissipation Theorem, Annals. of Phys. (N.Y.) 185:270, Eq. (A.14).

Götze, W. and Wölfle, P., 1972, Homogeneous Dynamical Conductivity of Simple Metals, Phys. Rev. B 6:1226.

Hansen, W., Smith III, T. P., Lee, K. Y., Brum, J. A., Knoedler, C. M., Hong, J. M. andKern, D. P., 1989, Zeeman Bifurcation of Quantum-Dot Spectra, Phys. Rev. Lett. 62:2168.

Hayashi, C., 1987, Ultrafine Particles, Phys. To-Day 40(12):44.

Hiramoto, T., 1988, The Quantum Interference Effect of Electron Waves in Semiconductor Quantum Wires Fabricated by Focused Ion Beam Implantation, Ph.D. Thesis, Tokyo University.

Hiramoto,T., Hirakawa, K., Iye, Y. and Ikoma, T., 1989, Phase Coherence Length of Electron Waves in Narrow AlGaAs/GaAs Quantum Wires Fabricated by Focused Ion Beam Implantation, Appl. Phys. Lett. 54:2103.

Howard, R. E., Jackel, L. D., Mankiewich, P. M. and Skocpol, W. J., 1986, Electrons in Silicon Microstructures, Science 231:346.

Hu, G. Y. and O'Connell, R. F., 1987, Quantum Transport for a Many-Body System Using a Quantum Langevin-Equation Approach, Phys. Rev. B 36:5798.

Hu, G. Y. and O'Connell, R. F., 1988, Strong Electric Field Effect on Weak Localization, Physica A153:114.

Hu, G. Y. and O'Connell, R. F., 1989a, Fluctuation Effects on the Cyclotron Resonance Spectrum for a Two-Dimensional Electron Gas, Phys. Rev. B 37:10391.

Hu, G. Y. and O'Connell, R. F., 1989b, Cyclotron Resonance in Two-Dimensional Electron-Phonon-Impurity Systems and Applications to Si Metal-Oxide-Semiconductor Systems, Phys. Rev. B 40:11701.

Hu, G. Y. and O'Connell, R. F., 1989c, Generalized Quantum Langevin Equations for High-Electric-Field Transport, Phys. Rev. B 39:12717.

Hu, G. Y. and O'Connell, R. F., 1989d, Electric Field Effect on Weak Localization in a Semiconductor Quantum Wire, Solid-State Electron. 32:1253.
Hu, G. Y. and O'Connell, R. F., 1990a, 1/f Noise: A Nonlinear-Generalized-Langevin-Equation Approach, Phys. Rev. B 41:5586.
Hu, G. Y. and O'Connell, R. F., 1990b, Weak Localization Theory for Lightly Doped Semiconductor Quantum Wires, J. Phys. C, in press.
Hu, G. Y. and O'Connell, R. F., 1990c, Electron-Electron Interactions in Quasi-One-Dimensional Electron Systems, Phys. Rev. B, in press.
Hu, G. Y. and O'Connell, R. F., 1990d, Dielectric Response of a Quasi-One-Dimensional Electron System, preprint.
Huang, K., 1987, Statistical Mechanics, 2nd ed., (Wiley), p. 90.
Ismail, K., Antoniadis, D. A. and Smith, H. I., 1989, One-Dimensional Subbands and Mobility Modulation in GaAs/AlGaAs Quantum Wires, Appl. Phys. Lett. 54:1130.
Kash, K., Scherer, A., Worlock, J. M., Craighead, H. G. and Tamargo, M. C., 1986, Optical Spectroscopy of Ultrasmall Structures Etched from Quantum Wells, Appl. Phys. Lett. 49:1043.
Khaikin, M. S., 1969, Magnetic Surface Levels, Adv. Phys. 18:1.
Landauer,R., 1970, Electrical Resistance of Disordered One-Dimensional Lattices, Phil. Mag. 21:863.
Landauer, R., 1987, Electrical Transport in Open and Closed Systems, Z. Phys. B68:217.
Landauer, R., 1989, Nanostructure Physics: Fashion or Depth? in Nanostructure Physics and Fabrication, Reed, M. A. and Kirk, W. P., eds. (Academic Press).
Lei, X. L. and Horing, N. J. M., 1989, Divergence in the Balance-Equation Theory of Resistivity, Phys. Rev. B40:5985.
Li, Q. and Das Sarma, S., 1989, Collective Excitation Spectra of One-Dimensional Electron Systems, Phys. Rev. B40:5860.
Mahan, G. D., 1981, Many-Particle Physics, (Plenum Press). Re the Kubo formula, see pps. 192 and 222.
Mori, H., 1965, Transport, Collective Motion, and Brownian Motion, Prog. Theor. Phys. 33:423.
Mott, N. F. and Kaveh, M., 1981, The Conductivity of Disordered Systems and the Scaling Theory, J. Phys. C 14:L659.
Mott, N. F. and Twose, W. D., 1961, The Theory of Impurity Conduction, Adv. Phys. 10:107.
O'Connell, R.F., 1982, Two Dimensional Systems in Solid State and Surface Physics: Strong Electric and Magnetic Fields Effects, J. de Physique, Colloque C2:81.
Pepper, M., 1988, Quantum Processes in Semiconductor Structures, Proc. R. Soc. Lond. A420:1.
Que, W. and Kirczenow, G., 1988, Theory of Collective Excitations in a Two-DimensionalArray of Quantum Dots, Phys. Rev. B 38:3614.
Ralph, D. C., Ralls, K. S. and Buhrman, R. A., 1989, Defect Motion, Electromigration and Conductance Fluctuations in Metal Nanocontacts, in Nanostructure Physics and Fabrication, Reed, M. A., and Kirk, W. P., eds. (Academic Press).
Reed, M. A., Randall, J. N., Aggarwal, R. J., Matyi, R. J., Moore, T. M. and Wetsel, A. E., 1988, Observation of Discrete Electronic States in a Zero-Dimensional Semiconductor Nanostructure, Phys. Rev. Lett. 60:535.
Ralls, K. S., Skocpol, W. J., Jackel, L. D., Howard, R. E., Fetter, L. A., Epworth, R. W., and Tennant, D. M., 1984, Discrete Resistance Switching in Submicrometer Silicon Inversion Layers: Individual Interface Traps and Low-Frequency (1/f?) Noise, Phys. Rev. Lett. 52:228.
Ralls, K. S. and Buhrman, R. A., 1988, Defect Interactions and Noise in Metallic Nanoconstrictions, Phys. Rev. Lett. 60:2434.
Robnik, M., 1986, Perimeter Corrections to the Landau Diamagnetism, J. Phys. A. 19:3619.
Rogers, C. T. and Buhrman, R. A., 1985, Nature of Single-Localized-Electron States Derived from Tunneling Measurements, Phys. Rev. Lett. 55:859.
Sikorski, C. and Merkt, U., 1989, Spectroscopy of Electronic States in InSb Quantum Dots, Phys. Rev. Lett. 62:2164.
Sivan, U., and Imry, Y., 1988, de Haas-van Alphen and Aharonov-Bohm-type Persistent Current Oscillations in Singly Connected Quantum Dots, Phys. Rev. Lett. 61:1001.
Skocpol, W. J., Jackel, L. D., Hu, E. L., Howard, R. E., and Fetter, L. A., 1982, One-Dimensional Localization and Interaction Effects in Narrow (0.1-mm) Silicon Inversion Layers, Phys. Rev. Lett. 49:951.
Skocpol, W. J., Mankiewich, P. M., Howard, R. E., Jackel, L. D., Tennant, D. M. and Stone, A. D., 1986, Universal Conductance Fluctuations in Silicon Inversion-Layer Nanostructures, Phys. Rev. Lett. 56:2865.
Smith III, T. P., Lee, K. Y., Knoedler, C. M., Hong, J. M. and Kern, D. P., 1988, Electronic Spectroscopy of Zero-Dimensional Systems, Phys. Rev. B :2172.
Thouless, D. J., 1977, Maximum Metallic Resistance in Thin Wires, Phys. Rev. Lett. 39:1167.
Thouless, D. J., 1980, The Effect of Inelastic Electron Scattering on the Conductivity of Very Thin Wires, Solid State Comm. 34:683.
Uren, M. J., Day, D. J. and Kirton, M. J., 1985, 1/f and Random Telegraph Noise in Silicon Metal-Oxide-Semiconductor Field-Effect Transistors, Appl. Phys. Lett. 47:1195.
Van Hove, L., 1957, The Approach to Equilibrium in Quantum Statistics: A Perturbation Treatment to General Order, Physica 23:441.
Van Wees, B. J., van Houten, H., Beenakker, C. W. J., Williamson, J. G., Kouwenhoven, L. P., and van der Marel, D., 1988, Quantized Conductance of Point Contacts in a Two-Dimensional Electron Gas, Phys. Rev. Lett. 60:848.
Wang, L., and O'Connell, R. F., 1986, Surface Effects on the Diamagnetic Susceptibility and Other Properties of a Low-Temperature Electron Gas, Phys. Rev. B 34:5160.
Webb, R. A. and Washburn, S., Dec. 1988, Quantum Interference Fluctuations in Disordered Metals, Phys. To-Day 41:12, pps. 46-53.
Wharam, D. A., Thornton, T. J., Newbury, R., Pepper, M., Ahmed, H., Frost, J. E. F., Hasko, D. G., Peacock, D. C., Ritchie, D. A., and Jones, G. A. C., 1988, One-Dimensional Transport and the Quantisation of the Ballistic Resistance, J. Phys. C 21:L209.

NATO ASI on Granular Nanoelectronics

Il Ciocco, Italy

23. July - 04. August 1990

RAMAN SPECTROSCOPY FOR THE CHARACTERIZATION OF LOW DIMENSIONAL SYSTEMS

Gerhard Abstreiter

Walter Schottky Institut

Techn. Univ. München

D-8046 Garching

The following aspects are discussed in the lecture

1. Growth aspects for the achievements of high quality Si/Ge heterostructures and superlattices

   - critical thickness
   - interface sharpness
   - phonon Raman spectroscopy as a sensitive local probe of various interface and superlattice properties
   - electronic and optical properties

2. Electronic excitations in low dimensional systems with emphasis on GaAs based heterostructures

   - basic concepts of single particle and plasmon excitations in 3,2,1, and 0 dimensional systems
   - q-dependence of electronic excitations in various dimensions
   - experimental light scattering techniques
   - discussion of available experimental results

3. Recent advances and new approaches for the achievements of lateral microstructures - direct analysis with spectroscopical techniques with high spatial resolution (Raman and luminescence)

From: SPECTROSCOPY OF SEMICONDUCTOR MICROSTRUCTURES
Edited by Gerhard Fasol, Annalisa Fasolino,
and Paolo Lugli
(Plenum Publishing Corporation, 1989)

PHONONS AND OPTICAL PROPERTIES OF SI/GE SUPERLATTICES

G. Abstreiter, K. Eberl, E. Friess, W. Menczigar
W. Wegscheider, and R. Zachai

Walter Schottky Institut, Technical University Munich
D-8046 Garching, Fed. Rep. of Germany

INTRODUCTION

Short period Si/Ge superlattices are new semiconductor materials whose band structure and consequently whose electrical and optical properties can be changed in a wide range. New device applications are expected on the basis of such layered structures [1]. Recent progress on low temperature molecular beam epitaxial growth [2,3,4,5] allows the realization of high quality Si/Ge superlattices with sharp interfaces and individual layer thicknesses of only a few monolayers. The large lattice mismatch of more than 4 % between the two constituents, however, still causes major problems for the achievement of sufficient total thickness, which is required for the application of new superlattice effects. The concept of strain symmetrization with certain buffer layers [4,6] might be one way to overcome this problem. Various basic properties of such new superlattice materials can be studied, however, also in relatively thin Si/Ge superlattices grown on Si, Ge, and SiGe substrates. In the present article results, obtained mainly in our group are reviewed. The excellent work of various other research groups from all over the world can be found in the literature [7].

In the next section we describe growth and structural properties which are studied "in situ" by LEED and Auger spectroscopy and "ex situ" by electron microscopy. Phonon properties, as analyzed by Raman spectroscopy, are discussed in the third section. Folded acoustical and confined optical modes lead, for example to information on period length, strain distribution interface sharpness. Phonons are also a sensitive tool for interdiffusion. The paper ends with a short discussion of optical properties of certain strain symmetrized Si/Ge superlattices. Photoluminescence experiments show evidence for new fundamental energy gaps due to band folding, resonant Raman scattering leads to information on higher energy gaps.

GROWTH AND STRUCTURAL PROPERTIES

A specially designed MBE system is used for low temperature

growth of Si/Ge superlattices. The main UHV chamber is equipped with Si and Ge evaporation sources, a quadrupole mass spectrometer for rest gas analysis, a special substrate holder and tools for surface analysis like LEED and AES. The growth conditions are very crucial, especially in the case of ultrashort period superlattices with individual layer thicknesses of only a few monolayers of pure Si and pure Ge. Characteristic growth conditions are for example growth rates of typically one to two atomic layers per minute and vacuum conditions in the low 10^{-10} mbar range during growth. The difference in lattice constant of about 4 % leads to a critical thickness of lattice matched and 2-dimensional growth for Si/Ge heterostructures. Ge can be grown pseudomorphic directly on Si (100), for example, only up to 3 to 4 monolayers at substrate temperatures of about 300°C to 350°C. After 3 atomic layers of Ge we observe already considerable surface roughness by evaluating the energy dependence of LEED spot profile. This indicates the beginning of strain relaxation by formation of misfit dislocations and 3-dimensional growth which is then clearly observed for thicknesses exceeding 6 monolayers of Ge. For the inverse situation of Si on Ge (100) substrate a slightly larger critical thickness is observed, and the crystalline quality is found to be much better with respect to surface flatness.

Fig. 1 shows a cross sectional transmission electron micrograph of a 20 period Si_3Ge_9 superlattice grown on a Ge (100) substrate. No misfit dislocations or defects are observed in the whole sample which was investigated. The growth temperature was as low as 310°C. The lateral lattice constant within the superlattice is equal to that of the Ge substrate. The thickness of the individual Si layers is below the critical value. Consequently the Si layers are extended by 4 % in lateral direction, whereas the Ge layers are unstrained. The total thickness of the superlattice is small enough, that also the second critical thickness, which arises from the extremely asymmetrical strain distribution between the layers, is not reached yet. In an equivalent structure with 120 periods (total thickness ~ 2000 Å) on the other hand, a lot of misfit defects are observed as shown in [5].

Fig. 1. Cross-sectional TEM micrograph of a 20 period Si_3Ge_9 superlattice on Ge (100).

To overcome the problem of thickness limitation in these strained layer superlattices, it was proposed to introduce Si_xGe_{1-x} alloy buffer layers, which provide a lateral lattice constant in between those of Si and Ge |3,8|. Recently, we have studied different types of strain adjusting buffer layers |6| which vary in composition and thickness.

Fig. 2 shows a TEM micrograph pf a 30 period Si_4Ge_4 superlattice grown on a Si (100) substrate with a relatively complicated multi layer buffer. It is composed of 30 ML ge, 150 ML Si_3Ge_3 superlattice and 40 ML Ge. The Si_3Ge_3 superlattice was introduced in order to achieve a surface smoothening effect. The spot separation of the diffraction patterns in LEED and TEM are used to determine the difference of the lateral lattice constants between the Si substrate and the Si_4Ge_4 superlattice. We obtain $\delta a/a_{Si}$ = 2.4 % ± 0.1 %, which means that the Si layers are extended in lateral direction by 2.4 %, whereas the Ge layers are compressed by about 1.8 %. With this multi-layer buffer we tried to keep the thickness of the region where $a_{\parallel}$ is changed as small as possible. The clear contrast of the individual layers in Fig. 2 reflects the high quality interfaces whose sharpness has been studied in detail also by Auger electron spectroscopy performed during growth interruptions |3,6|. Fig. 2 demonstrates that short period superlattices with different strain distributions can be achieved on Si, that the quality can be improved by introducing buffer layers, but that there is still a considerable amount of crystalline defects across the whole structure. Further improvement of strain adjusting buffer layers is necessary in order to reduce the defect density within the superlattice and to achieve a crystal quality which is required for improved optical and electrical properties.

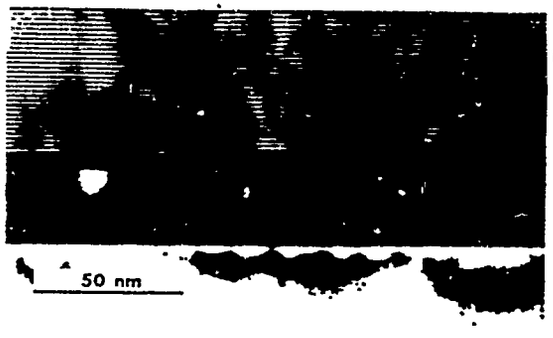


Fig. 2. Cross-sectional TEM micrograph of a 30 period Si_4Ge_4 superlattice grown on Si (100) with a (30 ML Ge/ 150 ML Si_3Ge_3 SL/40 ML Ge) buffer layer

PHONON PROPERTIES

First order phonon Raman spectroscopy of Si and Ge reveals one single narrow peak at the energy of the optical phonons close to the Brillouin zone center (= 520 cm^{-1} in Si, = 300 cm^{-1} in Ge). The main features of SiGe alloys consist of three asymmetrically broadened lines close to the Si and Ge mode and in between at about 400 cm^{-1}. The exact positions and intensities

depend on composition and strain |9|. The phonon spectra of Si/Ge superlattices are drastically different from the alloy spectra.

Fig. 3 shows the Raman spectrum of a Si_4Ge_{12} superlattice grown on a Ge (100) substrate. Various additional features are observable which reflect the artificial order of the sample. In the low energy range, below = 250 cm^{-1}, several, so called folded acoustical modes appear in the spectrum. In this energy range acoustical phonon branches of Si and Ge overlap, the modes can propagate through the alternating layers of the superlattice. Within the elastic continuum theory these modes can be understood by asuming an average sound velocity whose dispersion shows a backfolding and splitting at the new superlattice Brillouin zone boundary π/d and at the zone center |10,11|. The exact energy of the folded doubletts is very sensitive on the period length d. The simple model, however, is only valid for energies smaller than about 150 cm^{-1} to 200 cm^{-1}. Above these values the acoustic branches deviate from a linear behaviour |12,13,14|. This is important for very small periods which consist of only a few monolayers and for the higher order folded modes shown also in Fig. 3. Already the first folded doublet is in the nonlinear range for period lengths of the order of 10 monolayers or less. Additional information, which can be obtained from folded acoustical modes are for example the fluctuation in the period length which is reflected in a broadening |15| and the abruptness of interfaces, which determine the intensity decay of the higher order modes. The intensities are closely related to the Fourier coefficients of the concentration profile within the sample |10|.

In contrast to the acoustic phonons there is no overlap of the dispersion curve of the optical branches of bulk Si and Ge. Consequently optical modes cannot propagate but are confined within the corresponding slabs and thus provide information on the individual thin layers. This is not perfectly rue for the Ge modes which overlap with the longitudinal acoustical (LA) branch of Si and therefore are expected to have some dispersion in the superlattice direction. The spectrum of Fig. 3 shows a series of closely spaced peaks at about 300 cm^{-1}, the "confined" modes in the 12 monolayer thick Ge layers.

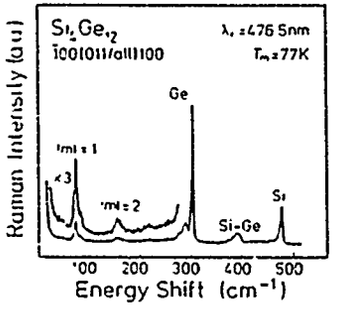


Fig. 3. Raman spectrum of a high quality Si_4Ge_{12} strained layer superlattice on a Ge (100) substrate (from Ref. |5|).

Such confined modes have been observed in Si/Ge superlattices in thin layers of Ge in (110) and (100) direction |16,3| and in Si grown in (100) direction |17|. In a simple model, similar as used for $GaAs/Al_xGa_{1-x}As$ short period superlattices |18|, the confined modes can be considered as standing waves which have to fit into the slab. Such an analysis leads to the bulk phonon dispersion of Si and Ge which in this way is accessable to the accurate Raman spectroscopy. One has to be careful, however, especially in the case of Ge where the coupling to the LA branch has to be taken into account |12|. Higher order confined modes are sensitve to interface roughness and are consequently also a good measure of the sample quality.

The sharp phonon mode at about 480 cm^{-1} in Fig. 3 originates from the 4 monolayer thick Si slabs. It is the first order confined optical mode. The major part of the large shift from the bulk value of about 520 cm^{-1} is, however, not due to confinement but due to the large built-in strain. This sample is grown on a Ge substrate. Consequently the lateral lattice constant of the Si layers is increased by about 4 %. This leads to a downward shift of the Si phonon modes of more than 30 cm^{-1}. The energetic position of the Si and Ge optical modes can be used as a sensitive tool to measure the strain distribution in short period Si/Ge superlattices |4,16|.

The weak structure at about 390 cm^{-1} in Fig. 3 is due to Si-Ge vibrations. The strength of this peak is extremely sensitive to interface roughness as can be seen in Fig. 4. A series of spectra of the optical phonon region is shown for a "strain-symmetrized" sample consisting of 12 monolayers of Si and 8 monolayers of Ge. Each spectrum was obtained after a different annealing step. A pronounced increase of the Si-Ge alloy mode, concomitant with a shift of the Ge mode reflects the intermixing at the originally sharp interfaces. From such studies one obtains also information on diffusivities of Si in Ge and Ge in Si, respectively, and how it depends on strain distribution |6,19|.

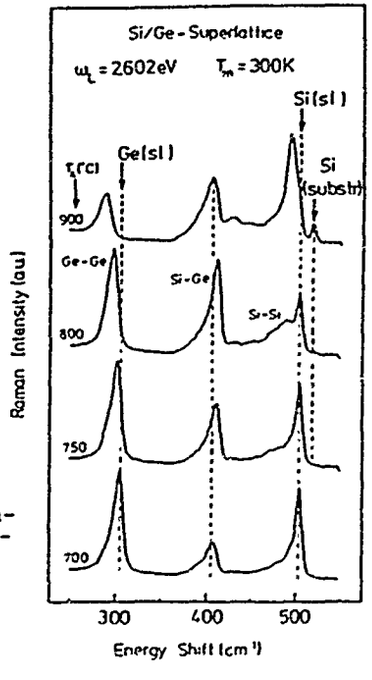


Fig. 4. Raman spectra of a strain-symmetrized $Si_{12}Ge_8$ superlattice after annealing steps (from |19|)

BAND STRUCTURE AND OPTICAL PROPERTIES

Strain, confinement and superlattice effects have a strong influence on the band structure of short period Si/Ge superlattices. Recent experimental results on electro reflectance |20|, photoconductivity |21| and photoluminescence |5,17,22| showed a variety of new optical transitions in strained Si/Ge superlattices in the range from 0,7 eV to 3,5 eV. Band structure calculations |23-27| try to explain these new transitions due to a combination of band folding and strain effects. The assignments, however, are still controversial. Under certain conditions a new quasi-direct fundamental energy gap is expected in the Si/Ge system. This has been proposed already in 1974 by Gnutzmann and Clausecker |28| and more recently by People and Jackson |29| on the basis of zone-folding arguments. Most of the relevant features and the dependence of the energy gap on strain distribution, Ge and Si thickness ratio, and period length can be seen already from a simple picture of band offsets, which takes the splitting of the conduction and valence bands due to strain into account. A quasi direct energy gap requires a folding of the twofold-degenerate Si conduction band minimum along the growth direction ($\Delta_\perp$) back to the zone center. This is best achieved for period lengths of 5, 10 or 20 monolayers, because the Δ minimum is at about 0.8 times the Brillouin zone boundary wave vector in (100) direction. It is however also necessary that the $\Delta_\perp$ minima are lower in energy than the fourfold degenerate in-plane minima ($\Delta_\parallel$). This can be achieved by tensile biaxial strain in the Si layers. The strain has also a large influence on the band alignment. This is shown in Fig. 5 where the band offsets are compared for different lateral lattice constants ($a_\parallel = a_\perp(Si)$, $a_\parallel = a_\perp(Si_{0.5}Ge_{0.5})$ and $a_\parallel = a_\perp(Ge)$). The top of the valence band is always highest in Ge. The band offset for the heavy hole valence band is between 750 meV and 800 meV, nearly independent of strain distribution. The band discontinuity for light holes is decreasing with increasing strain in Si and the average energy is shifting upwards. The conduction band is lowest in Si in all cases (staggered band line-up). The two-fold $\Delta_\perp$ minima shift to low energies with increasing lateral lattice constant, a quasi direct energy gap is expected. The expected fundamental energy gap is decreasing with increasing strain in Si.

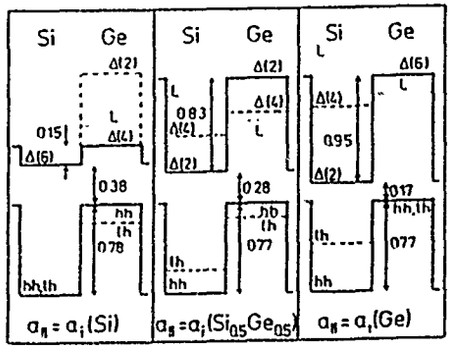


Fig. 5. Band alignment in Si/Ge superlattice with lateral lattice constants $a_\parallel$ of Si, $Si_{0.5}Ge_{0.5}$, and Ge in (100) direction (from |17|)

First evidence for a quasi-direct energy gap was recently published |5,17,22| for strain symmetrized Si_6Ge_4 superlattices. Strong photoluminescence was observed at about 850 meV from a 2000 Å thick superlattice on a $Si_{0.4}Ge_{0.6}$ buffer layer. The actual strain distribution in the sample was 1.4 % tensile in Si and about 2.7 % compressive in Ge, as determined from Raman spectroscopy. Very recently similar luminescence signals were observed in Si_6Ge_4 superlattices with different strain distribution |22|. Results are shown in Fig. 6. As expected from theory, the luminescence peak energy shifts from 0.84 eV (1.4 % strain in Si) to about 0.77 eV for 2.4 % strain in Si. This strongly supports the ideas of new quasi-direct energy gaps in strained Si/Ge superlattices, however, more work on these structures is necessary to prove unambigiously that the observed radiative recombination is due to intrinsic band structure effects.

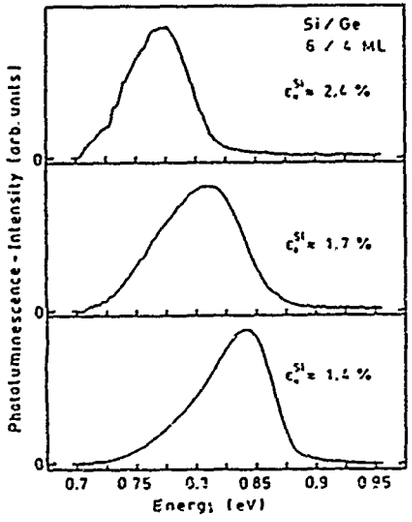


Fig. 6. Photoluminescence of Si_6Ge_4 superlattices with different strain distribution (from |22|)

Electroreflectance has been used as a powerful tool to investigate new energy gaps in Si/Ge superlattices in a wide range |20|. New gaps have been observed both in the near infrared region (direct, indirect) and above 2 eV where the E_1 and $E_1 + \Delta_1$ related gaps are expected. Electroreflectance cannot distinguish between confined and extended electronic states in the superlattice. Recently resonant Raman scattering has been used to get detailed information on the higher energy gaps |30, 31|. Due to the strong confinement of the optical phonons in Si and Ge the resonance enhancement can probe confined electronic states in each layer separately and can distinguish between confined and extended electronic states. The resonance enhancement of the Si and Ge optical phonons for two Si_3Ge_3 superlattices grown on Ge (100) substrates below and above the total critical thickness is shown in Fig. 7. For comparison we show also the resonance curves of Si and Ge like modes of a $Si_{0.25}Ge_{0.75}$ alloy, which was grown on a Ge (100) substrate. The intensities of each Raman peak were measured relative to that of a bulk Si sample mounted next to each superlattice.

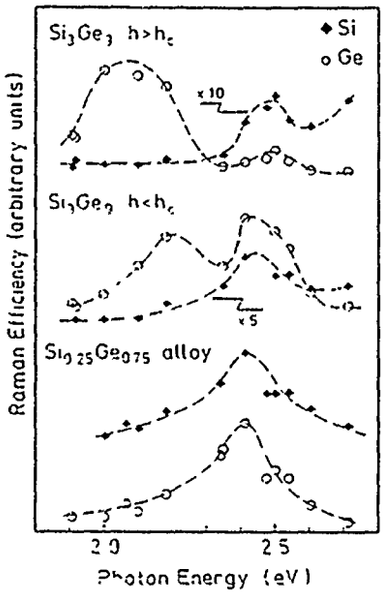


Fig. 7. Resonance enhancement of Si and Ge optical phonons in Si_3Ge_3 superlattices and $Si_{0.25}Ge_{0.75}$ alloy

Two distinct resonance features are observed for the superlattice samples. The Ge optical phonons resonate at about 2.2 eV and 2.45 eV for the asymmetrically strained Si_3Ge_3 superlattice (total thickness 330 Å). The partially relaxed superlattice (total thickness 2200 Å) shows resonance peaks at about 2.1 eV and 2.5 eV. The Si phonon mode resonates only at higher energies. Thus we conclude that the lower resonance corresponds to a energy gap with electronic states confined strongly in the Ge layers. Both modes show a resonance at the higher energy, which is evidence for more extended bands in this energy range. The phonons of the $Si_{0.25}Ge_{0.75}$ alloy show only one resonance peak at about 2.4 eV for all three modes, known as the E_1 and $E_1 + \Delta_1$ resonance. The superlattice resonances evolve from the E_1 gap of the corresponding alloy. A similar splitting has been observed in electroreflectance when the superlattice band structure develops. The Raman resonances of the more extended band gap shifts to higher energies with decreasing period length and/or increasing strain in the Ge layers |31|. The splitting between the two resonances on the other hand is increasing with increasing period length or increasing strain in the Ge layers |31|.

The results reported here are first spectroscopical investigations on details of the new superlattice band structure grown on various substrates. They clearly demonstrate that it is indeed possible to realize new semiconductors by growing short period superlattices, whose bandstructure evolves from the bandstructures of the constituents, but can have properties which are widely different from the corresponding alloys.

ACKNOWLEDGEMENTS

It is our pleasure to thank E. Kasper and his coworkers at the AEG research laboratories in Ulm for the fruitful collaboration especially with respect to the strain symmetrized short period superlattices. The strain symmetrized samples grown by H. Kibbel at AEG were the first one, which showed strong photoluminescence in the infrared spectral region. We had also an excellent collaboration with H. Oppolzer and H. Cerva at the Siemens research laboratories in München-Neu Perlach, where the high

quality TEM micrographs were produced. Part of the work was supported financially by the Siemens AG via SFE.

REFERENCES

1. see for example: E. Kasper, SiGe/Si superlattices strain influence and devices, in: "Heterostructures on Silicon: One step further with Silicon," Y. I. Nissim, and E. Rosencher, eds., Nato ASI Series Vol. 160, Dordrecht, (1989).
2. J. C. Bean, L. C. Feldmann, A. T. Fiory, S. Nakahara, and J. K. Robinson, Ge_xSi_{1-x}/Si strained-layer superlattices grown by molecular beam epitaxy, J. Vacuum Sci. Technol. A2: 436 (1984).
3. K. Eberl, W. Wegscheider, E. Friess, and G. Abstreiter, Realization of short period Si/Ge strained layer superlattices, in: ref. 1.
4. E. Kasper, H. Kibbel, H. Jorke, H. Brugger, E. Friess, and G. Abstreiter, Symmetrically strained Si/Ge superlattices on Si substrates, Phys. Rev. B 38: 3599 (1988).
5. G. Abstreiter, K. Eberl, E. Friess, W. Wegscheider, and R. Zachai, Silicon/Germanium strained layer superlattices, Journal of Crystal Growth 95:431 (1989).
6. K. Eberl, E. Friess, W. Wegscheider, U. Menczigar, and G. Abstreiter, Improvement of structural properties of Si/Ge superlattices, E-MRS Meeting, Strasbourg (1989), to be published in: Thin Solid Films.
7. see for example publications in conference proceedings of Refs. 1 and 6.
8. E. Kasper, H. J. Herzog, H. Jorke, and G. Abstreiter, Strain adjustment in Si/Ge superlattices, Mat. Res. Soc. Symp. Proc. Vol. 102:393 (1988).
9. M. A. Renucci, J. B. Renucci, and M, Cardona, Raman scattering in Ge-Si alloys, in: "Light scattering in Solids," M. Balkanski, ed., Flammarion, Paris (1971).
10. C. Colvard, T. A. Gant, M. V. Klein, R. Merlin, R. Fischer, H. Morkoc, and A. C. Gossard, Folded acoustic and quantized optic phonons in (GaAl)As superlattices, Phys. Rev. B31:2080 (1985).
11. H. Brugger, G. Abstreiter, H. Jorke, H. J. Herzog, and E. Kasper, Folded acoustic phonons in Si-Si_xGe_{1-x} superlattices, Phys. Rev. B33:5928 (1986).
12. E. Molinari, A. Fasolino, Calculated phonon spectra of Si/Ge (001) superlattices: Features for interface characterization, Appl. Phys. Lett. 54:1220 (1989).
13. M. I. Alonso, F. Cerdeira, D. Miles, M. Cardona, E. Kasper, and H. Kibbel, Raman spectra of Si_nGe_m superlattices: theory and experiment, preprint.
14. J. White, G. Fasol, R. Ghanbari, C. J. Gibbings, and C. G. Tuppen, Calculation of energies and Raman intensities of confined phonons in Si-Ge strained layer superlattices, in: Ref. 6.
15. M. Ospelt, K. A. Mäder, W. Bacsa, J. Henz, and H. von Känel, Unstrained vs. strained layer epitaxy: Thick Ge-layers and Ge/Si superlattices on Si(100), in: Ref. 1.
16. E. Friess, H. Brugger, K. Eberl, G. Krötz, and G. Abstreiter, Confined optical modes in short period (110) Si/Ge superlattices, Solid State Communications 69:899 (1989).

17. R. Zachai, E. Friess, G. Abstreiter, E. Kasper, and H. Kibbel, Band structure and optical properties of strain symmetrized short period Si/Ge superlattices on Si (100) substrates, in: "19th International Conference on the Physics of Semiconductors," W. Zawadzki, ed., Insitute of Physics, Polish Academy of Sciences, Warsaw (1988).
18. see for example A. K. Sood, J. Menéndez, M. Cardona, and K. Ploog, Resonance Raman scattering by confined LO and TO phonons in GaAs-AlAs superlattices, Phys, Rev. Lett. 54:2111 (1985).
19. H. Brugger, E. Friess, G. Abstreiter, E. Kasper, and H. Kibbel, Annealing effects in short period Si-Ge strained layer superlattices, Semicond. Sci. Technol. 3:1166 (1988)
20. T. P. Pearsall, J. Bevk, L. C. Feldmann, J. M. Bovar, J. P. Mannaerts, and A. Ourmazd, Structurally induced optical transitions in Si-Ge superlattices, Phys. Rev. Lett. 58:729 (1987), T. P. Pearsall, Germanium-Silicon alloys and Heterostructures - Optical and electronic properties, CRC Critical Rev. of Solid State and Materials Sciences, in press.
21. D. V. Lang, R. People, J. C. Bean, and A. M. Sergent, Measurements of the band gap of Ge_xSi_{1-x}/Si strained layer heterostructures, Appl. Phys. Lett. 47:1333 (1985).
22. R. Zachai, K. Eberl, and G. Abstreiter, Photoluminescence in Si/Ge superlattices with different strain distributions, to be published.
23. S. Froyen, D. M. Wood, and A. Zunger, Structural and electronic properties of epitaxial thin-layer Si_nGe_n superlattices, Phys. Rev.B 37:6893 (1988).
24. S. Satpathy, R. M. Martin, and C. G. van de Walle, Theory of electronic properties of the (100) Si/Ge strained-layer superlattices, Phys. Rev. B 38:13237 (1988).
25. M. S. Hybertsen and M. Schlüter, Theory of optical transitions in Si/Ge (001) strained-layer superlattices, Phys. Rev. B 36:9683 (1987).
26. I. Morrison and M. Jaros, Electronic and optical properties of ultrathin Si/Ge (001) superlattices, Phys. Rev. B 37: 916 (1988).
27. M. A. Gell, Effect of buffer-layer composition on new optical transitions in Si/Ge short-period superlattices, Phys. Rev. B 38:7535 (1988).
28. U. Gnutzmann and K. Clausecker, Theory of direct optical transition in an optical indirect semiconductor with a superlattice structure, Appl. Phys. B 36:1310 (1987).
29. R. People and S. A. Jackson, Indirect, quasi-direct, and direct optical transitions in the pseudomorphic (4 x 4)-monolayer Si-Ge strained -layer superlattices on Si (001), Phys. Rev. B 36:1310 (1987).
30. F. Cerdeira, M. I. Alonso, D. Niles, M. Garriga, M. Cardona, E. Kasper,and H. Kibbel, Resonant Raman scattering in short-period Si_nGe_m superlattices, preprint.
31. U. Menczigar, E. Friess, K. Eberl, and G. Abstreiter, Resonant Raman scattering in ultrathin Si/Ge strained-layer superlattices, to be published.

6th Int. Winterschool, Mauterndorf (1990)

Inelastic Light Scattering by Electrons in Microstructured Quantum Wells

Gerhard Abstreiter and Thomas Egeler
Walter Schottky Institut, Techn. Universität München
D–8046 Garching

Abstract. Experimental results of electronic excitations in microstructured GaAs quantum wells are discussed. Micro–Raman spectroscopy and grating coupler effects are used to extend the accessible wave vector range of in–plane single particle and plasmon excitations in multi–layered two–dimensional electron systems. In lateral quantum wire structures a strongly anisotropic dispersion of the plasmon resonances is observed, which can be understood in terms of plasmons confined normal to the wires.

1. Introduction

The first observation of inelastic light scattering by free carriers in semiconductors was reported by Mooradian and Wright [1] in doped n–type GaAs more than twenty years ago. In this pioneering work plasmons of a degenerate electron gas, coupled to LO–phonons were studied. Shortly afterwards Mooradian [2] also reported light scattering by single particle excitations in n–GaAs. The early work stimulated extensive light scattering research of electron and hole plasmas in semiconductors (for reviews see Refs. [3,4,5]). Especially resonant inelastic light scattering experiments [6] showed that the method is sensitive enough to observe electronic excitations with densities of the order of 10^{11} cm^{-2}. This was followed by the proposal [7] and the first observations of resonant light scattering by quasi–two–dimensional (2D) electron systems in selectively doped GaAs–(AlGa)As heterojunctions [8] and multiple quantum wells [9] about ten years ago. Numerous publications have appeared since then, which demonstrate the versatility of light scattering for the study of various properties of 2D carrier systems. These include single–particle and collective intersubband transitions in various potential wells which allow a separation of the depolarization shift, 2D plasmons in layered electron systems and in–plane single particle excitations. The light scattering work of 2D–systems of the past ten years has been reviewed for example in [10,11]. So far the dispersion of the different excitations could be studied only in a limited wave vector range ($q \leq 1.7 \times 10^5$ cm^{-1}) in normal

backscattering geometry from the surfaces. In the present paper we first review shortly new developments in electronic light scattering by 2D systems with large scattering wave vectors which become accessible by micro–Raman scattering [12] and by using grating coupler effects [13,14]. In the last part we concentrate on first light scattering results in wire structured electron systems in GaAs quantum wells [15,16].

2. Electronic excitations in systems of different dimensionality

Elementary excitations of free carrier systems fall into two main categories: collective excitations and single particle excitations. Resonant electronic light scattering allows a separation of the two kinds of excitations by analyzing the polarization of incident and scattered light. Polarized spectra (incident and scattered light polarized parallel to each other) usually display collective excitations, i.e. plasmons which are coupled to LO–phonons in polar semiconductors. Single–particle excitations are observed in crossed polarizations. They are characterized by uncorrelated excitations of electrons below the Fermi energy to empty states above the Fermi energy. In backscattering geometry the scattering wave vector is given by $q \simeq 2 \cdot 2\pi\eta/\lambda$, which is about $7 \cdot 10^5$ cm^{-1} for GaAs in the red spectral range (η is the refractive index and λ the laser wavelength). It leads to the kind of excitations shown in Fig. 1a for an isotropic 3D–electron system. The cutoff frequency of the single particle excitations is at about $q \cdot v_f$, where v_f is the Fermi velocity. In 3D they have approximately triangular lineshape. The plasmon frequency is proportional to $\sqrt{n_{3D}}$, with n_{3D} being the three–dimensional carrier concentration. It exhibits a weak dispersion which in random phase approximation (RPA) is given by $\omega_p^2(q) = \omega_p^2(0) + 3q^2 v_f^2/5$.

In 2D–systems, which are realized for example in narrow quantum wells or heterostructures, the motion of the carriers is quantized normal to the layers and is free in parallel direction ($q_{\parallel}$). Typical excitations of such systems are shown in Fig. 1b. In usual backscattering geometry normal to the layers ($q_{\parallel} = 0$) collective and single particle intersubband excitations are observed, which reveal the subband splittings ω_{01} and the depolarization shift. Recent experiments on extremely high mobility samples seem to indicate that depolarized spectra (scattering via spin density fluctuation) do not give exactly the one particle energies ω_{01}, but are subject to measureable exciton–like shifts to lower energies [17], which were neglected in most of the earlier experiments in GaAs. In–plane single–particle and collective excitations are only observed for finite $q_{\parallel}$. Due to the high refractive index of GaAs the maximum achievable in–plane wave vector in backscattering from the surface is given by $q_{\parallel} = (4\pi/\lambda) \sin\alpha \lesssim 1.7 \times 10^5$ cm^{-1} for laser energies around the E_0 and $E_0 + \Delta_0$ energy gap. The angle α is defined between the incident (scattered) light and the

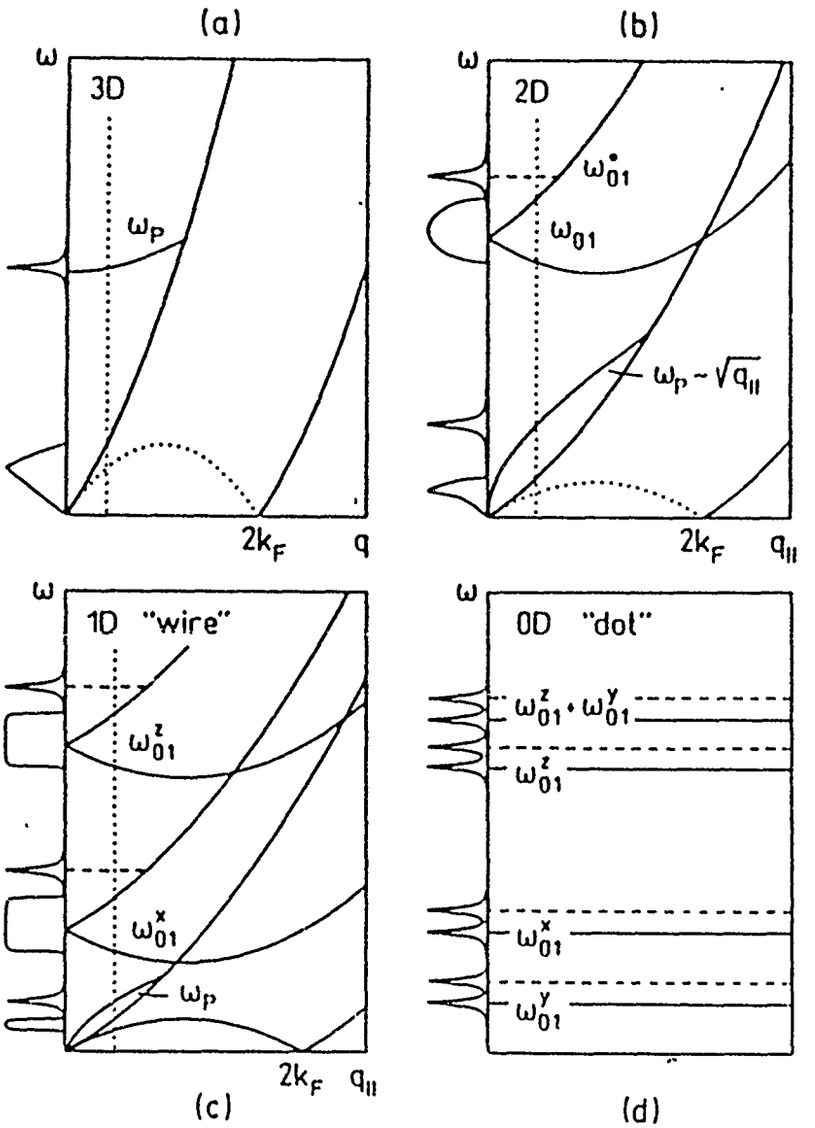


Fig. 1 Single particle and collective excitations in 3, 2, 1 and 0 dimensional electron systems (schematically)

normal to the sample surface. The dispersion of both plasmons and single particle excitations of layered electron systems have been studied in this limited wave vector range [10,11,18–21]. An extension of the accessible in–plane wave vector is discussed below (see also [12,13,14]).

In Fig. 1c and 1d the possible excitations in 1D and 0D systems are shown schematically. In quantum dots, confinement exists in all three space directions which finally results in an atomic–like situation with no dispersion in any direction. In 1D systems (quantum wires) free motion of carriers is still possible along one direction ($q_{||}^y$). For the other two directions different confining potentials are assumed in

Fig. 1c. This leads to different single particle and collective excitations in all three directions. The restriction of free motion in one dimension also changes the single particle excitation along the wire drastically. No low frequency excitations are possible for a finite wave vector $q_{\parallel}^{y}$ along the wire. First experimental results of electronic excitations in wire structured quantum wells or heterostructures have been published recently [15,16]. Our results, especially on confined plasmons, will be discussed in this paper. To the best of our knowledge, no results have been published so far for quantum dots.

3. In–plane excitations in layered electron systems

Until recently in–plane excitations in 2D systems were observed only in a relatively small wave vector range. In backscattering from the surface of a layered electron gas the total wave vector is given by $q = 2 \cdot 2\pi\eta/\lambda$, which divides into $q_{\parallel} = (4\pi/\lambda)\cdot\sin\alpha$ and $q_{\perp} = (4\pi\eta/\lambda)\sqrt{1-\sin^2\alpha/\eta^2}$, depending on the angle α. A maximum in–plane wave vector is achieved in back–scattering geometry from the edge of the layered structure (cleavage plane). Such geometries have been realized recently by using low temperature micro–Raman scattering [12]. Results of single–particle excitations and plasmon excitations in cleaved and wedged GaAs multilayer structures are shown in Fig. 2 and 3. The maximal achievable in–plane wave vector is $q_{\parallel} = 4\pi\eta/\lambda \simeq 7 \cdot 10^5$ cm^{-1} using a red laser line for excitation. The sample consists of 25 GaAs quantum wells with a thickness of 60 Å and a carrier concentration of about 9×10^{11} cm^{-2} per well. The in–plane single particle excitations shown in Fig. 2 exhibit the characteristic lineshapes of 2D systems. The solid lines are calculated lineshapes using the Lindhard–Mermin dielectric function of the 2D electron systems. All spectra can be fitted by assuming a temperature of the electron gas of $T_F \simeq 13$ K. Scattering from the as grown surface ($q_{\parallel} = 1.3 \cdot 10^5$ cm^{-1} for $\alpha = 45°$) and from the cleavage plane reveal a carrier concentration of about 9×10^{11} cm^{-2} and an electron damping of $\Gamma_F \simeq 1$meV. The polished surfaces show somewhat lower quality which results in a slightly reduced carrier concentration and larger damping.

Plasmons in layered electron systems are characterized by the Coulomb interaction between the layers. The dispersion of a purely 2D system is given by $\omega_{2D}^2(q_{\parallel}) = 2\pi n_{2D} e^2 q_{\parallel}/m^*\epsilon$, where n_{2D} is the 2D carrier concentration and ϵ the dielectric constant. The coupling between the layers leads to a plasmon band which is determined for an infinite system by $\omega(q_{\parallel},q_{\beta}) = \omega_{2D}(q_{\parallel})\, S\,(q_{\parallel},q_{\beta})$, where $S\,(q_{\parallel},q_{\beta}) = \sqrt{\sinh q_{\parallel} d/(\cosh q_{\parallel} d - \cos q_{\beta} d)}$ is a structure factor. For a finite number N of electron gas layers N plasmon branches are obtained [22] which can be approximated by the above–mentioned formula with $q_{\beta} = (2\pi/Nd)\beta$ with $\beta = 0,1,\ldots N-1$ for a large number

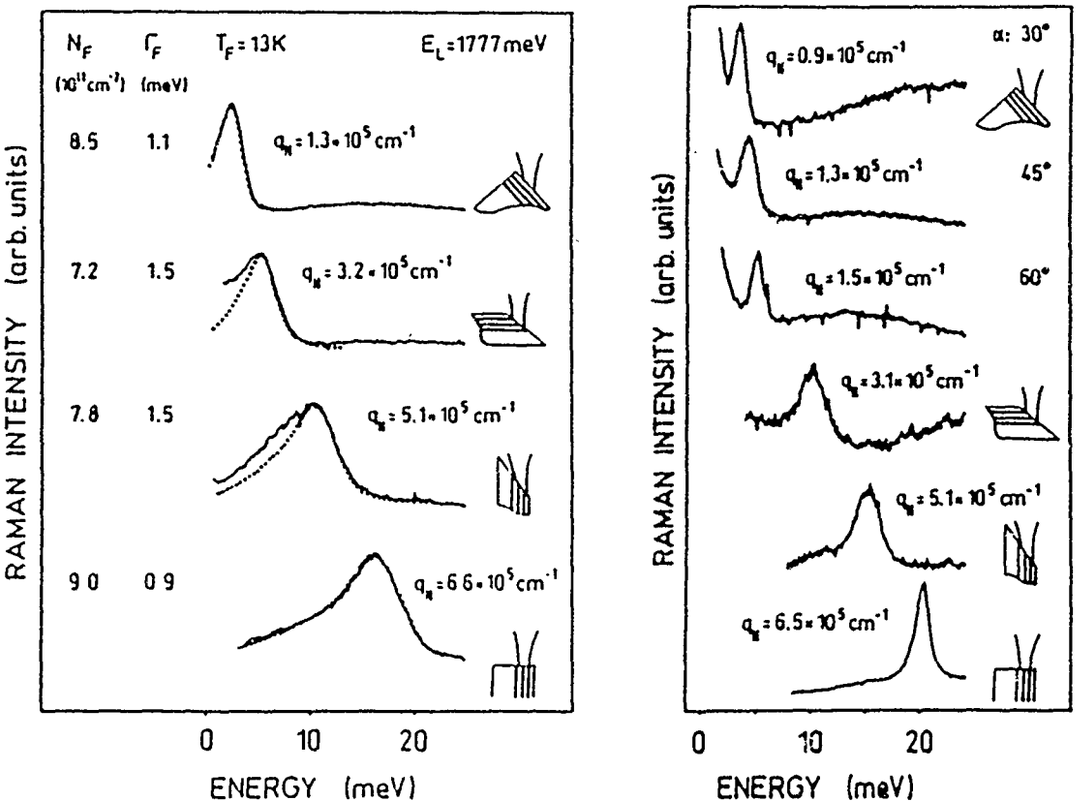


Fig. 2 Depolarized Raman spectra of layered electron systems. The dotted lines are theoretical fits to the observed single–particle excitations using the parameter given in the figure. The in–plane wave vector is determined by the scattering geometries which can be realized with micro–Raman spectroscopy as shown schematically.

Fig. 3 Plasmon excitations (polarized spectra) of layered electron systems measured with micro–Raman spectroscopy

of layers. The branches of the plasmon band which are probed by inelastic light scattering depend on the ratio $q_\parallel/q_\perp$. In backscattering from the as–grown surface $q_\perp$ is large so the lower (acoustic) plasmon modes are observed. From the cleavage plane $q_\perp = 0$, therefore the more 3D upper branch of the plasmon ($\beta = 0$) is probed. We see again from the spectra of Fig. 3 that the edges polished under an angle have less quality compared to the grown or cleaved surfaces. The plasmon dispersion, however, can be measured up to $q_\parallel \simeq 7 \times 10^5$ cm^{-1}.

Another way to extend the accessible wave vector range is the so–called grating coupler effect [13,14]. A periodic structure on top of the layered electron system can couple larger in–plane wave vector into the underlying semiconductors due to diffraction. The total wavevector is then given by $q_\parallel^m = q_\parallel + m \cdot g$ where $m = m_I + m_S$ is the diffraction order of incident and scattered light and $g = 2\pi/a$ is the reciprocal lattice constant of the periodic grating. Examples of collective in–plane excitations in GaAs quantum wells are shown in Fig. 4 [14]. A periodic etch mask of

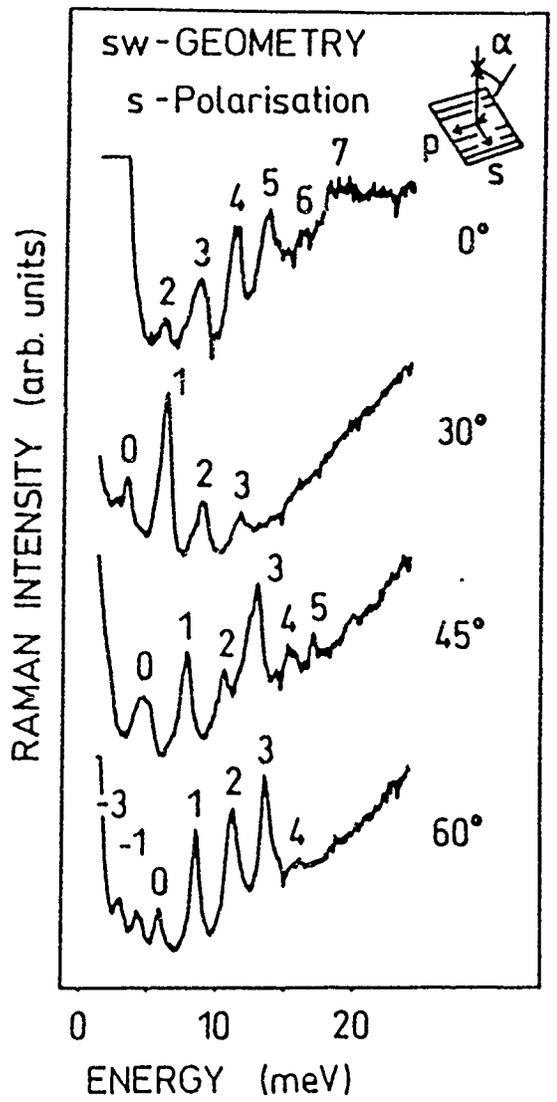


Fig. 4: Plasmon excitations measured with resonance Raman spectroscopy in layered quasi 2D electron systems. The different plasmon peaks correspond to different in–plane wave vectors $q_{||}$ which are accessible via diffraction of light due to the grating coupler which is etched into the sample. The numbers label the grating order. The 0 order peaks are similar to the plasmons observed in the upper three spectra of Fig. 3. The seventh order corresponds to an in–plane wave vector of $q_{||} = 5.5 \times 10^5$ cm^{-1} (from [14]).

photo resist stripes with a period of 8100 Å was produced by holographic lithography. The grating was etched into the sample in an optimized reactive ion etching process using $SiCl_4$ plasma [23]. Eighteen quantum wells remained underneath of this grating with a carrier concentration of about $5 \cdot 10^{11}$ cm^{-2}. Plasmons can be excited up to the seventh order now, even in backscattering with normal incidence on the surface. The peaks in Fig. 4 are labeled by the suitable diffraction order index m. For $\alpha = 0°$ the scattering wavevector is $q_{||}^m = mg$. With decreasing α we have to add $q_{||} = (4\pi/\lambda)\sin\alpha$ to obtain the total in–plane wavevector. With this grating coupler technique, the dispersion of layered plasmons could be measured up to about $q_{||} = 5.5 \times 10^5$ cm^{-1}[14].

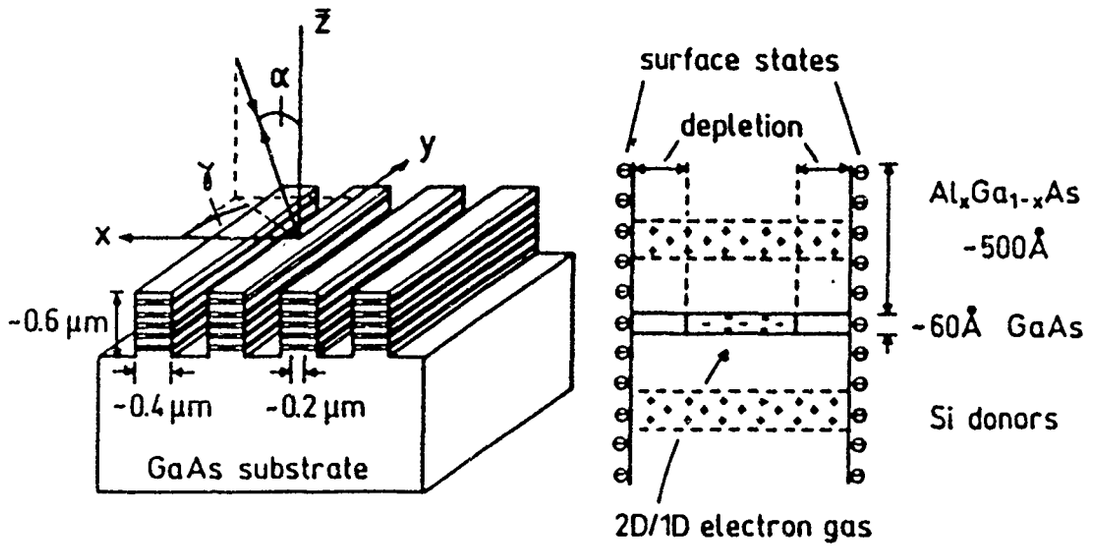


Fig. 5 Typical geometry of the wire structured samples achieved by holographic lithography and reactive ion etching. Each wire contains five modulation doped GaAs quantum wells. Details of a wire are shown schematically on the right–hand side of the figure.

4. Electronic excitations in wire structured quantum wells

In the last section we discuss very recent results on inelastic light scattering by free carriers in lateral quantum wire superlattices. The samples are prepared in a similar way as the etched grating coupler. The resulting periodic wires are shown schematically in Fig. 5. The geometrical width of the individual wires is about 4000 Å. Each wire contains five sheets of electrons in GaAs quantum wells with a thickness of 60 Å and a 2D carrier concentration of about $5 \cdot 10^{11}$ cm^{-2}. The actual lateral width of the electron systems is expected to be considerably smaller than the geometrical width of the wires due to surface depletion. This is shown schematically on the right–hand side of Fig. 5. The Fermi level pinning at the surface due to surface states and damage causes a depletion layer on each side, which leads to an approximately parabolic potential well toward the surface with a flat part in the middle [24]. A subband splitting due to this confinement in x–direction of about 1 to 2 meV is expected.

We have investigated both collective and single–particle excitations in such wire structured system by resonant inelastic light scattering. The dispersion of these excitations were studied in backscattering by tilting the laser parallel and perpendicular to the wires. In Fig. 6 polarized spectra are shown for three tilt angles α in each direction. Several plasmon excitations are observed in each spectrum. No shift of the peak position is observed with tilt angle α normal to the wires ($\gamma = 0°$). On the

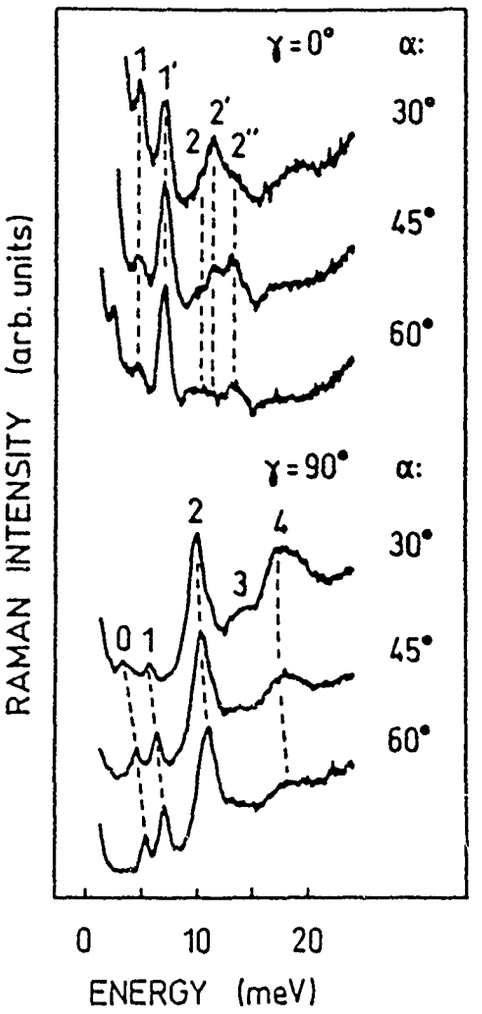


Fig. 6: Typical plasmon excitations in the wire structured sample for various tilt angles α normal to the wires ($\gamma = 0°$) and along the wires ($\gamma = 90°$). The peak positions remain unchanged for $\gamma = 0°$ (no dispersion). With $\gamma = 90°$ a considerable dispersion of the plasmon modes is observed, which is strongest for the lower order modes.

other hand the peaks shift to higher energies when α is increased in the direction along the wires. The dispersion is stronger for the low energy excitations. The vanishing dispersion perpendicular to the wires shows that Coulomb coupling between neighbouring wires is very weak and we observe resonances determined by the isolated wires. The spectrum of plasmon modes obtained parallel to the wires reflects qualitatively all the predictions made by Eliasson et al. [25] for a laterally bounded single layer 2D electron gas. To include the multilayer structure, we use the simplified model of confined plasmons. The interpretation of the plasmon resonances is based on the dispersion shown in Fig 7. The dotted lines are the five plasmon branches calculated for $n_{2D} = 5.1 \cdot 10^{11}\ cm^{-2}$. We keep this plasmon dispersion of the layered 2D electron gas and impose standing wave conditions for the plasmons in direction perpendicular to the wires. Then the in–plane component of the plasmon perpendicular to the wires is restricted to values $q_m = m \cdot \pi / w_e$, where w_e is the effective wire width which takes the partial lateral depletion into account. The component parallel to the wires remains unrestricted: $q_u = \sqrt{q_m^2 + q_u^y(\alpha)^2}$. These values are represented as vertical lines in Fig. 7. The peaks are assigned to the

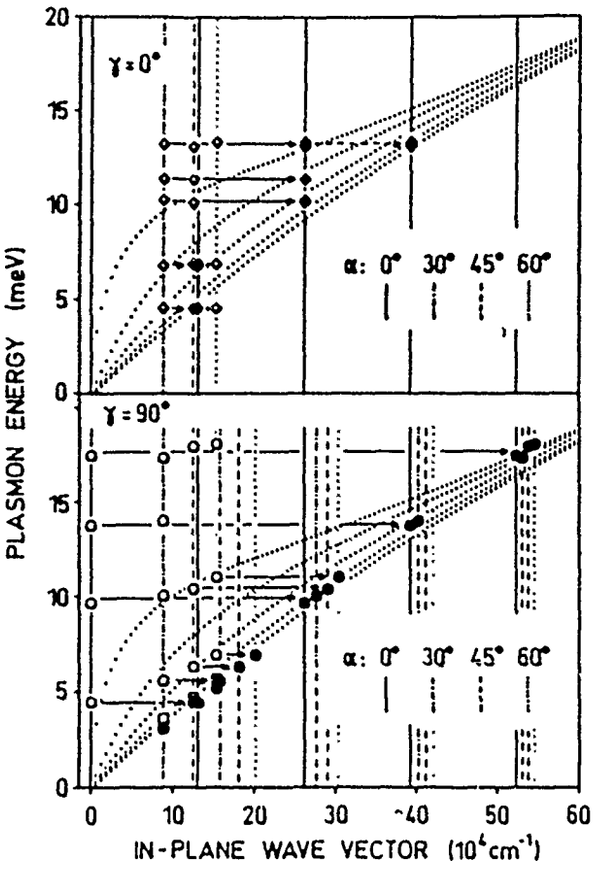


Fig. 7: Dipersion of plasmon excitations in the wire structured sample for $\gamma = 0°$ and $\gamma = 90°$. The allowed values for the in–plane momentum for the different angles α are represented by the vertical lines. The effective wire width determined from this analysis is $w_e \simeq 2400$ Å.

quantized momenta, for which their energy coincides with one of the plasmon dispersion branches. The peaks in Fig. 6 are labeled according to the quantization index m. Different branches of the plasmon dispersion are distinguished by primes. From the lowest confined plasmon, labeled by 0, we obtain a perpendicular momentum of 1.3×10^5 cm^{-1} which results in an effective wire width of $w_e \approx 2400$ Å. These experiments represent the first experimental study of plasmon dispersion and plasmon confinement in lateral quantum wire superlattices by resonant electronic light scattering. A detailed analysis of the results is presented in [16].

In conclusion, we have shown that resonant electronic light scattering is a versatile tool to study the dispersion of single particle and plasmon excitations in electron gases of different dimensionality in a wide wavevector range. Especially the possibilities opened by micro–Raman spectroscopy and grating couplers, and the new properties of quantum wires and quantum dots, which become now accessible, will lead to stimulating and exciting developments in the near future.

Acknowledgements. The work reported here was performed in close collaboration with G. Weimann, W. Schlapp, T. Demel and D. Heitmann. We are grateful for the experimental and technological helps and for many stimulating discussions. The project is supported financially by the Deutsche Forschungsgemeinschaft via Schwerpunkt "Physikalisch–technische Grundlagen von III–V Halbleiterstrukturen" and via the "Gottfried–Wilhelm–Leibniz Programm".

References

[1] A. Mooradian and G.B. Wright, Phys. Rev. Lett 16, 999 (1966)
[2] A. Mooradian, Phys. Rev. Lett. 20, 1102 (1968)
[3] A. Mooradian, in "Festkörperprobleme" IV, ed. by O. Madelung, (Pergamon Vieweg, Braunschweig 1969) p. 74
[4] M.V. Klein, in "Light Scattering in Solids" ed. by M. Cardona, Topics Appl. Phys., Vol. 8, 2nd ed. (Spinger, Berlin, Heidelberg 1983) p. 147
[5] G. Abstreiter, M. Cardona, and A. Pinczuk, in "Light Scattering in Solids" IV, ed. by M. Cardona, G. Güntherodt, Topics Appl. Phys., Vol. 54 (Springer, Berlin, Heidelberg 1984) p. 5
[6] A. Pinczuk, G. Abstreiter, R. Trommer, and M. Cardona, Solid State Commun. 30, 429 (1979)
[7] E. Burstein, A. Pinczuk, and S. Buchner, in "Physics of Semiconductors", ed. by B.L.H. Wilson (The Institute of Physics, London, 1979) p. 1231
[8] G. Abstreiter and K. Ploog, Phys. Rev. Lett. 42, 1308 (1979)
[9] A. Pinczuk, H.L. Störmer, R. Dingle, J.M. Worlock, W. Wiegmann, and A.C. Gossard, Solid State Commun. 32, 1001 (1979)
[10] G. Abstreiter, R. Merlin, and A. Pinczuk, IEEE J. Quantum Electron. QE 22, 1771 (1986)
[11] A. Pinczuk and G. Abstreiter, in "Light Scattering in Solids" V, ed. by M. Cardona and G. Güntherodt, Topics Appl. Phys. Vol. 66 (Springer, Berlin, Heidelberg 1989) p. 153
[12] T. Egeler, S. Beeck, G. Abstreiter, G. Weimann, and W. Schlapp, Superlattices and Microstructures, 5, 123 (1989)
[13] T. Zettler, C. Peters, J.P. Kotthaus and K. Ploog, Phys. Rev. B 39, 3931 (1989)
[14] T. Egeler, G. Abstreiter, G. Weimann, T. Demel, D. Heitmann, and W. Schlapp, Surface Science, in press
[15] J.S. Weiner, G. Danan, A. Pinczuk, J. Valladares, L.N. Pfeiffer, and K. West, Phys. Rev. Lett 63, 1641 (1989)
[16] T. Egeler, G. Abstreiter, G. Weimann, T. Demel, D. Heitmann, and W. Schlapp, submitted for publication
[17] A. Pinczuk, S. Schmitt–Rink, G. Danan, J.P. Valladares, L.N. Pfeiffer, K.W. West, Phys. Rev. Lett. 15, 1633 (1989)
[18] D. Olego, A. Pinczuk, A.C. Gossard, and W. Wiegmann, Phys. Rev. B 25, 7867 (1982)
[19] G. Fasol, N. Mestres, H.P. Hughes, A. Fischer, and K. Ploog, Phys. Rev. Lett. 56, 2517 (1986)
[20] A. Pinczuk, M.G. Lamont, A.C. Gossard, Phys. Rev. Lett. 56, 2092 (1985)
[21] G. Fasol, N. Mestres, M. Dobers, A. Fischer, and K. Ploog, Phys. Rev. B 36, 1536 (1987)
[22] J.K. Jain and P.B. Allen, Phys. Rev. Lett. 54, 2437 (1985)
[23] T. Demel, D. Heitmann, P. Grambow, and K. Ploog, Appl. Phys. Lett. 53, 2176 (1988)
[24] S.E. Laux, F. Stern, Appl. Phys. Lett. 49, 91 (1986)
[25] G. Eliasson, Ji–Wei Wu, P. Hawrylak, J.J. Quinn, Solid State Commun. 60, 41 (1986)

Nanofabrication of Quantum Wires and Rings

Steven P Beaumont
Nanoelectronics Research Centre
Department of Electronics and Electrical Engineering
University of Glasgow
Glasgow G12 8QQ
Scotland UK

Outline:

Dimensional requirements

Lithography for nanostructure definition

Pattern transfer techniques

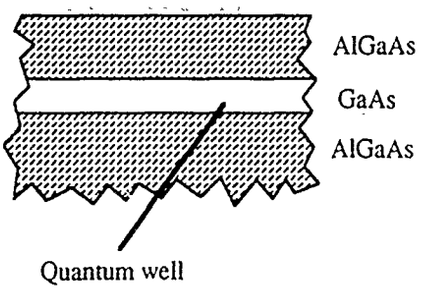
(Direct fabrication - wait for next lecture)

Issues in Fabrication of Quantum Confinement Devices

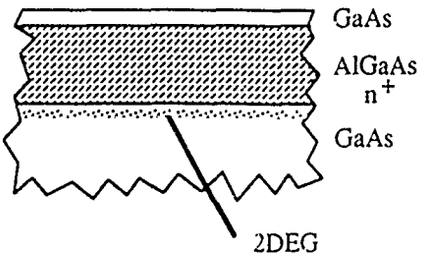
- Dimensional requirements
- Preservation or improvement of structural resolution during pattern definition and transfer
- 'Cleanliness'
  - start with structure with high degree of perfection, usually grown by MBE. Critical, active regions are buried below surfaces, interfaces are of high (~monolayer smooth) quality
  - we do something nasty to this beautifully perfect structure using electrons, ions, dopants, etchants, contacts, chemicals.....
  - and then hope that we have a perfect system in which to demonstrate some new physics or make smart devices

Fabrication of Structures for Quantum Confinement

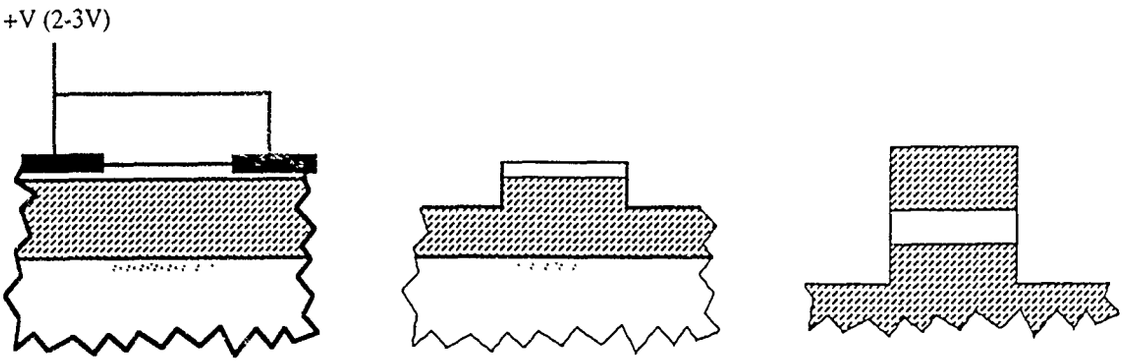
Starting materials:



1. Quantum well layer



2. Modulation doped layer



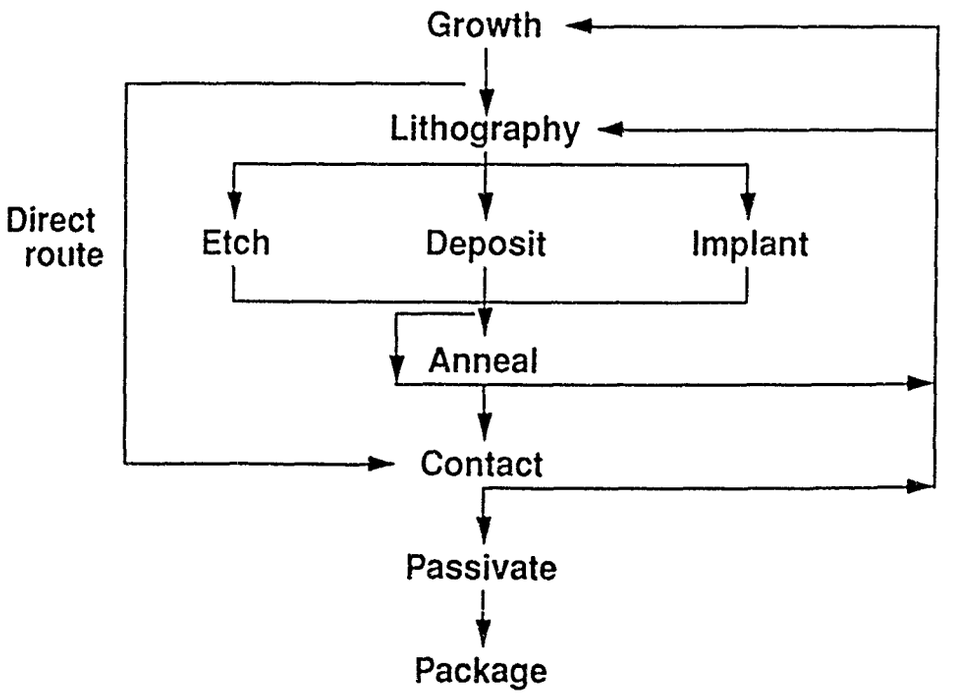
a. Gating

b. Shallow etching

c. Deep etching

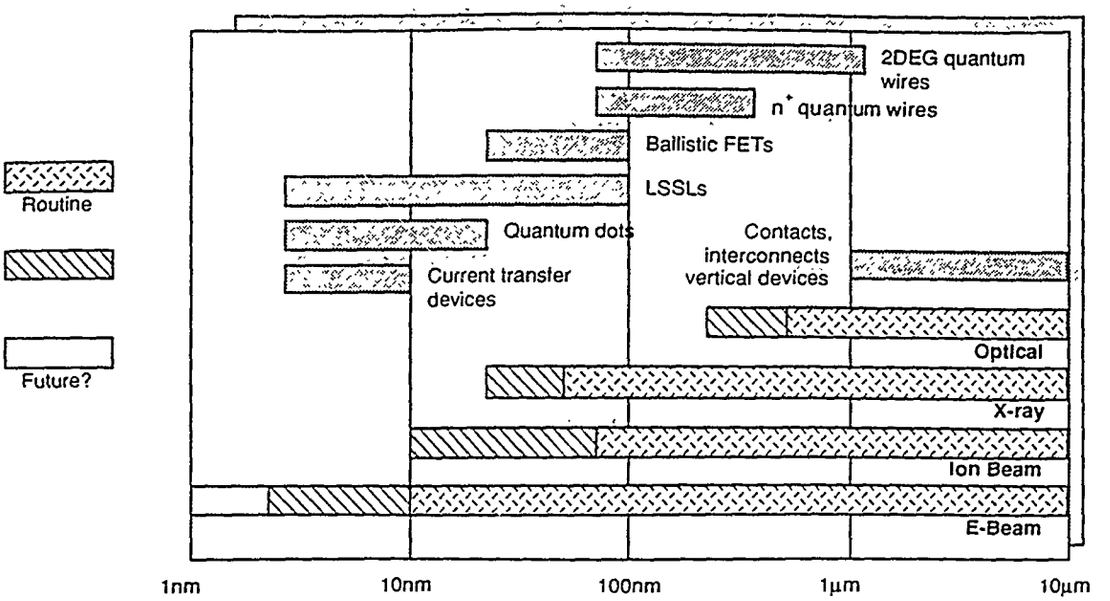
The Device Process Cycle

• Structure fabrication usually requires a multi-step approach



• All steps in the process require to preserve or enhance the resolution of their predecessors

Lithography for <100nm dimensions



•Suitable techniques are

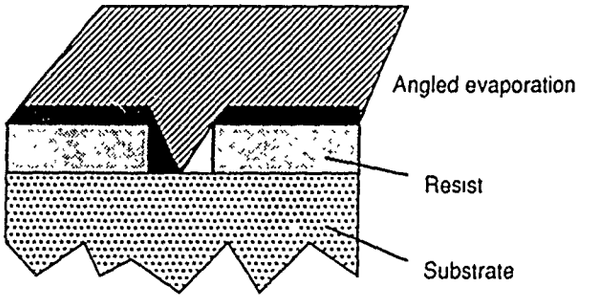
* x-ray lithography: but this is a contact printing process relying on other methods to make masks
* Ion beam lithography
* Electron bea lithography

Photolithography

Photolithography

- Limited to ~ 0.4 - 0.5µm (arguably)
- But specialist techniques or tricks can achieve smaller dimensions e.g.

- Shadowing



- Holography

Overlapping laser beams generates interference fringes with period

$$\Lambda = \frac{\lambda_0}{2 \sin \theta'}$$

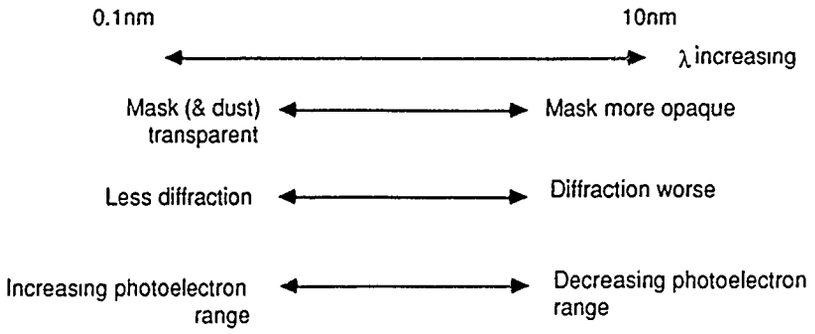
X-Ray Lithography I

If diffraction effects are a limitation, move to shorter wavelengths e.g. X-rays

Problem: no lenses. Must use contact or proximity printing process

- need some other non-optical method of mask making e.g. E-Beam

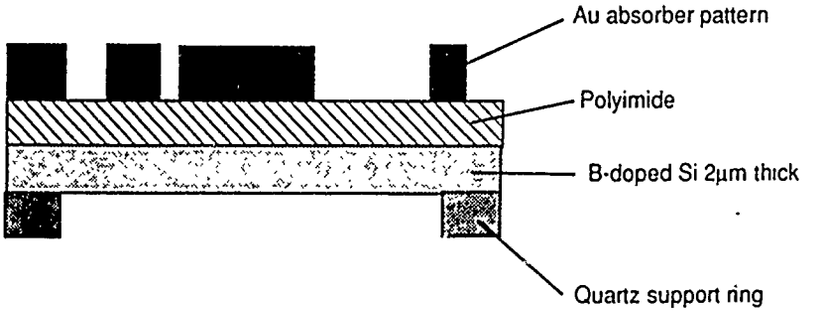
Choice of wavelength: Depends on degree of contact and mask transparency



X-ray Lithography II

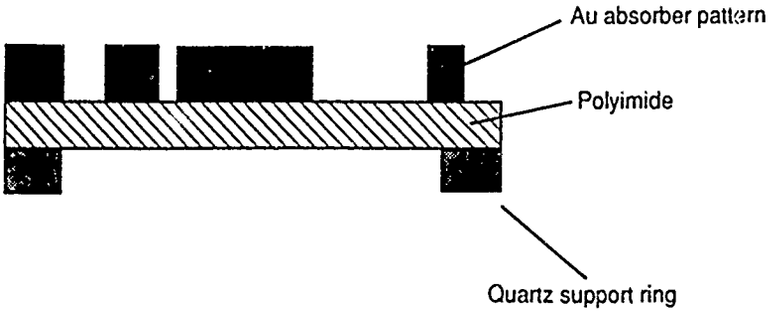
Mask Technology

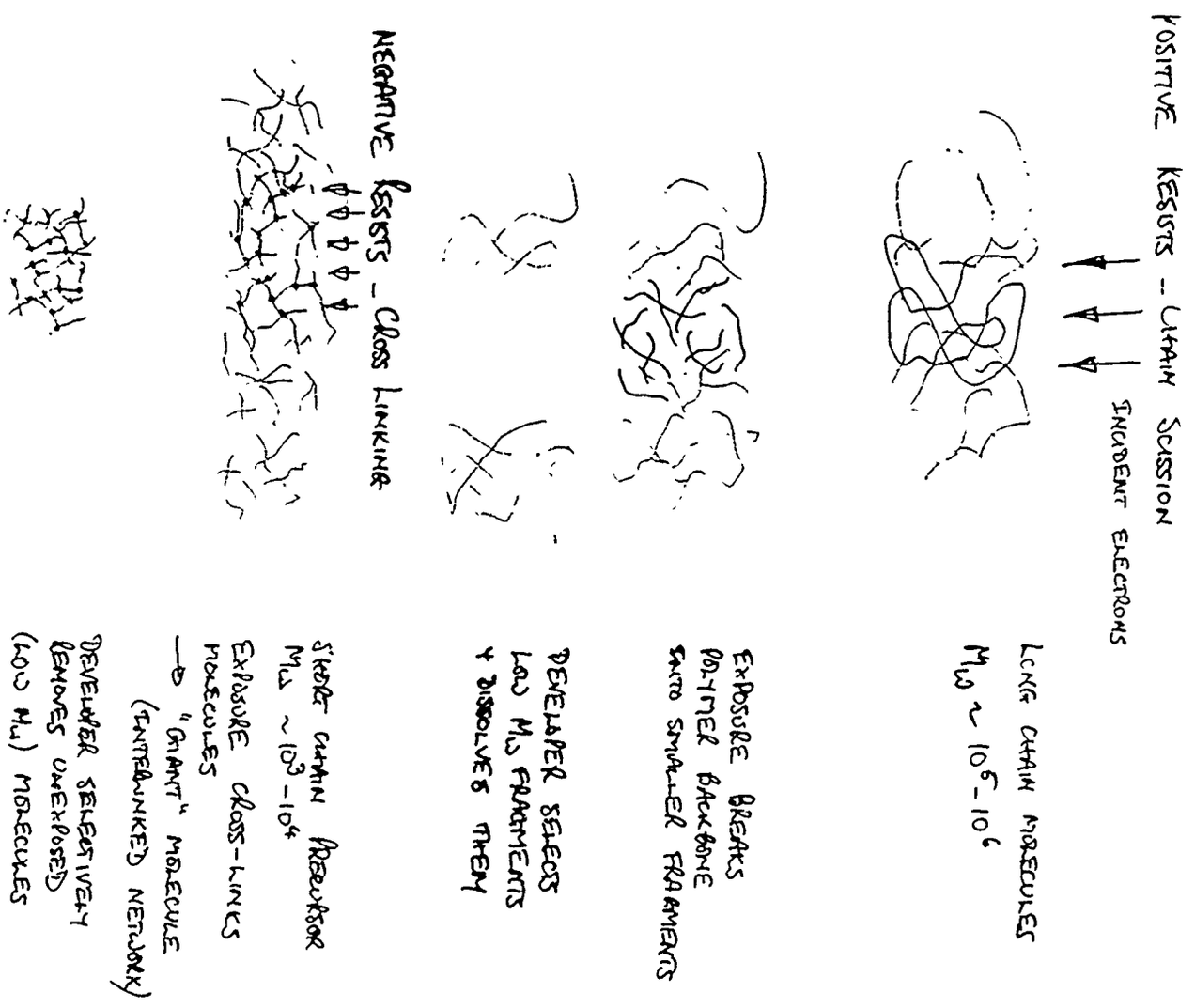
Si + Polyimide + Gold Absorber



Polyimide + gold (conformable mask for contact)

- Transparent to long wavelength radiation
- Flexible, electrostatic chucking





Inorganic Resist Techniques for <10nm Lithography

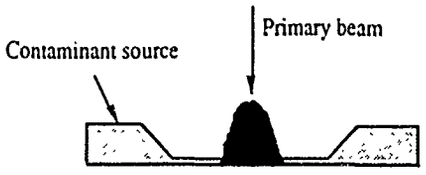
Attempts at higher resolution resist processes

Contamination

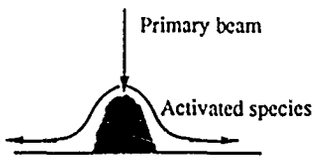
- Decomposition of vacuum oil or other carbonaceous material
- Carbon deposit can be used as an etch mask
- 8nm resolution achieved

Drawbacks

- Surface migration of undecomposed contaminant causes depletion effect limiting packing density
- Known that activated species migrate after interaction with beam: linewidth resolution limitation



Depletion of contaminant source



Migration of activated species before 'fixing'

NB Watch out for these effects in proposals to carry out beam-controlled growth or deposition. Presumably can control with substrate temperature

Inorganic Resist Techniques for <10nm Lithography II

Another promising line of attach involves "hole-burning" in inorganic resists.

Thin films are prepared by evaporation and exposed to very high doses of electrons

Holes are "burned" in the films

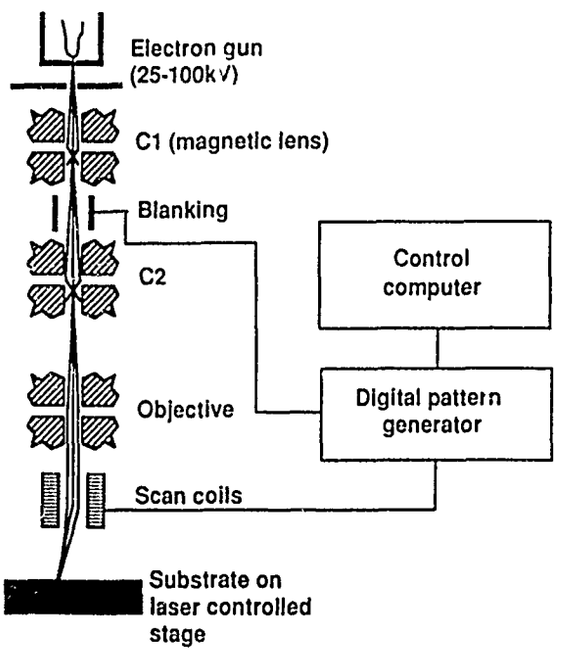
- Examples of materials include magnesium oxide, aluminium/strontium/cadmium fluoride (last two are epitaxial on GaAs)
- Resolution <~2nm has been achieved
- Require very heavy dose (10's of C/cm^2) for exposure (cf PMMA ~10^-4 - 10^-3 C/cm^2)
- DOSE RATE DEPENDENT (unlike ordinary resists): will not expose at low current

Problem: material does not clear completely in smallest structures

Practical resolution (after transfer) no better than best polymers

Interesting nonetheless: basis for all-vacuum process

Electron Beam Nanolithography



Resolution determined by:
- Probe diameter (can be 0.5nm)
- Electron scattering (minimise with high energy beam)
- Secondary electron generation (3nm?)
- Long-range chemistry, molecule size effects

Advantages:
- Direct computer control of pattern
- High resolution

Drawbacks:
- Low throughput
- Limited field size

Applications:
- Mask making
- Direct write <0.5μm

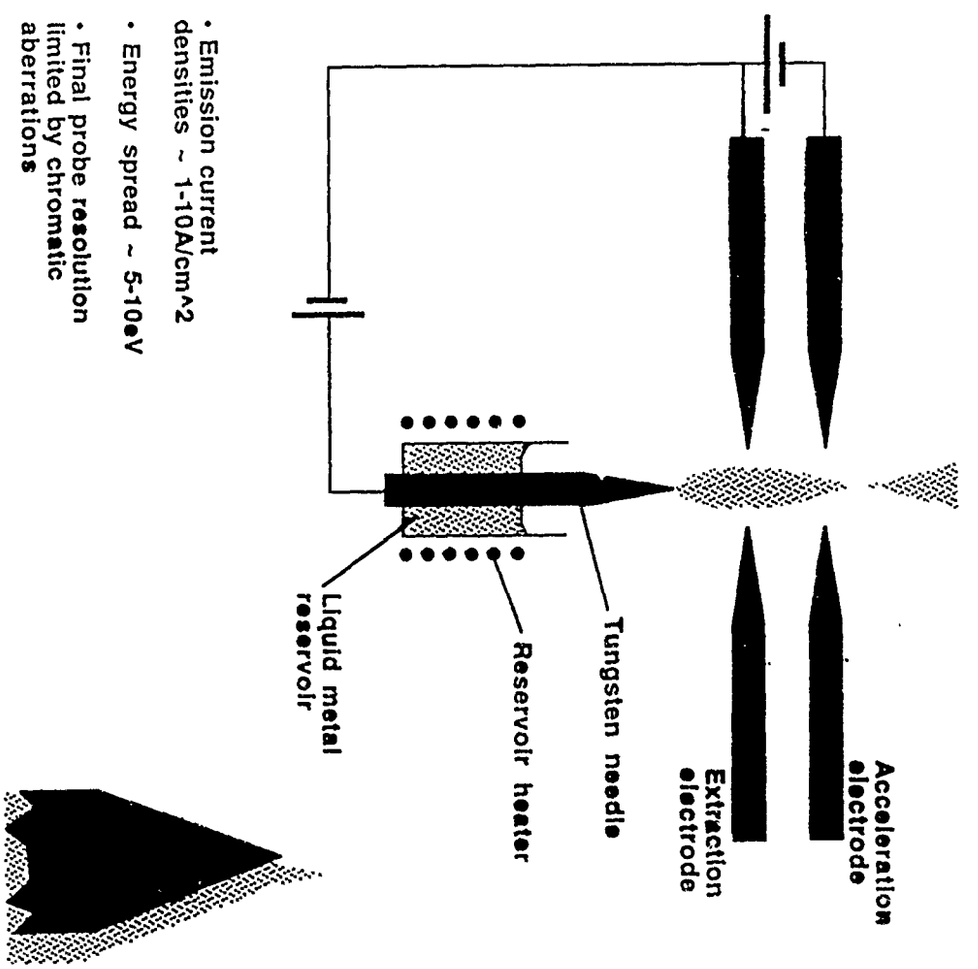
Focussed Ion Beams

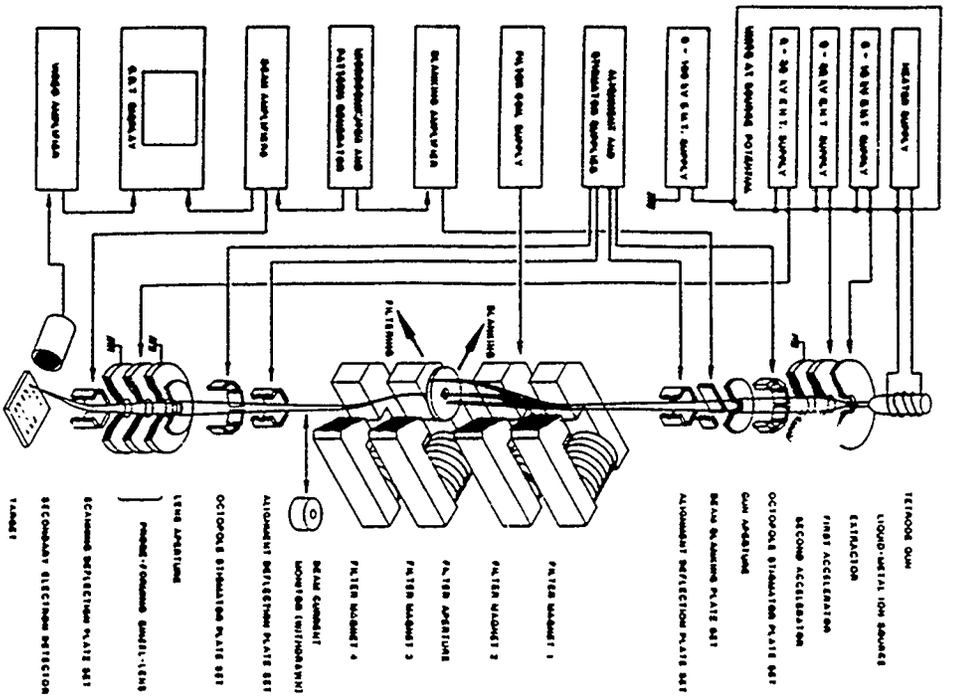
Why Focussed Ion Beams?

- Much less primary elastic scattering of ions than electrons especially in resist (less deviation/lateral scattering)
- Can be arrested by resist - no penetration or damage to the substrate
- A focussed ion beam offers other fabrication opportunities:
  - Direct Implantation
  - Direct Etching/Sputtering
  - Localised damage for quantum wire isolation/definition

But:

- The technology is immature: viable, bright sources have only recently emerged
- The resolution is limited by chromatic aberrations (energy spread on the beam) to 10-30nm



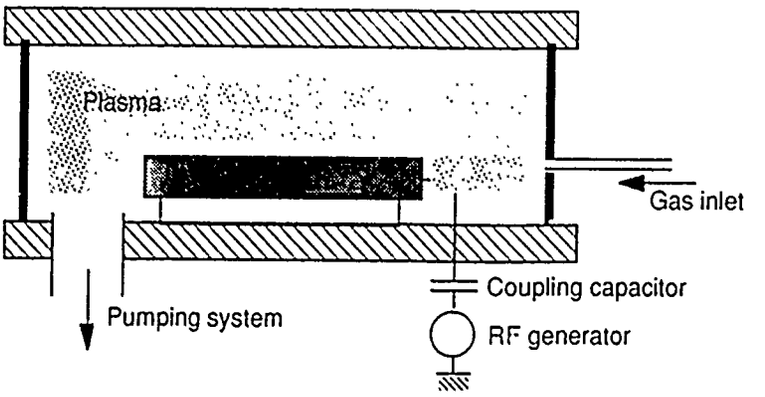


Ion implantation: schematic of 100 kV, 40 nm focused ion beam system

(Cleaver, Heard and Ahmed, 1983, *Microcircuit Engineering 83* (Academic Press) 135)

J. R. A. Cleaver

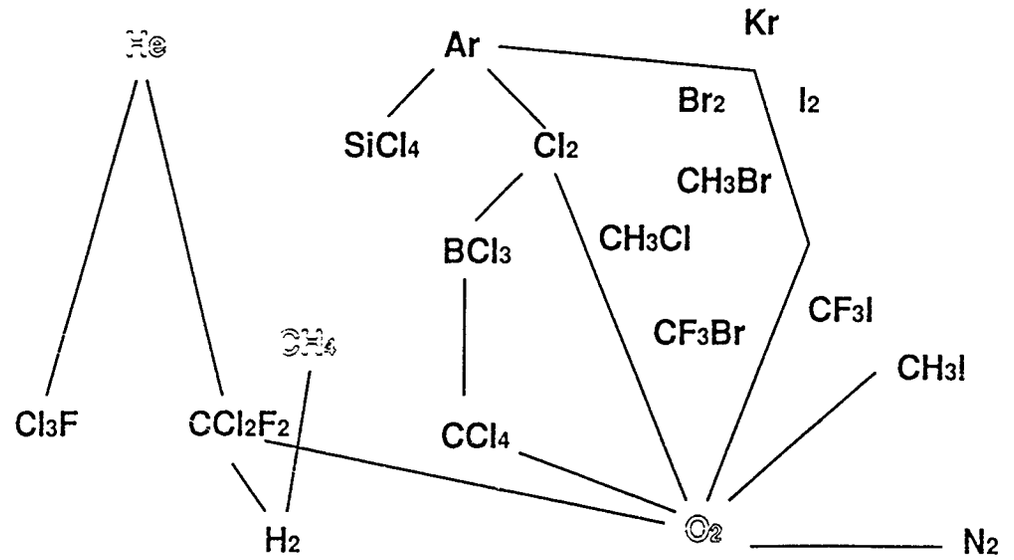
Reactive Ion Etching



Schematic RIE System

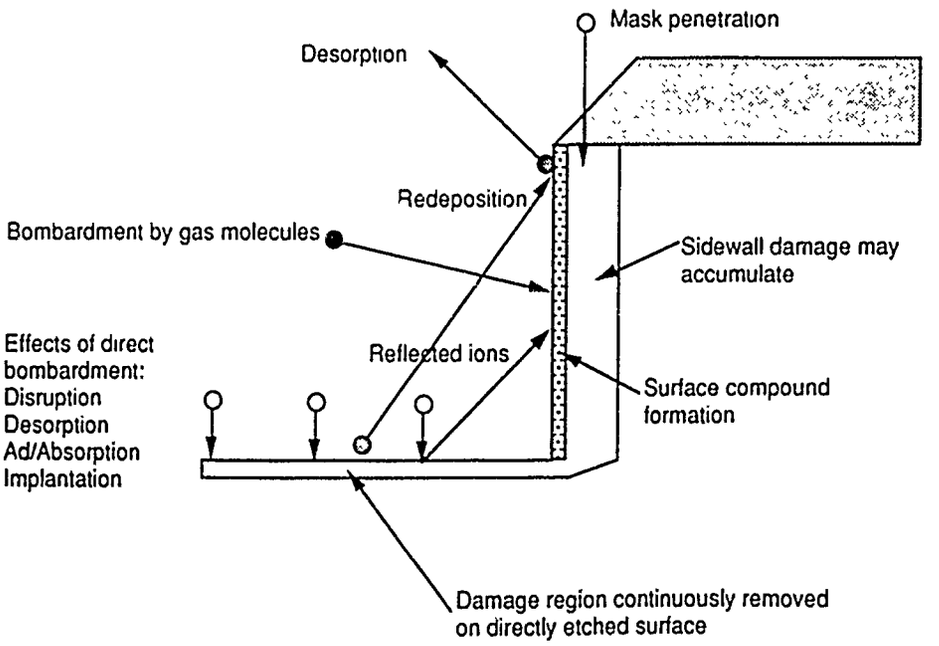
- Sample bombarded by ions from rf excited plasma accelerated across dark space by dc bias (20-1000V)
- Gas may be inert (sputtering) or chemically active
- Chemical reactions enhanced by ion bombardment
- Products of chemical reactions should be volatile at substrate temperature
- Resolution = resolution of mask: but damage to material results

Gases reported for III-V Etching

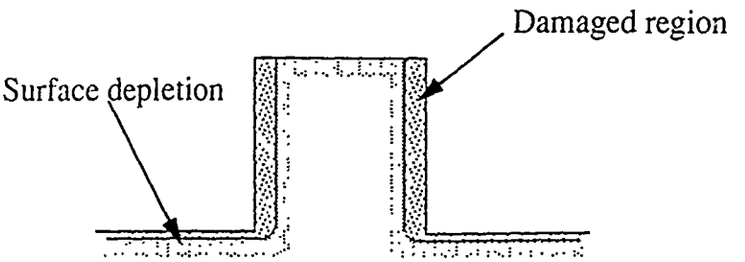


Gas only as mixture
Gas alone

Sources of Dry Etch Damage



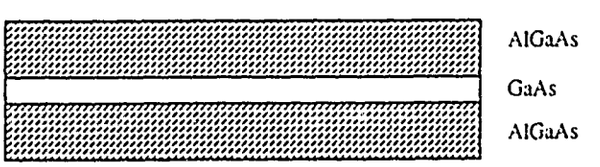
Electrical Study of Sidewall Damage



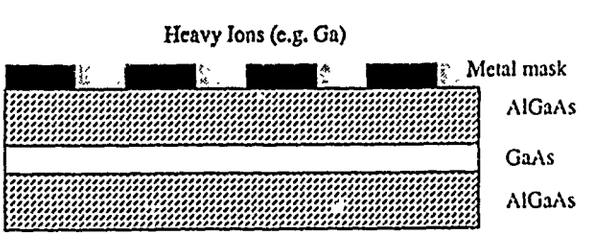
- Measure conductance versus width
- Perfect wire cuts off at 2 x surface depletion depth
- Damaged wire has larger cutoff
- Characterisation of the damage depth allows better modeling/design of quantum wire structures

Impurity Induced Disorder

- Diffuse/Implant impurity into quantum well structure
- Wells randomise/intermix/interdiffuse
- Bandgap increased, optical absorption decresed
- Can use to produce:
  - low loss optical waveguides interconnecting MQW lasers, modulators and detectors
  - laser mirrors
  - lateral confinement
  - quantum wires and dots
- Avoid free carrier absorption by use of neutral impurities, laser-induced disorder



1. Original Quantum Well Structure

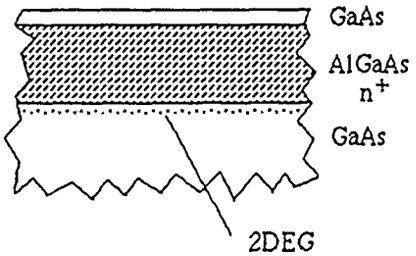


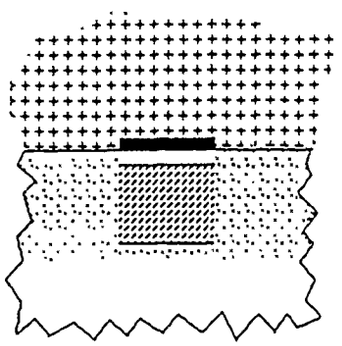
2. Implant ions (Ga, Si, B, F) and anneal

3. Al intermixed into GaAs well

Damage Confinement Technique

Method due to Scherer and Roukes, developed by Thornton

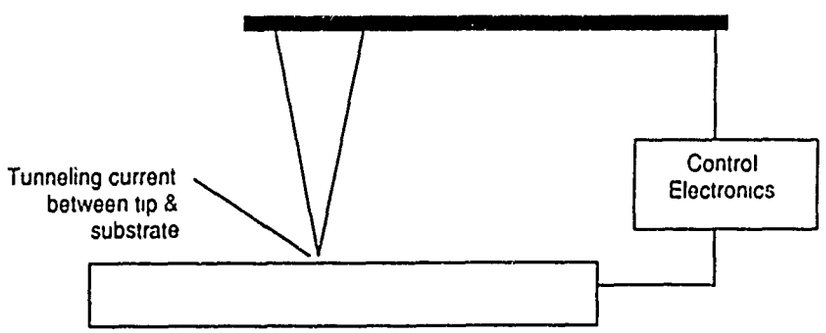




- 2DEG starting material
- Deposit mask (resist, lifted off inorganic, e.g. Strontium Fluoride
- Bombard with low energy ions of noble gas (Ar, Ne)
- In-situ monitoring of conductivity
- Confinement by mobility reduction and loss of carriers to traps
- If carefully controlled, lithographic width = electrical width

---

Fabrication using the Scanning Tunneling Microscope



Tunneling current can be used to:

- **Dissociate (expose) resist**
- **Deposit contamination**
- **Dissociate metallorganics**
- **Directly modify surface**

Possibility of atomic scale fabrication??

Conclusions

- Quantum confinement structures: straddle a size regime where fabrication is difficult
- Cleanliness, or structural perfection, is a major issue
  - In doped structures, depletion from surfaces and interfaces may help to confine electrons from imperfect regions. Transport devices can be very 'clean' (eg. gated point contacts)
  - Depletion can also be utilised to obtain an extra degree of confinement: the drawback being the difficulty of exercising fine control of the shape and size of the active region due to fringing effects
  - In optical quantum structures in which carriers are photoexcited, there is greater sensitivity to processing effects
- 'Traditional' processing still works quite well for transport experiments: more ingenious techniques may be needed for optical devices

Optical Properties of Arrays of MicroDevices (Quantum Wires and Dots)

Steven P Beaumont
Nanoelectronics Research Centre
Department of Electronics and Electrical Engineering
University of Glasgow
Glasgow G12 8QQ
Scotland UK

Outline:

Dimensional requirements

Studies of etched and overgrown quantum wires & dots: the role of surfaces

Evidence for quantum confinement

Clean fabrication techniques

Ingenious experiments

Conclusions

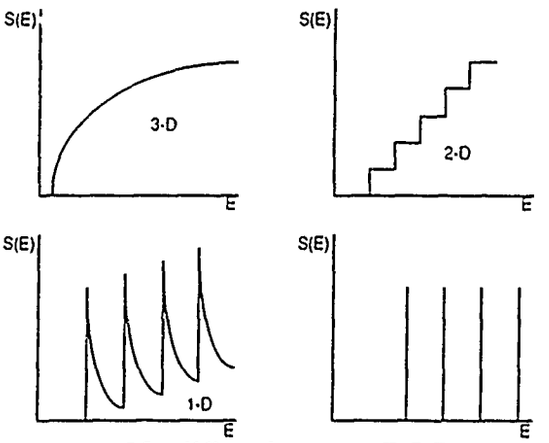
Introductory

Increasing confinement from 2- to 1- to 0-D

Typically by starting from 2-D structure such as QW

Pattern laterally to create confining potentials:

- surfaces (etch)
- modify alloy composition (IID)
- apply external potentials



Modifications to DOS

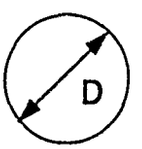
Optical transitions restricted to few discrete energies

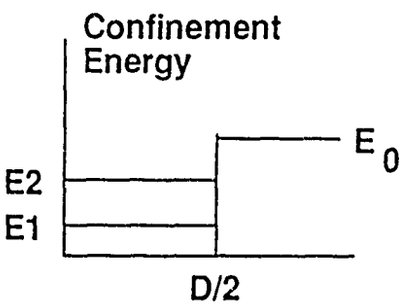
- enhanced oscillator strength
- narrow transitions

Size constraints

(After Vahala, J. Quant. Electr. 24,523,1988)

Lower limit





If D too small, no confined states in dot

$$D > D_{crit} \approx \sqrt{\frac{10\hbar^2}{m_e E_0}}$$

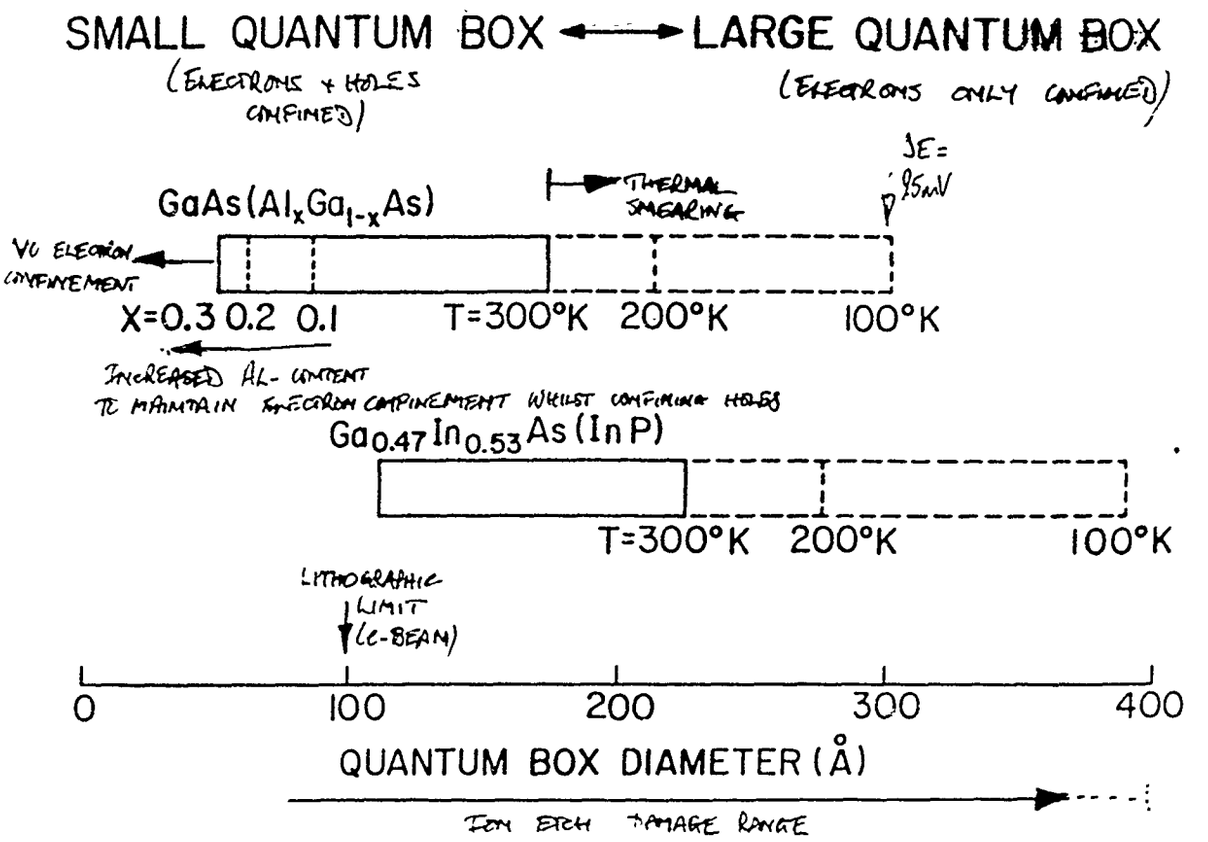
Upper limit

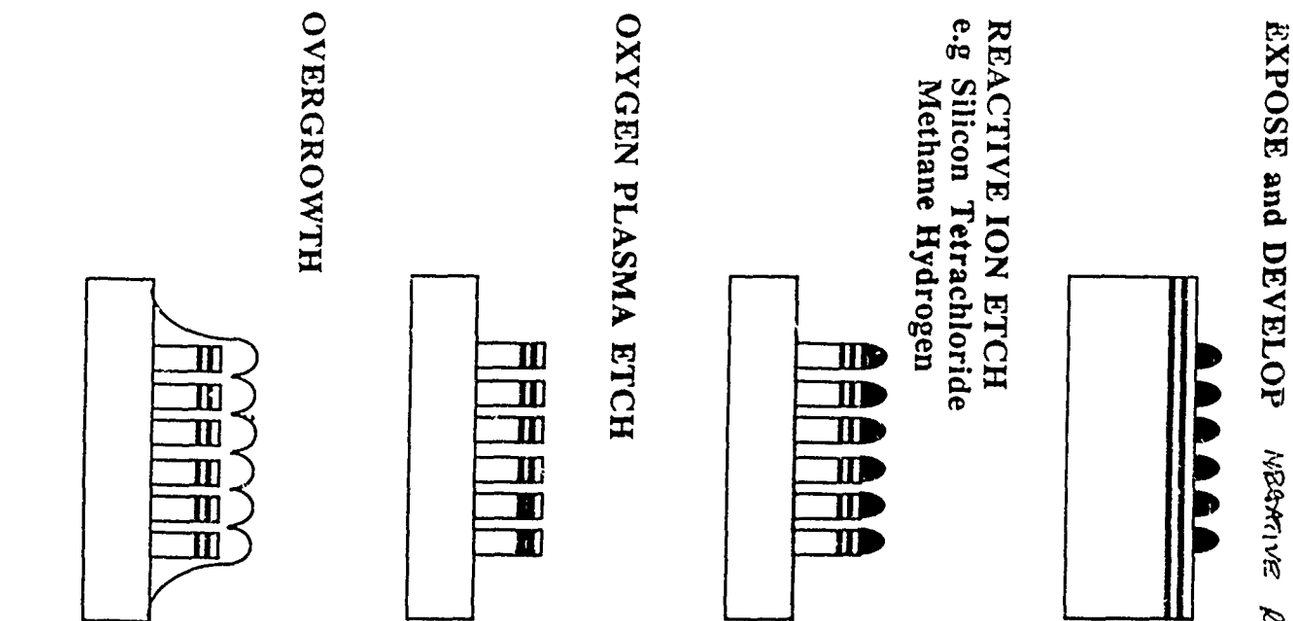
If D too large, transitions smeared by thermal excitation

$$D < D_{max} \approx \sqrt{\frac{10\hbar^2}{6 m_e kT}}$$

Also detectability and fabrication tolerance limitations

AFTER VAHALA, J. QUANT. ELECTR. 24, 523, 1988





(Typical)

Material Structure

MBE

GaAs Cap	1000Å of 30% AlGaAs
14Å QW	200Å of 30% AlGaAs
28Å QW	200Å of 30% AlGaAs
57Å QW	200Å of 30% AlGaAs
115Å QW	200Å of 30% AlGaAs
	GaAs Buffer Layer
	n+ GaAs substrate

Depth Standard (Å)

MOCVD

GaAs Cap	500Å of 30% AlGaAs
100Å QW	500Å of 30% AlGaAs
	GaAs Buffer Layer
	semi-insulating GaAs substrate

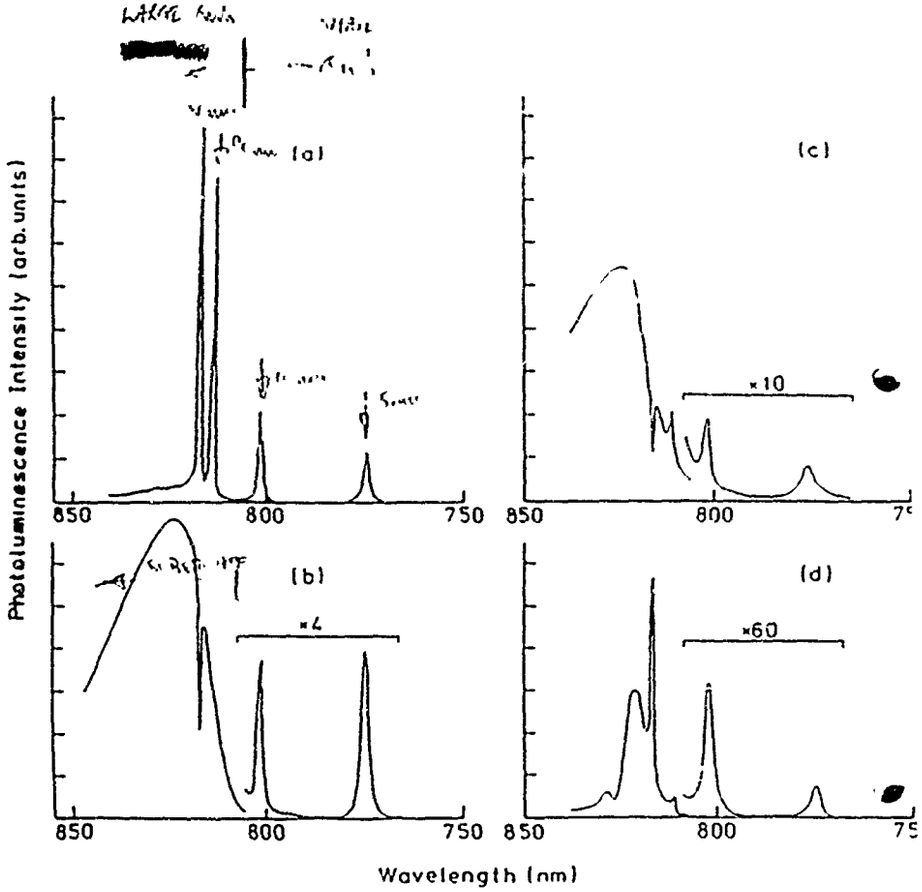


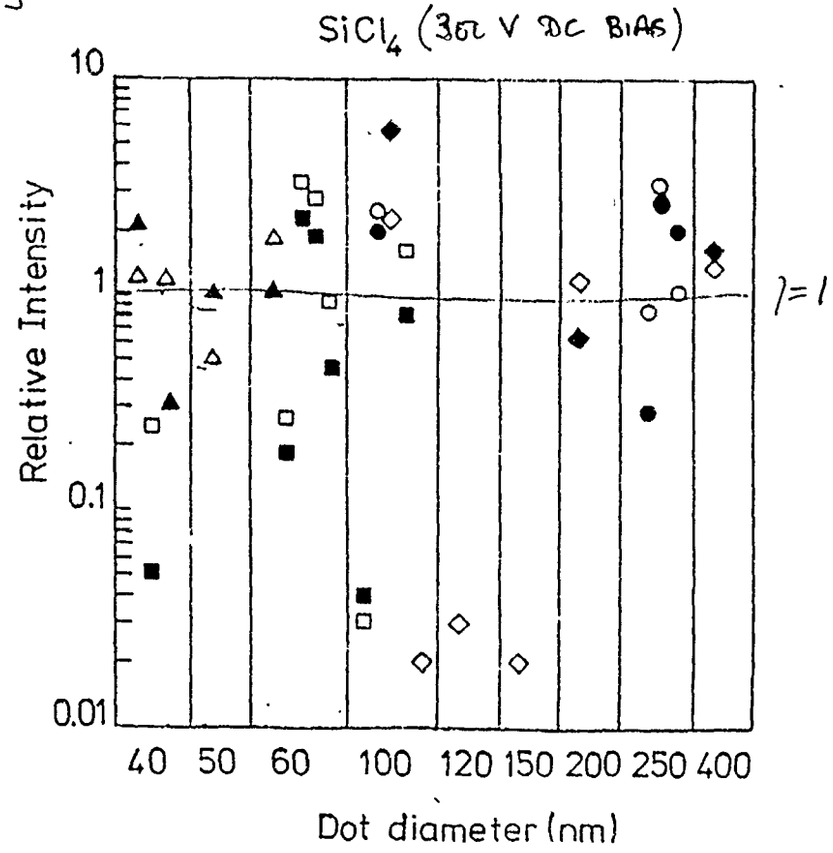
Figure 4.2 Representative PL spectra obtained at 5K with excitation at 1.65eV and 1Wcm^{-2} for GaAs/AlGaAs quantum well sample before patterning (a) and after patterning into (b) 60nm pillars (250nm pitch) etched to depth of 0.2μm using $SiCl_4$ RIE (c) 60nm pillars (200nm pitch) etched to depth of 0.12μm using Ar ion milling, (d) 60nm pillars (180nm pitch) etched to depth of 0.1μm using CH_4/H_2 RIE

LUMINESCENCE OF QD's
NORMALISED TO CONTROL MESA

OPEN SYMBOLS – 5nm QW
FILLED " – 10nm QW

△ □ } HRN MASK

○ ◇ } METAL MASK

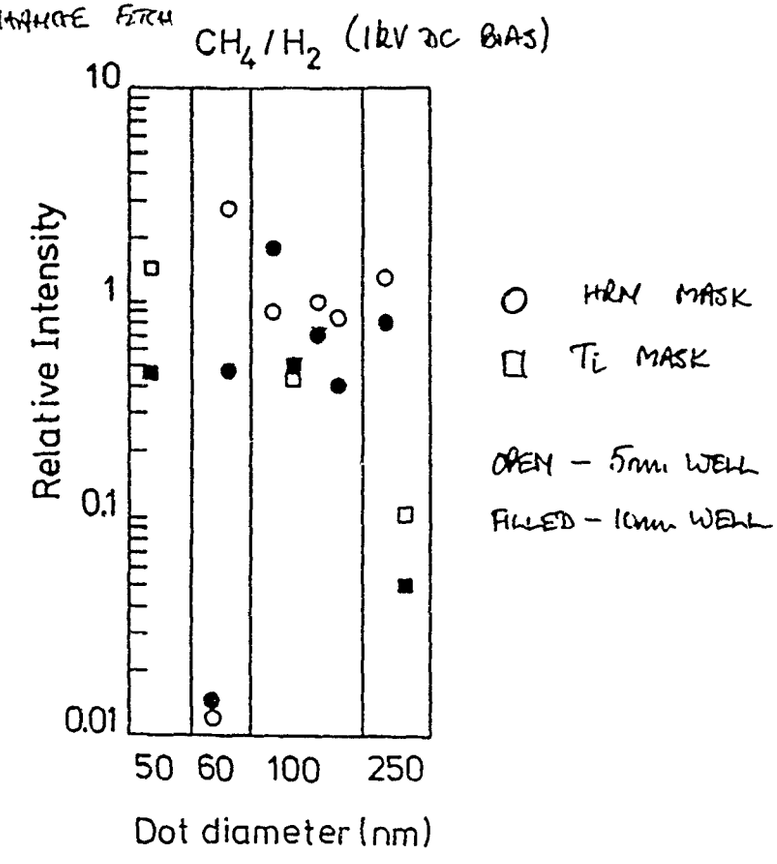


WHERE IS NON-RADIATIVE SURFACE RECOMBINATION?

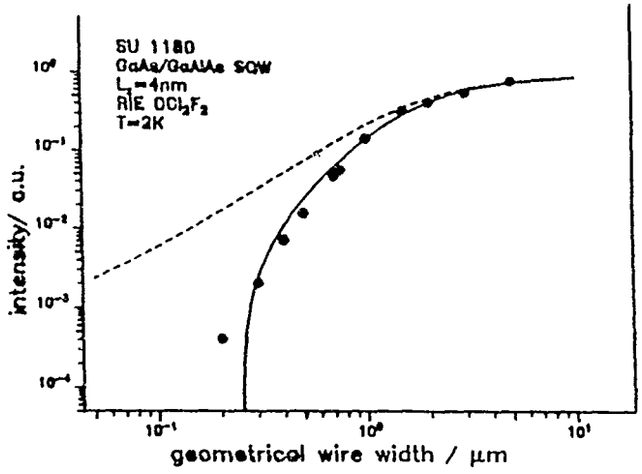
IS IT SUPPRESSED/PASSIVATED BY PROCESS?

IS IT PASSIVATED BY CHLORIDE LAYER?

CHANGE ETCH



LUMINESCENCE INDEPENDENT OF PROCESS/GAS



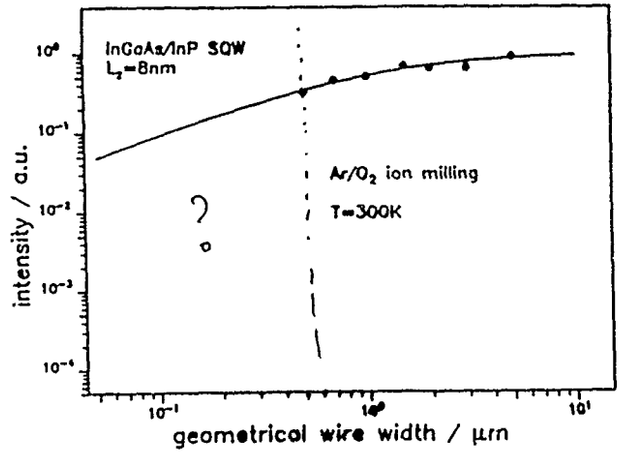
BUT IN SOME CASES IT MAY BE:

FORCHEL'S EARLY RESULTS

NOTE STRONG CUTOFF ABOVE 200 nm LINEWIDTH

POSTULATES "DEAD" LAYER DUE TO REE DAMAGE

Fig. 8 Photoluminescence intensity of etched GaAs/GaAlAs wires versus the wire width. Hexagons: experimental data; broken line: calculated width dependence with a surface recombination velocity of $5 \cdot 10^5$ cm/s and no depleted surface layers; solid line: calculation including a depleted surface layer of 130 nm depth.



THINGS ARE BETTER IN InGaAs/InP

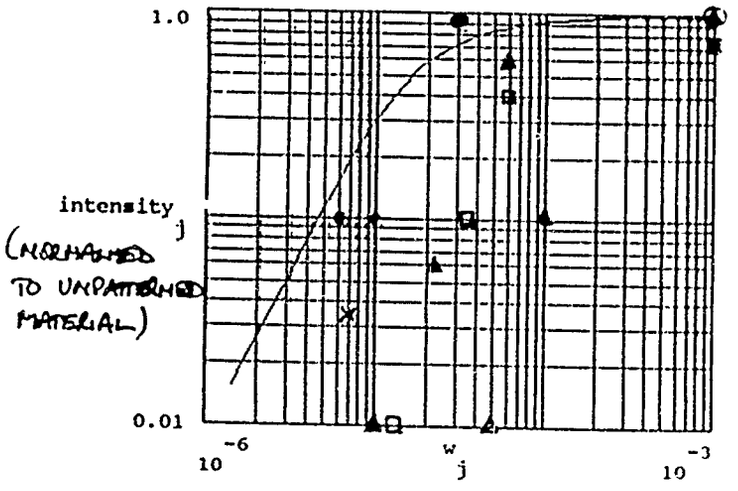
- THIS ION MILLING IS HIGH-DAMAGE PROCESS

Fig. 9 Lateral width dependence of the excitonic emission from etched InGaAs/InP wires. Hexogons: experimental data; solid line: calculation using a surface recombination velocity of $5 \cdot 10^4$ cm/s.

FORCHEL et al. FESTKÖRPERPROBLEME 28, 99, 1988

COMPARISON OF LUMINESCENCE EFFICIENCY / WIDTH (DIMENSION) RESULTS

LUMINESCENCE η OF DOTS AND WIRES COMPARED FOR FIRST TIME.



Forchel data:
D = 0.4
τ = 200ps
S = 10^6cm/s

					TEMP.
Δ	FORCHEL	WIRES	(1988)	ETCH $SiCl_4$	2K.
□	CLAUSEN	DOTS	(1989)	ETCH BCl_3	20K
X	HEITMANN	WIRES	(1989)	ETCH $SiCl_4$	4K
•	ARNOT	DOTS	(1990)	ETCH $SiCl_4$ }	SAME SAMPLE, SAME EXPERIMENT, 4K
▲	ARNOT	WIRES	(1990)	ETCH $SiCl_4$ }	

ALL POINTS ARE FOR :-

GaAs/AlGaAs QW MATERIAL

E-BEAM LITHOGRAPHY

REACTIVE ION ETCHING

1. TECHNOLOGY (of FORCHEL, HEITMAN, ARNOT WIRES)
2. DRAMATIC DIFFERENCES BETWEEN WIRES + DOTS (SAME SAMPLE, PROCESS) OF EQUAL DIMENSION.

NOMINALLY IDENTICAL

MATERIALS (GaAs/AlGaAs)

STRUCTURES (WIRES)

PROCESSING ($SiCl_4$ ETCHING EBLITHOGRAPHY)

→ ORDERS OF MAGNITUDE

DIFFERENCES IN RESULTS FOR LUMINESCENCE EFFICIENCY.

INTER-LAB PROCESS DIFFERENCES ARE IMPORTANT

(CLEANLINESS, PREPARATION TECHNIQUES ETC)

(Conclusions) - Free Standing Etched Dots

- Majority of data points scattered about relative intensity of 1
- Luminescence scales approximately with volume of quantum well material excited
- No blue shift - dots not small enough
- INDEPENDENT OF PROCESSING CONDITIONS (MASK, ETCH ROUTE)

Issues

- What is role of surface recombination?
  - Unimportant here?
  - Dots etched with Ar show low luminescence efficiency
  - Chemical passivation effect?
  - Linear incident power dependence supports low surface recombination rate OR DIMINISHED ROLE FOR SURFACE RECOMBINATION
- → Can carriers diffuse to surface? ←
- Variability
  - Mask variations - cathodoluminescence
- APPARENTLY NO EFFECT OF SIDEWALL DAMAGE

CAN EXCITONS DIFFUSE TO SURFACE AT LOW TEMP? (4K)?

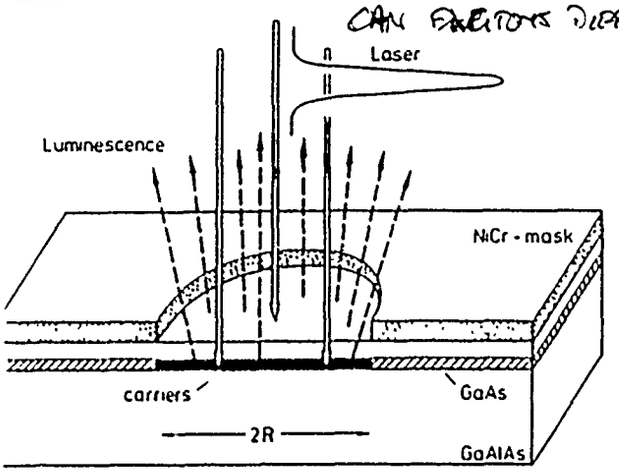


FIG. 1 Schematic design of the mask sample combination used for the lateral transport studies.

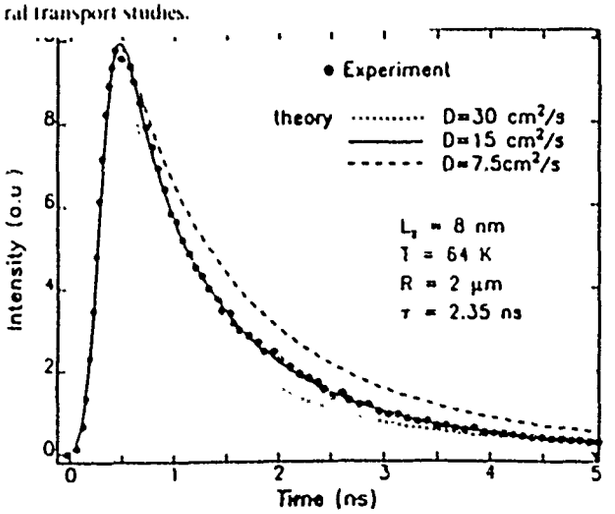


FIG. 4 Experimental transport profile (dots) together with three calculat-

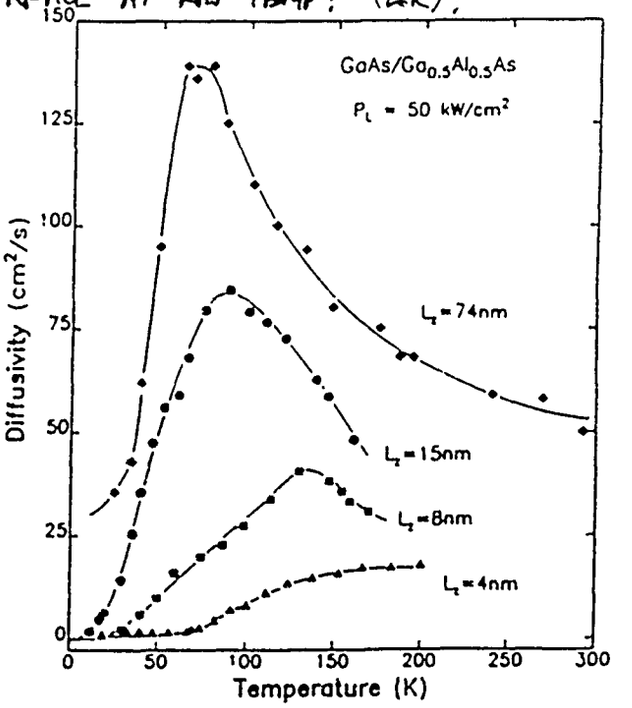


FIG. 5. Diffusivity as a function of temperature for three different QW widths and a 3D GaAs reference layer.

EXCITON TRANSPORT AT LOW TEMPERATURE: ARE EXCITONS LOCALIZED IN DOTS & PREVENTED FROM MIGRATING TO SURFACE STATES ?

(HILLMER et al. APL 53, 1937, 1988

TEMPERATURE DEPENDENCE OF QD LUMINESCENCE

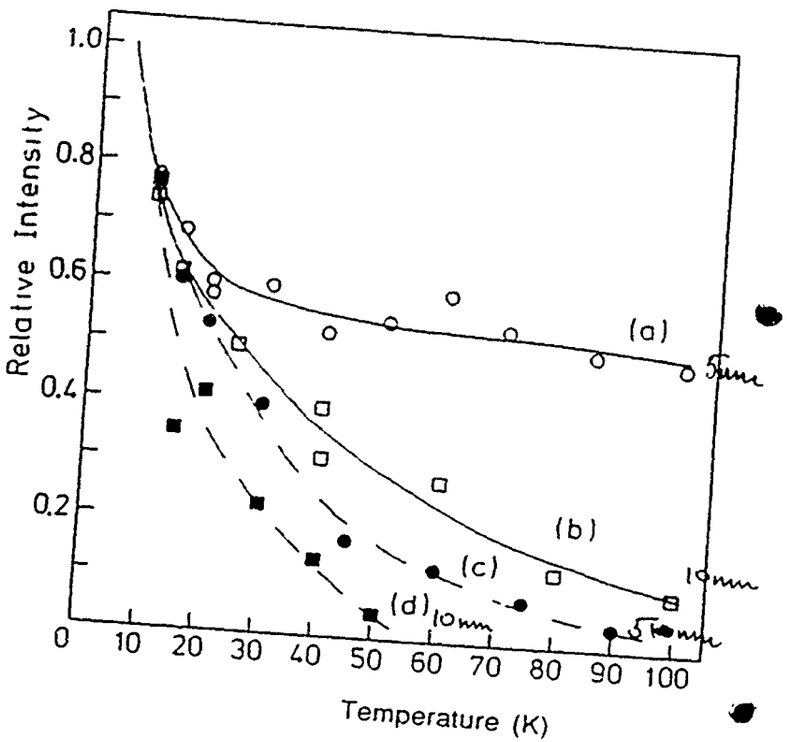
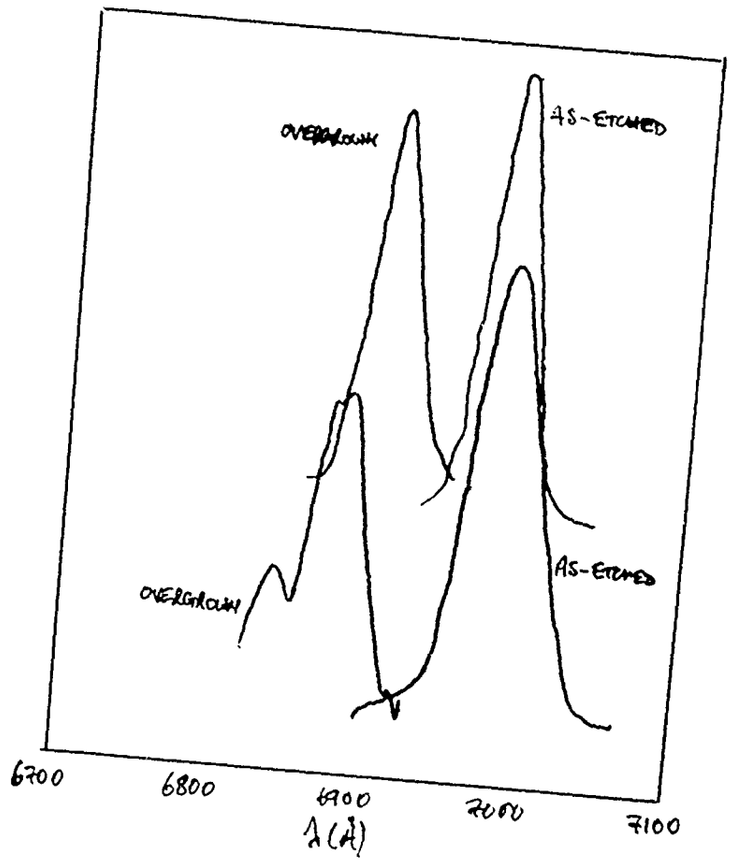
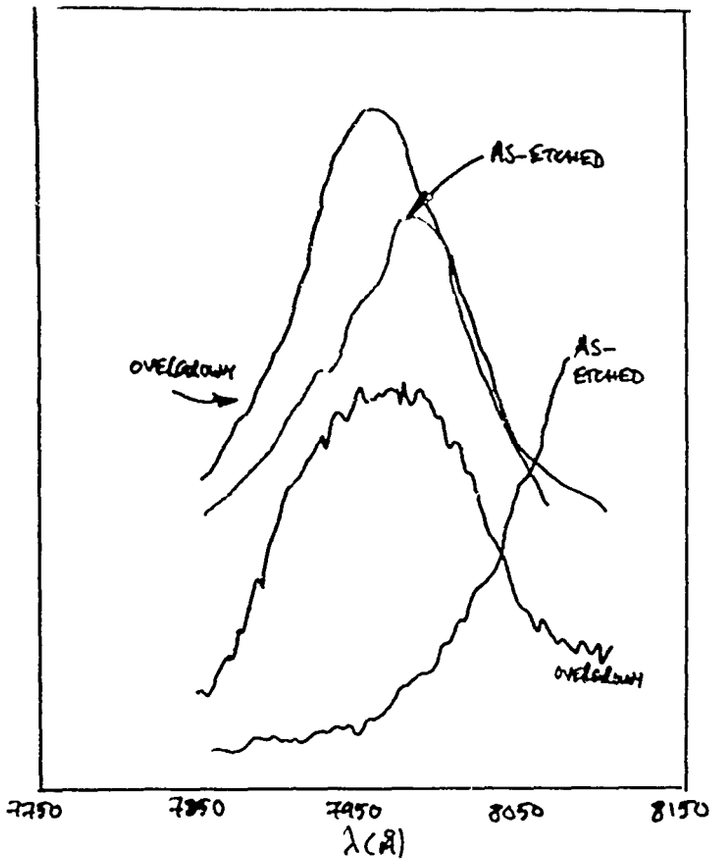


Figure 4.7 Temperature dependence of integrated luminescence intensity relative to that at 5K in (a) 5nm quantum well of GaAs/AlGaAs sample before processing and (c) after patterning into an array off 60nm pillars by CH4/H2 RIE. Curves (b) and (d) similarly before and after processing for 10nm well.

THERMALLY-ACTIVATED DELOCALISATION. FITS LOW-TEMP LOCALISATION MODEL



MBE MATERIAL (14Å QW)
LUMINESCENCE @ 4.2K

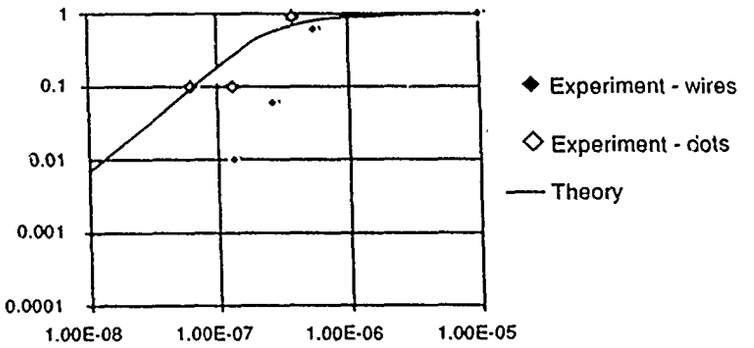


MOCVD MATERIAL
10 nm QUANTUM WELL
LUMINESCENCE @ 60K.

Quantum dot/wire lum[illegible]escence

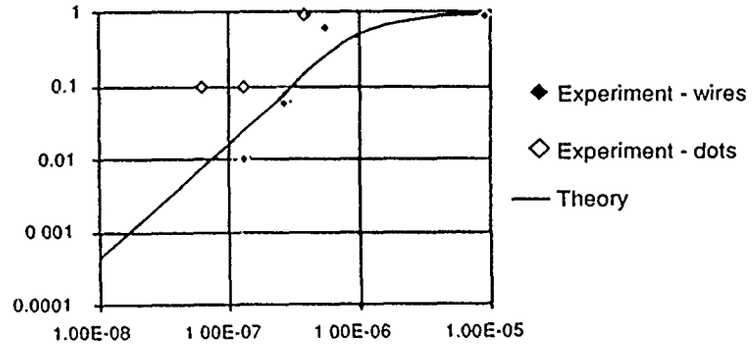
Dots:

τ = 0.2ns



Wires:

τ = 3ns



Assuming diffusion coefficient D = 1e-4 m^2/s and surface recombination velocity S=10^4 m/s for both wires and dots

Only the lifetime is changed.

CONCLUSIONS

* FREE STANDING DRY ETCHED WIRES DOWN IN GaAs/AlGaAs LUMINESCE DETECTABLY TO DIMENSIONS WHERE QUANTUM CONFINEMENT SHOULD BE OBSERVED

* PROVIDED THE PATTERNING PROCESS IS OPTIMISED, WELL CONTROLLED AND DEEP LATERAL DAMAGE TO AVOIDED.

* LUMINESCENCE IS QUENCHED BY EXCITON DIFFUSION AND NONRADIATIVE SURFACE RECOMBINATION
  - TEMP. DEPENDENCE OF LUMINESCENCE EFFICIENCY
  - OVERGROWTH EXPERIMENTS

  BUT IS WEAK IN DOTS AT LOW TEMPERATURE

* LUMINESCENCE OF QD's IS PRESERVED TO SMALLER DIMENSIONS THAN QW's FABRICATED IN THE SAME SAMPLES AND PROCESSED TOGETHER
  - EVIDENCE OF REDUCED RADIATIVE LIFETIME IN QD's DUE TO ENHANCED OSCILLATOR STRENGTH ??

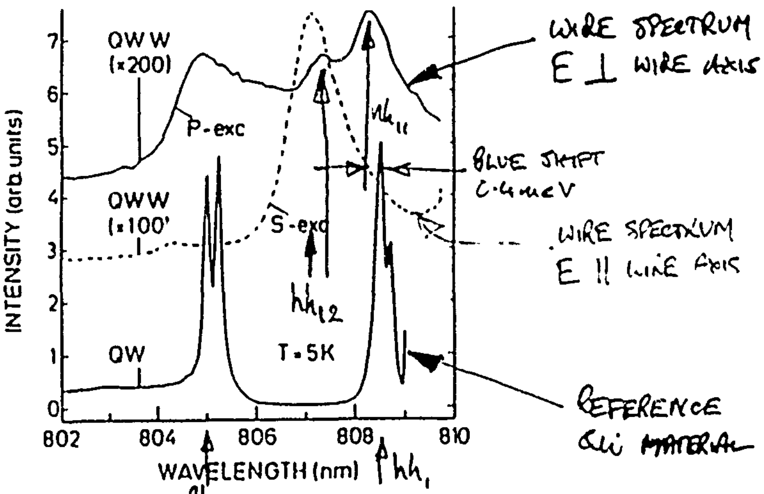


FIG. 1. PLE spectra at zero magnetic field of a QWW with $l \approx 70$ nm and of the corresponding reference QW. The spectra are shifted vertically with respect to each other for clarity and show the dependence on the polarization of the exciting laser light. For s-polarized (p-polarized) light **E** is parallel (perpendicular) to the wires. The excitation intensity was about 100 mW/cm^2.

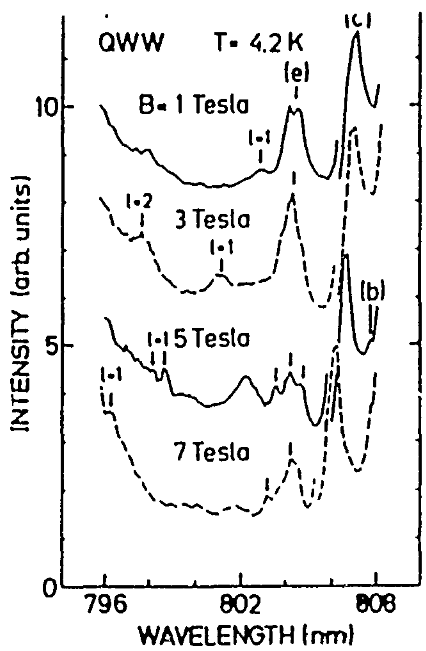


FIG. 2. Magnetic-field-dependent PLE spectra of a $l \approx 7$(nm QWW. (b), (c), and (e) denote the hh_{11}, hh_{12}, and lh_1 transitions, respectively; the index l labels the order of the inter-Landau-level transitions. The spectra are shifted vertically with respect to each other for clarity. The excitation intensity was about 100 mW/cm^2.

[redacted] Kath, Heitmann et al. PRL 63, 2124, 1989.

70nm WIDE WIRES BY HOLOGRAPHIC LITHOGRAPHY AND $SiCl_4$ DRY ETCHING.

a) BLUE SHIFT

b) SECOND 1D SUBBAND EXCITATION

c) POLARISATION DEPENDENCE

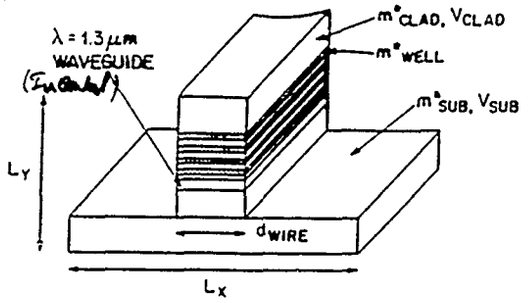


FIG 1 Schematic cross-sectional drawing of the quantum wire structure. Parameters used in the calculation of confined energy levels and wave functions are indicated.

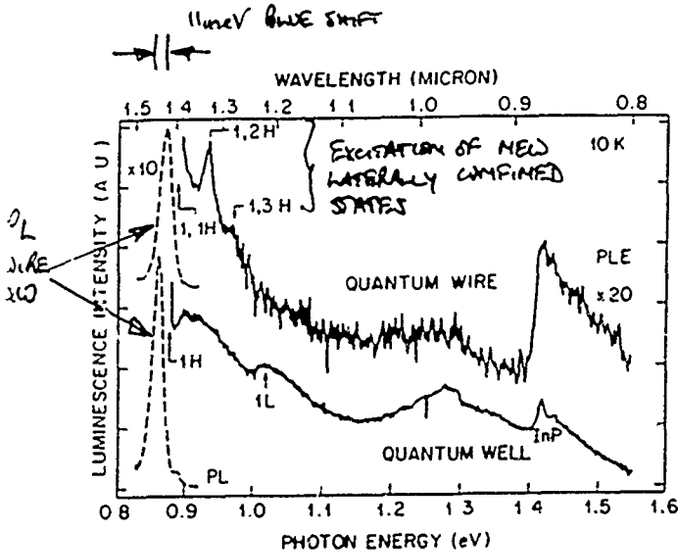
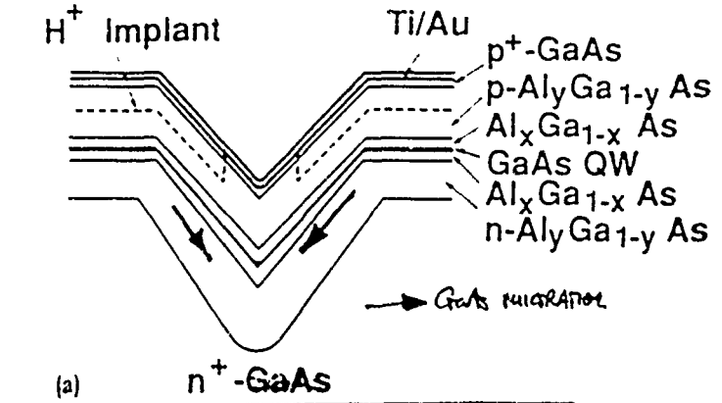
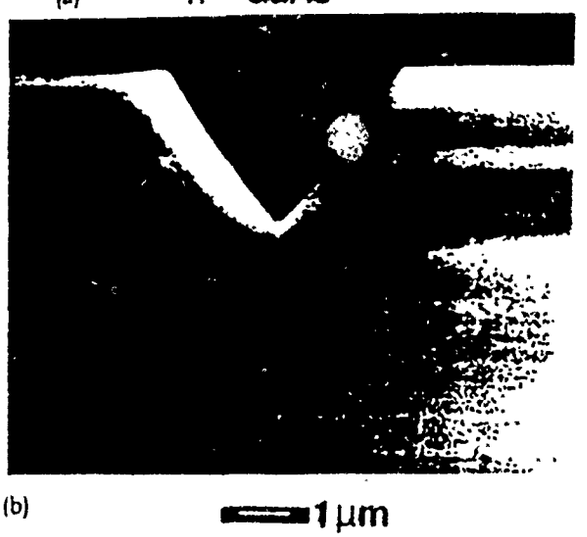
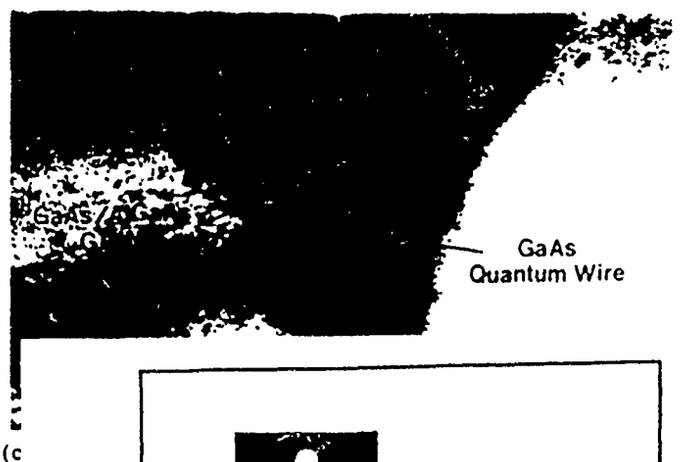
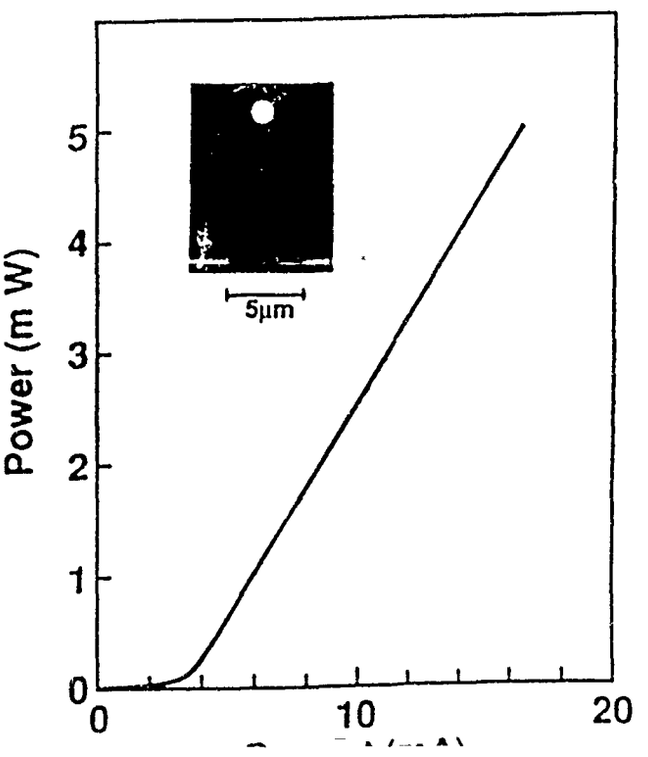


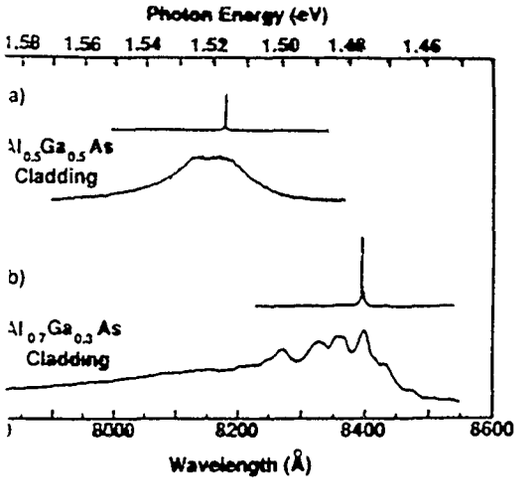
FIG. 2. Low-temperature PL (dashed lines) and PLE (solid lines) spectra of ~350-Å-wide quantum wires.











3. Emission spectra of quantum wire diode lasers with different Al fraction y in the cladding layers and different guiding layer thickness t Fig. 1). (a) $y = 0.5$, $t = 0.2\,\mu$m; lower: 40 mA, upper: 50 mA (cavity h $L = 2.07$ mm, $I_{th} = 41$ mA). Lasers with longer cavities (as long as im) lased at same wavelength. (b) $y = 0.7$, $t = 0.15\,\mu$m; lower: 22 mA, r: 25 mA ($L = 3.48$ mm, $I_{th} = 23$ mA).

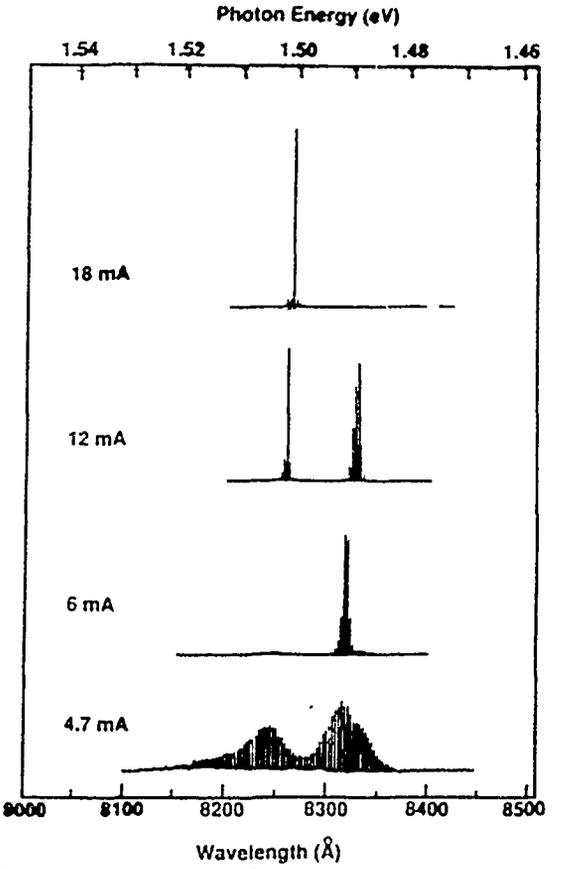
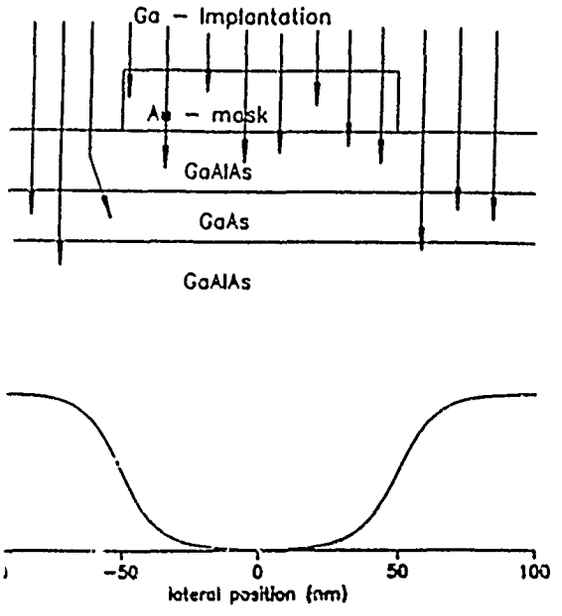
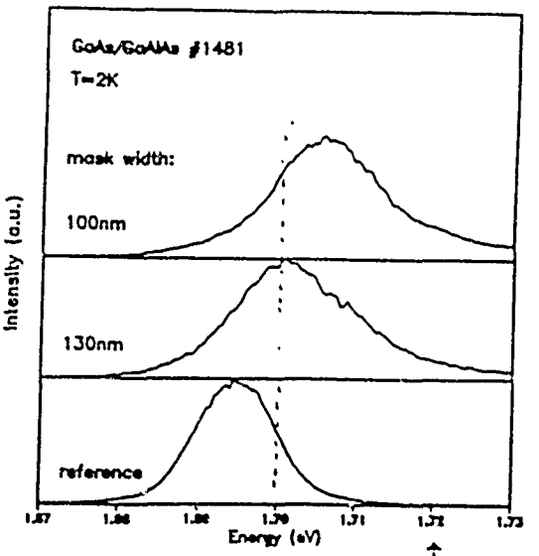


FIG 4 Evolution of emission spectra of a quantum wire semiconductor laser showing lasing at two adjacent quasi-one-dimensional subbands. $L = 270\,\mu$m, $I_{th} = 4.3$ mA.

KAPON et al. APL 55, 2716, 1989

- SINGLE QWW WIRES
- DIRECT GROWTH IN V-GROOVE
- STRUCTURE IN SPONTANEOUS EMISSION SPECTRUM
- SELECTIVE PUMPING OF WIRE STATES BY BAND-FILLING

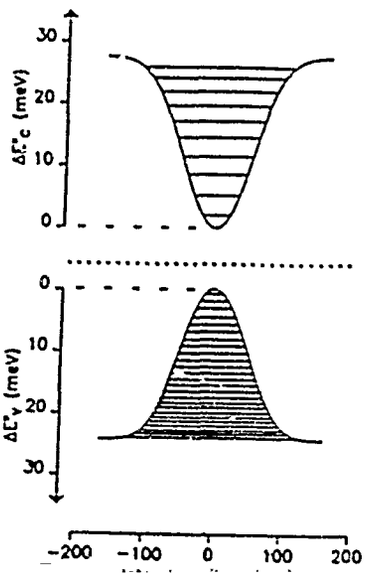




LUMINESCENCE FROM MASKED AREAS.

BLUE SHIFT DUE TO CONFINEMENT? (OR UNINTENTIONAL MIXING?)

a) Schematic sketch of the quantum well sample and wire mask. The typical wire widths the order of 100 nm. b) Lateral distribution of defects due to ion straggling under the the quantum well plane.



IMPURITY-INDUCED DISORDER (IID)
IMPLANT eg Ga MASKED OR FOCUSSED ION BEAM
ANNEAL
WELLS INTERMIXED IN IMPLANTED REGIONS LOCALLY INCREASING BAND GAP

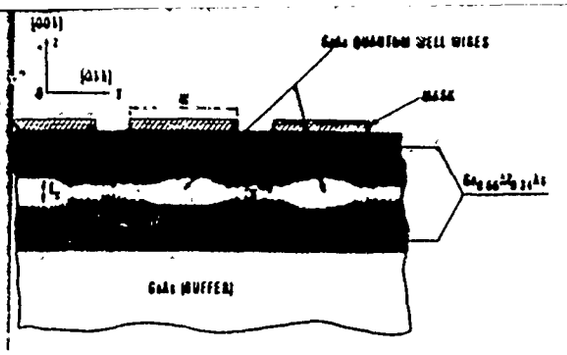


FIG. 1. Schematic not to scale of the structure after implantation and annealing. The mask position during implantation is also indicated. The wires are actually 1 μm apart.

CIBERT, PEYRAFF et al. APL 49, 1275, 1986

- IID AFTER Ge-IMPLANTATION
- BLUE SHIFT
- NEW FEATURES

  ? ELECTRONS DO NOT RELAX TO LOWEST LEVEL BECAUSE OF SMALL 1-D DOS + LONGER RELAXATION TIMES ?

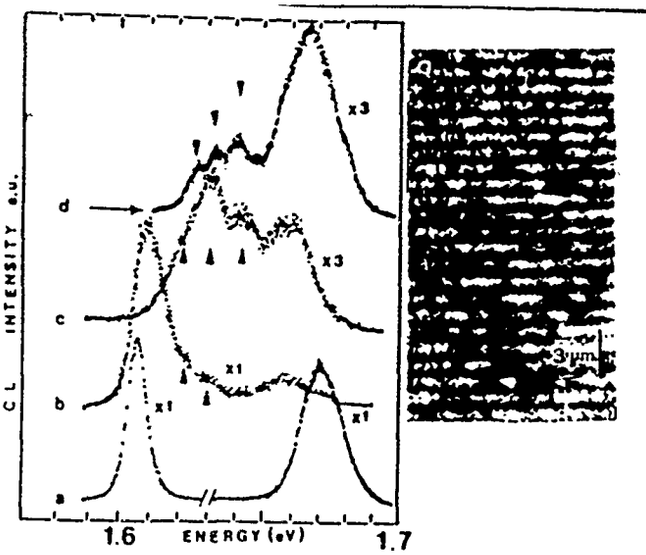


FIG 2 Cathodoluminescence spectra ($T < 15$ K) with the electron probe (~ 200 Å in diameter) on: a large masked area [(a) left], a large interdiffused area [(a) right], quantum well wires with mask sizes of 4500 Å (b), 1700 Å (c), and 1400 Å (d). (e) is a CL micrograph obtained by selecting the luminescence energy (1.647 eV) of the QW wires which appear as bright areas. The mask size for these QW wires is 600 Å.

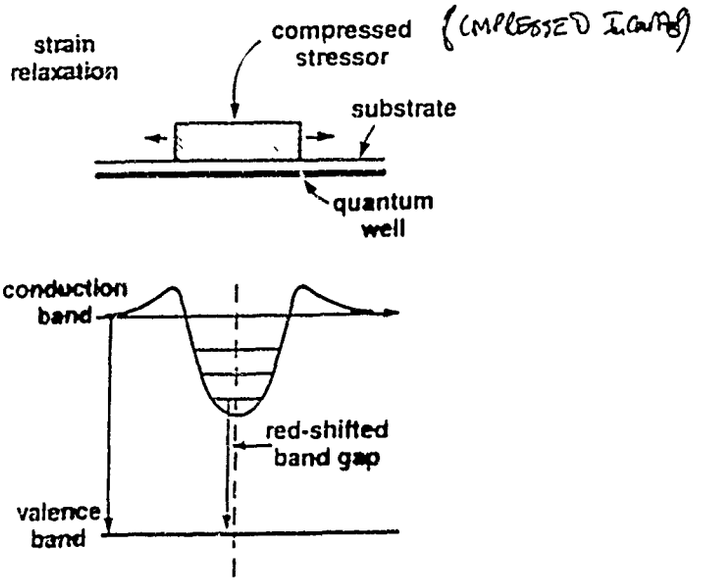


FIG. 1. Schematic diagram of the stressed wire and substrate, conduction- and valence-band modulation, and exaggeration of the red shift in the band gap under the wire.

Kash et al. APL 55, 681, 1989.

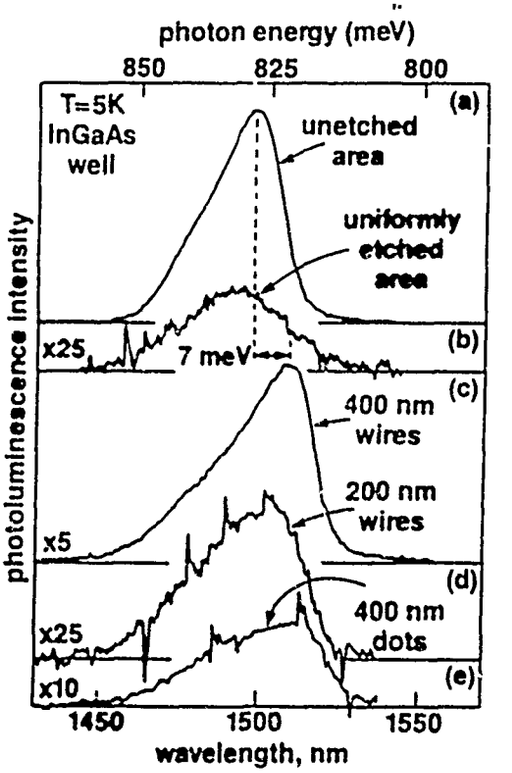
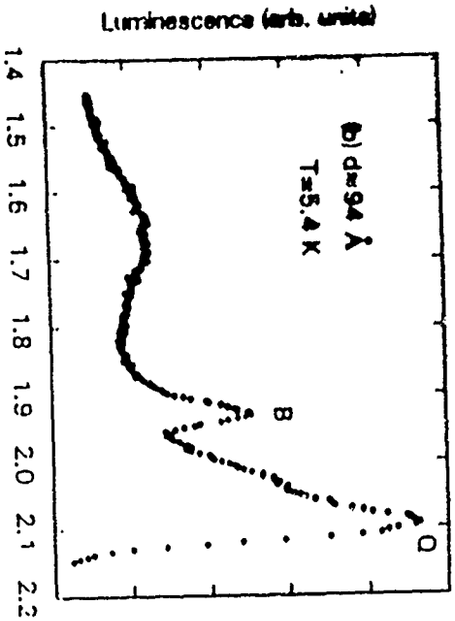


FIG. 2. Photoluminescence spectra of the InGaAs quantum well underlying (a) an unetched portion of the sample, (b) a uniformly etched region, and arrays of (c) 400-nm and (d) 200-nm-wide wires, and (e) 400-nm-wide (square) dots. All regions of the sample shown here were located within a few hundred microns of each other. Relative intensity scales are shown at the bottom left of each spectrum. Note the $\sim$9 meV red shift of the 400 nm wires, and slightly smaller and larger shifts for the 200 nm wires and 400 nm dots, respectively. Sharp spikes on the high-sensitivity spectra are from cosmic rays, to which the Ge detector is sensitive.



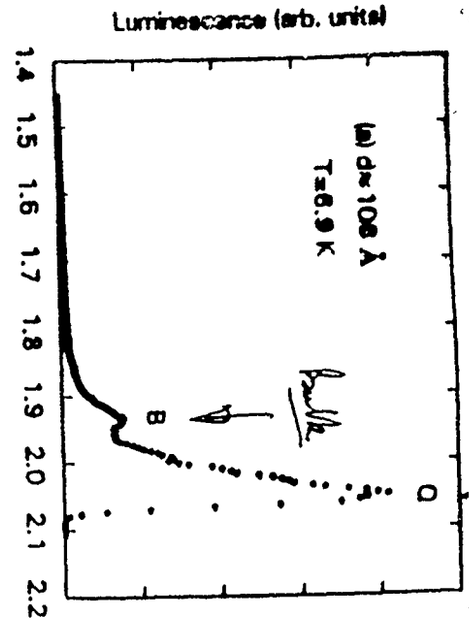


FIG. 1. (a) Low temperature luminescence spectrum from Corning Glass optical filter 2-59. Peaks *Q* and *B* label the microcrystallite and bulk luminescence, respectively. The crystallite size *d* was measured from the transmission electron microscopy negatives. (b) Same as (a), but for filter 2-61.

Biosynthesis of cadmium sulphide quantum semiconductor crystallites

C. T. Dameron*, R. N. Reese*, R. K. Mehra*, A. R. Kortan†, P. J. Carroll†, M. L. Steigerwald†, L. E. Brus† & D. R. Winge*

* University of Utah Medical Center, Salt Lake City, Utah 84132, USA
† AT&T Bell Laboratories, Murray Hill, New Jersey, 07974, USA

NANOMETRE-SCALE semiconductor quantum crystallites exhibit size-dependent and discrete excited electronic states which occur at energies higher than the band gap of the corresponding bulk solid[1–4]. These crystallites are too small to have continuous energy bands, even though a bulk crystal structure is present. The onset of such quantum properties sets a fundamental limit to device miniaturization in microelectronics[5]. Structures with either one, two or all three dimensions on the nanometer scale are of particular interest in solid state physics[6]. We report here our discovery of the biosynthesis of quantum crystallites in yeasts *Candida glabrata* and *Schizosaccharomyces pombe*, cultured in the presence of cadmium salts. Short chelating peptides of general structure $(\gamma\text{-Glu-Cys})_n\text{-}Gly$ control the nucleation and growth of CdS crystallites to peptide-capped intracellular particles of diameter 20 Å. These quantum CdS crystallites are more monodisperse than CdS particles synthesized chemically. X-ray data indicate that, at this small size, the CdS structure differs from that of bulk CdS and tends towards a six-coordinate rock-salt structure.

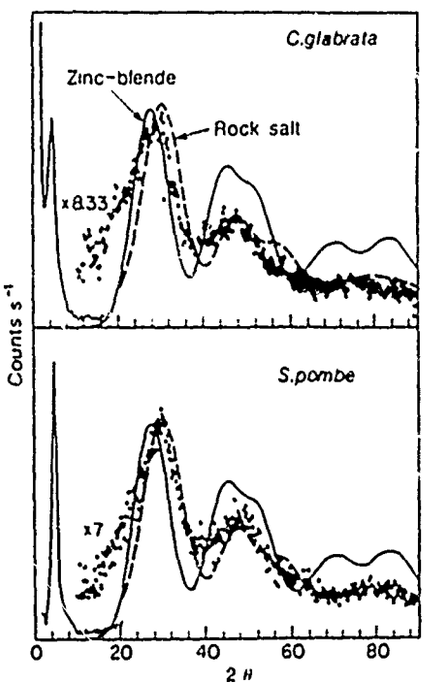


FIG. 1 1.54 Å powder X-ray patterns of desalted lyophilized Cd-γ-Glu peptide complexes from *C. glabrata* (top) and *S. pombe* (bottom). The samples were purified by a combination of ion-exchange and gel-filtration chromatography as described previously[9,15]. The samples were desalted by chromatography on Sephadex G-25 equilibrated and eluted with deionized water. Desalted samples were dried in X-ray tubes under vacuum. Solid and dashed lines are expected patterns for zinc-blende and rock-salt lattices, respectively, with an 8 Å coherence length. The sharp peak at low angle (4°) represents particle spacing in the dense isolate. The nearest-neighbour distances are 22 Å and 20 Å for isolates from *C. glabrata* and *S. pombe* respectively.

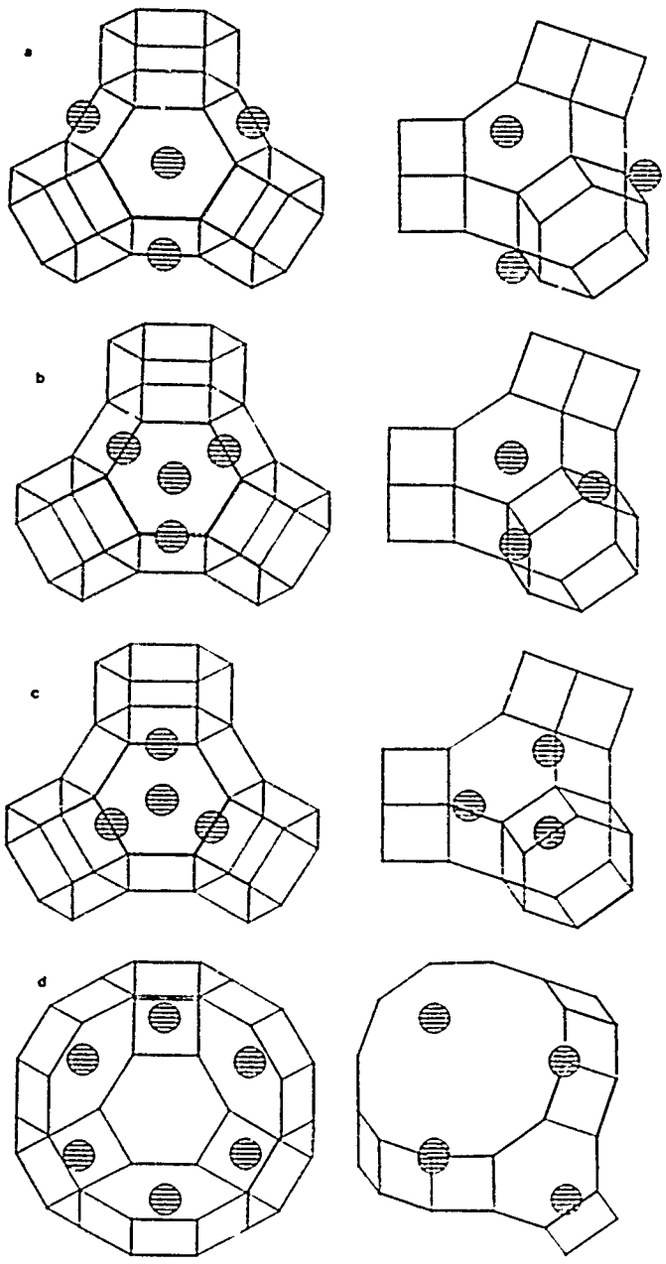


Figure 5. Extra-framework ion sites of zeolite Y shown as hatched circles. The framework is represented by sticks connecting the T atoms and omitting the oxygen atoms: (a) view of sodalite cage illustrating the SII site, (b) view of sodalite cage illustrating the SI′ site, (c) view of sodalite cage illustrating the SII′ site, (d) view of a supercage illustrating the SIII site.

- QUANTUM DOT/WIRE FABRICATION BY CONVENTIONAL MEANS IS DIFFICULT: PROCESSING REACHING RESOLUTION LIMITS.
- EVIDENCE OF CONFINEMENT IN ETCHED STRUCTURES
- "CLEANER" PROCESSING REQUIRED AND PROGRESS IS BEING MADE. 'PLANAR' TECHNOLOGIES ARE EMERGING.
- INGENIOUS METHODS FOR BEATING 10nm RESOLUTION BARRIER.

Introduction to quantum transport in electron waveguides

J. R Barker
Nanoelectronics Research Centre
Department of Electronics and Electrical Engineering
University of Glasgow
Glasgow G12 8QQ, Scotland, UK.

1. Introduction

A number of exciting low-dimensional semiconductor devices and structures have been fabricated recently where the key feature is the confinement of the conducting electrons to narrow channels which have dimensions comparable to or smaller than the inelastic coherence length at an appropriate temperature. The most interesting devices possess some feature sizes commensurate with the de Broglie wavelength. Examples of such structures include quantum wires, ring structures and split-gate squeezed channel devices. To a certain extent these structures may be pictured as *electron waveguides*. Originally of interest for the construction of electron interferometer devices a number of phenomena have been discovered in electron waveguides which bring out more of the classical picture of the electron than had been originally appreciated: particulary focussing effects and the cycloidal motion peculiar to magnetic edge states. As we shall see in this lecture even the phenomenon of conductance quantisation in quantum point contacts depends more on the quantisation of the carrier supply function than on an intrinsic quantum transport process. This paradox may be understood on the basis of rigorous quantum transport theory. Although practical devices are much further from development than had been hoped for in the mid-nineteen eighties, electron waveguide structures offer one route to the study and possible application of granular electronics. Indeed they provide an environment for exposing some outstanding difficulties with manipulating devices containing very few carriers The present lecture is aimed at linking the more specialised and detailed lectures in the School by developing a full quantum ballistic transport picture which reveals both the classical and non-classical features of practical electron waveguides. Edge state effects are not treated explicitly although we shall have cause to comment on their "classical" success.

Until recently electron waveguide structures were modelled by transmission-line like models using *one-dimensional*, single channel or multi-channel conductance techniques based on calculations for the electron transmission matrix. Interference phenomena such as the Aharonov-Bohm effect cannot be explained quantitatively by these models because the true electron confinement potential is at least two-dimensional and gives rise to effects such as the variation of zero-point energy and transitions between lateral confined modes along the channels. Added to this is the problem that the true confinement potential contains large non-self-averaging spatial fluctuations and in itself is a difficult target for modelling

2. Electron waveguides

2.1. Structure

Most heterostructure electron waveguides are based on a re-structuring of the two-dimensional electron gas using direct etching or a patterned split-gate to squeeze down the electron gas into a narrow channel or 'quantum wire' . This quasi-one-dimensional structure may be further restricted to form a quantum point contact or a quantum dot .The basic principles are illustrated in figure 1

Ideally, the resulting channel should have only one transverse confined state occupied corresponding to a monomode waveguide The basic quantum wires can be combined into more complex structures: multi-port waveguides, tunnelling structures and interferometer devices such

as Aharonov-Bohm rings and stub tuners. A selection of structures which derive from the HEMT or MODFET geometry are shown in figures 2 and 3.

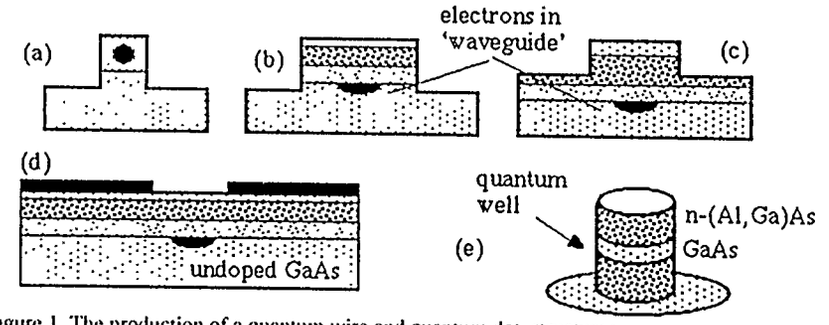


Figure 1. The production of a quantum wire and quantum dot structures
(a) epitaxial n+ GaAs wire; (b) modulation doped deep-etched GaAs wire; (c) modulation doped shallow etchedGaAs wire; (d) split-gate wire; (e) quantum dot.

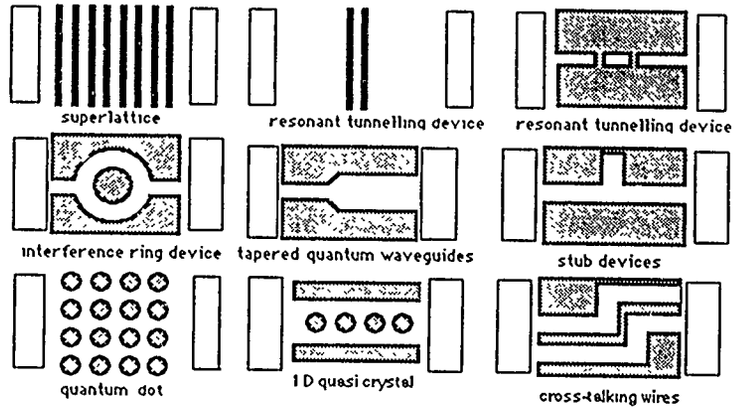


Figure 2 : Gate patterns for lateral nanostructures.

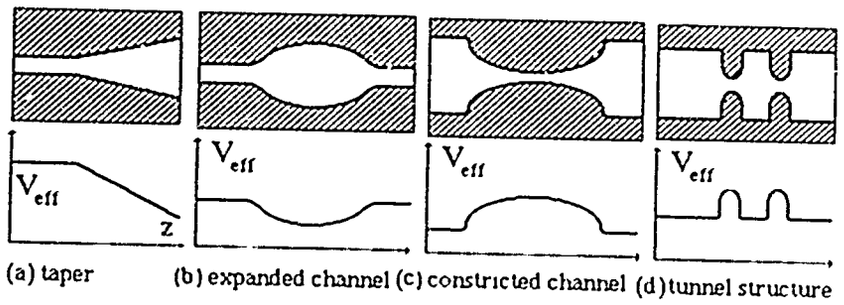


Figure 3 Non-uniform electron waveguides and the corresponding effective potentials derived from the variation of *zero* point energy with position

2.2. The confinement potential

For simplicity let us assume a parabolic effective-mass Hamiltonian to describe electrons confined to a quasi-two dimensional layer (x-z plane) in a semiconductor heterostructure by a *confinement potential* $V_c(y)$

$$H = (p_x^2 + p_y^2 + p_z^2)/2m^* + V_c(y) = (p_x^2 + p_z^2)/2m^* + (p_y^2/2m^* + V_c(y)) \qquad (1)$$

In conventional analyses the confinement potential V_c is calculated from the Schrodinger equation and Poisson's equation by considering the fields arising from the band-edge discontinuities at the heterojunction(s), charge spill over and the *mean field* of the remote donors and surface (including gate) charges Consequently, V_c is a relatively smooth idealised potential Its most elementary form is a triangular quantum well

Choosing the z-axis as the axis of our prototype electron waveguide we restrict the lateral (or transverse) degree of freedom in the x direction by imposing further confinement forces The corresponding model Hamiltonian is

$$H = p_z^2/2m^* + (p_x^2 + p_y^2)/2m^* + V_c(R, a_\mu(z)) + V(R,z) = p_z^2/2m^* + H_c + V(R,z) \qquad (2)$$

where $V_c(R, a_\mu(z))$ is the full mean confinement potential for the quantum waveguide and $V(R,z)$ is an external applied potential The vector R lies in the x-y plane We have allowed the confinement potential to depend on a finite number of parameters $a_\mu(z)$, $\mu=1$ N which may vary with position z along the guide

Typically, as sketched in figure 4(a) V_c will have a roughly parabolic profile throughout, or at least at the boundaries of, the transverse region as sketched in Figure 4(c) (see the work of Laux and Stern [1] andKumar, Laux and Stern[2]) The addition of a uniform external electric field aligned along the z-direction results in the total potential $V_c + V$ If this total potential is a general quadratic function of the coordinates it may be termed a *gutter potential* V_g sketched in figure 4(b) , it is parametrised by four constants $a_1 = V_0$, $a_2 = E_c$, $a_3 = \omega_x$, $a_4 = \omega_y$ and provides a good approximation for many real situations

$$V_g = V_0 + eE_c z + (1/2)m^*[\omega_x^2 x^2 + \omega_y^2 y^2] \qquad (3)$$

In section 4 2 we shall show that non-dissipative transport along a gutter potential is *classical*: but quantum effects can occur in the extracted current flow due to the quantised injection of charge into the discrete channel states. We shall use the expression - *inhomogeneous gutter potential*- to refer to the situation where ω_x, ω_y are functions of position z along the waveguide.

V_c may be sufficently abrupt that a square well or multiple square well model is appropriate.

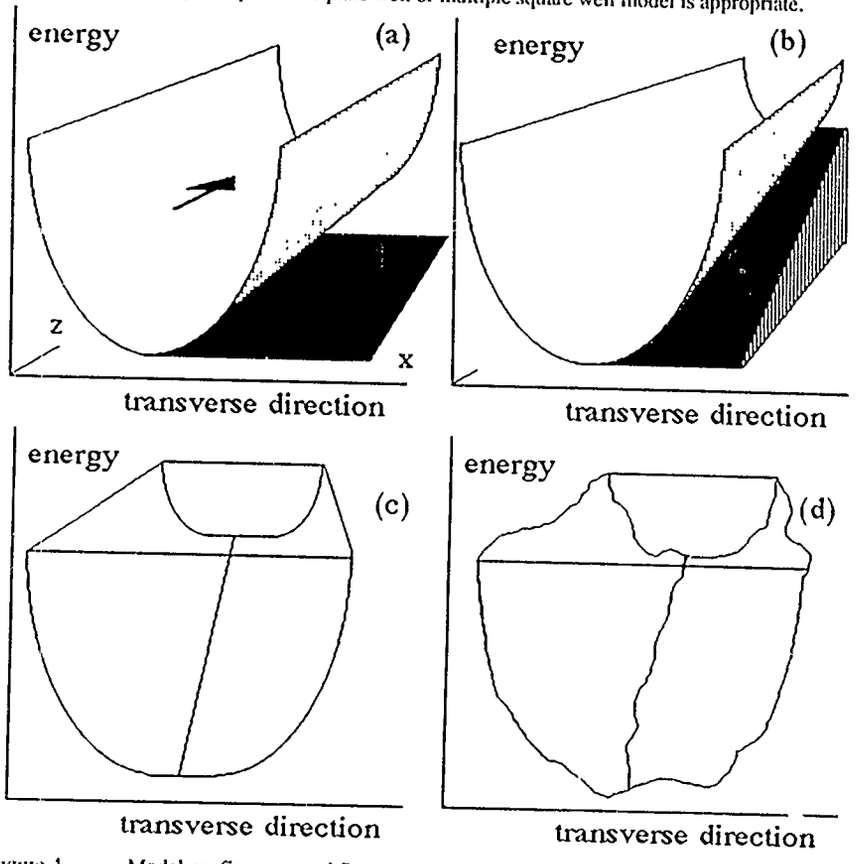


Figure 4 Model confinement and fluctuation potentials for a quantum wire.
(a) Idealised parabolic cross-section; (b) the gutter potential - parabolic transverse potential plus linear longitudinal potential; (c) real quantum wires have a flatter than parabolic confinement potential; (d) the fluctuation potential produces significant distortions to the idealised models.

Although expression (2) is perfectly general it is best-suited to a quantum wire that is long and oriented in the z direction but which is inhomogeneously narrow in cross section. With suitable parameterisation this model will describe more complicated situations such as a loop in a long wire or a Y-junction and other multiply-connected geometries (see figure 5 and [3,4]) It is generally the

case that the confinement in the y direction (direction perpendicular to the Q2DEG) is much stronger and more localised than the lateral confinement (x direction) because of the higher precision afforded by heterolayer formation In many cases this leads to the useful approximation that the y-motion is restricted to the lowest allowed discrete energy level and the guide becomes essentially two dimensional

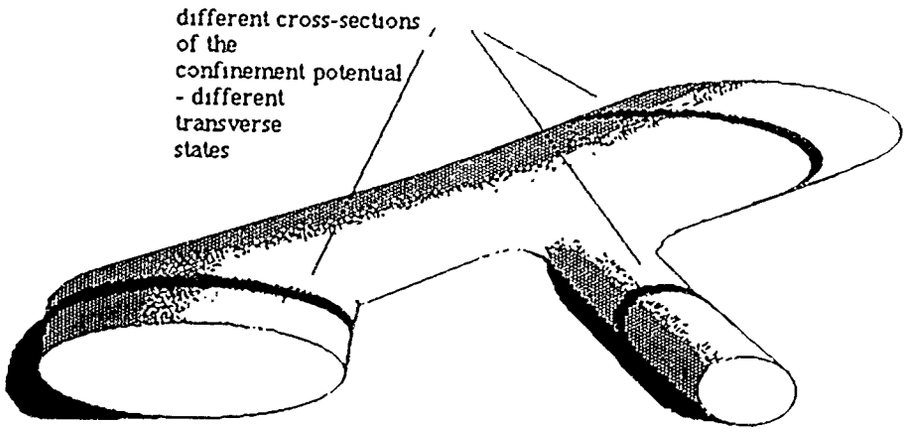


Figure 5 A generic electron waveguide - potential contour in 3-D real space showing different cross sections

2.3. The fluctuation potential

The confinement potential has so far been assumed to derive from the mean field of the donors and various image charges in the host heterostructure It has been shown that it must be supplemented by a *fluctuation potential* V_f which arises because of the *discrete random* spatial distributions of donors [5,6] The effect is to distort the *full confining potential* $V_{cf} = V_c + V_f$ into a *random* structure which is illustrated in figure 4(d)

3. Semi-classical particle model

In the ballistic transport regime it is convenient to use transmission matrices for the description of electron wave-propagation in the confinement potential of the quantum wires As with all quantum phenomena the details depend critically on the boundary conditions Most recent studies have involved electron waveguide structures which merge into relatively large thermal reservoirs in the 'contact' or 'electrode' regions (see the generic picture in figure 6) Such systems have proved to be well-described by a set of formulae for the effective conductance between the connected reservoirs known as the Landauer[7] or Landauer-Buttiker formulae (reviewed in [8] and which rely on the computation of appropriate transmission coefficients The simplest case arises for just two reservoirs connected by a single uniform quantum wire It can be understood by a simple one-dimensional quasi-classical argument [9] which we now outline

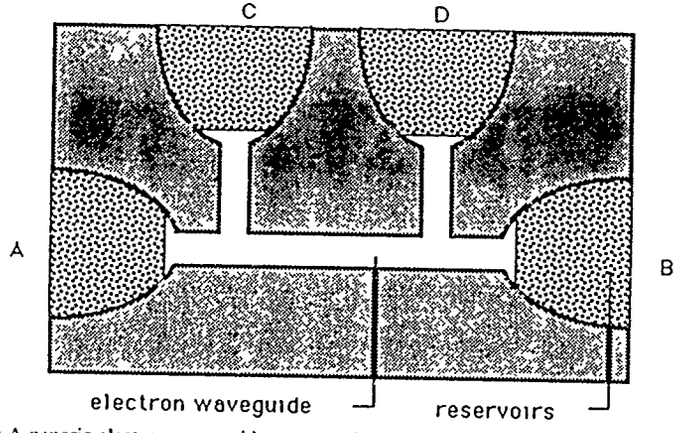


Figure 6 : A generic electron waveguide connected to four thermal reservoirs of electrons.

3.1. The classical Liouville equation

Classical ballistic transport may be described by phase-space techniques based on Hamilton's equations of motion. Because classically, a particle can be described uniquely by its instantananeous position R and momentum P it is convenient to describe one or more non-interacting particles by a phase space probability distribution f(R,P,t) normalised to the number density of particles and which physically is conserved according to Liouville's theorem df/dt = 0. Since f is a perfect differential we have mathematically

$$[\ \partial_t + dR/dt\ \partial_R + dP/dt\ \partial_P\ ]f(R,P,t) = 0 \qquad (4)$$

or physically, by substitution from Hamilton's equations,

$$dR/dt = \partial_P H\ ; \qquad dP/dt = -\partial_R H \qquad (5)$$

$$[\ \partial_t + \partial_P H\ \partial_R\ -\partial_R H\ \partial_P\ ]f(R,P,t) = 0 \qquad (6)$$

For a simple Hamiltonian $H = P^2/2m^* + V(R)$ we set $F = -\partial_R H$ to obtain the most simple form of Liouville's equation or the collisionless Boltzmann equation

$$[\ \partial_t + P/m\ \partial_R + F\ \partial_P\ ]f(R,P,t) = 0 \qquad (7)$$

If the solution to (6) or (7) is known one may compute the current density according to the statistical prescription

$$j(R,t) = e \int d^3P(P/m)f(R,P,t) \qquad (8)$$

(where we might also put in a factor of two for spin degeneracy)

Any point initial distribution in phase space, eg $f = \delta(R-R_0)\delta(P-P_0)$ will unfold as a unique phase space trajectory which satisfies Hamilton's equations Any area of phase space will be mapped by Hamilton's equations into an equal area of phase space at a later time We may model the injection and extraction of particles by inserting suitable generation and recombination terms (source and

sink descriptions) on the right hand side of the Liouville equation which is a convenient way to manage open systems

3.2 Perfect channel

Let us apply the classical formalism to the electron waveguide problem, using initially, a one-dimensional model for simplicity. Referring to figure 6 , suppose we ignore contacts C and D and suppose that the two reservoirs A and B are identical except that a potential difference V exists between them (figure 7) The region between the reservoirs is conservative and carriers are injected or extracted at thermal sources G and perfectly absorbing sinks R located at the reservoir boundaries x = 0 and x = L. A phase space portrait of this system is shown in figure 7

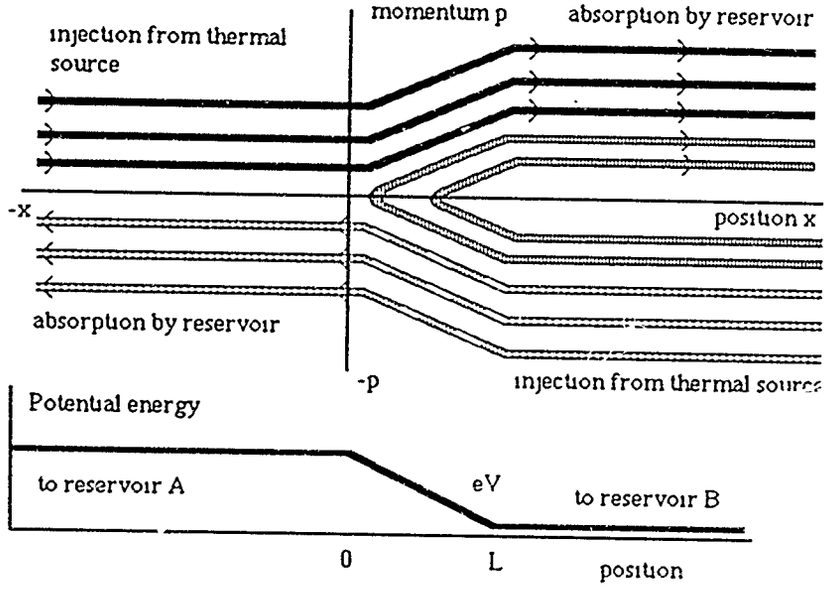


Figure 7 Phase space portrait of trajectories through a guide with a uniform field

The carrier phase-space distribution f(x,p,t) then satisfies the open-system Liouville equation

$$[\partial/\partial t + v(p)\partial/\partial x + F(x)\,\partial/\partial P\,]f(x,p,t) = G - R \qquad (9)$$

$$G = [\theta(v)v\delta(x)\,f_0(\varepsilon_p) + \theta(-v)v\delta(x-L)\,f_0(\varepsilon_p + eV)]\rho_P \qquad (10)$$

$$R = \theta(v)v\delta(x-L)\,f(x,p,t) + \theta(-v)v\delta(x)\,f(x,p,t) \qquad (11)$$

where θ is the unit step function, F is the local conservative force in the waveguide and f_0 is the Fermi-Dirac distribution for the reservoirs. ρ_P is the density of states in momentum space for the injected carriers; we take this as $1/h$ Here lies our only concession to quantum mechanics we assume that the injecting reservoirs are maintained in thermal equilibrium with quantum distributions $f_0\,\rho_P$ The current may be found from

$$J = 2e\int dp\,v(p)f(x,p,t) \qquad (12)$$

where the 2 accounts for spin

The most striking property of very high mobility electron waveguides is the quantization of the conductance of short channels in units of $2e^2/h$, a result only discovered in 1988 (reviewed in [10]) and which may be understood on the Landauer picture mentioned previously. Conductance quantisation may also be described on our semi=classical picture by using the quantum supply functions inherent in equation (10)

Equation (9) may be solved by path-variable techniques (just using an integrating factor in this trivial case) since we have conservative flow described by phase-space trajectories. Choosing f(x,p,t=0) = 0 within the wire and letting $f_0(\varepsilon_p)$ = the low temperature limit = $\theta(\mu - \varepsilon)$ and solving for long times we find the current as :

$$J = (2e/h)\{\int d\varepsilon\,\theta(\varepsilon-\mu-eVx/L) - \theta(\varepsilon-\mu-eVx/L+eV)\}$$

Finally,

$$J = (2e^2/h)\,V \qquad (13)$$

We might call this a 'classical' Landauer formula It should be stressed that the transport is classical but the supply of carriers by the reservoirs is quantised

3.3. Multiple channels

For a two-dimensional channel with transverse states ε_n at the entry and exit to the guide we replace f_0 by

$$f_0 = \Sigma\, f_n\,, \quad f_n = \theta(\mu-\varepsilon_n-\varepsilon)$$

the formula (13) trivially generalises to $J = (2N\,e^2/h)\,V$ whence the conductance is found as

$$G = (2N\,e^2/h) \qquad (14)$$

where N is the number of occupied sub-bands The quantisation of conductance is thus demonstrated.

3.4. Reflective channels

If we add a conservative 'scattering potential' of maximum energy W into the guide with a classical transmission coefficient $T(\varepsilon)$ shown in figures 8, 9 we find the simple generalisation of (13) to be

$$J = (2\,e^2/h)\,T(\mu)\,V \qquad (15a)$$

The extension to a multi-channel system may be made by including a transmission coefficient T_{nn} for each channel and a coefficient T_{nm} for possible transmission from one channel to another.

$$J = (2\,e^2/h)\sum_n\sum_m T_{nm}\,V \qquad (15b)$$

Technically, T_{nm} is the probability of flux input in channel n being transmitted out in channel m and the sums are over all input modes with energies below the Fermi energy

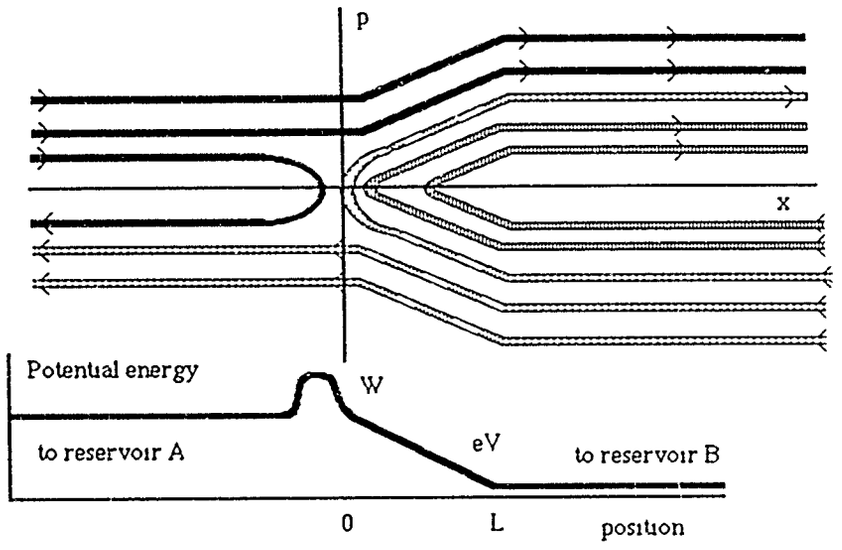


Figure 8 Addition of a scattering potential to the guide

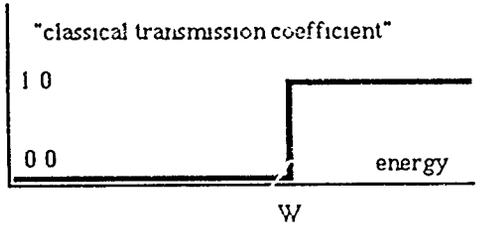


Figure 9 Classical transmission coefficient for a barrier of height W

The familiar finite-temperature Landauer formulae can be recovered similarly The transition to the familiar mobility formulae which obtain from Boltzmann's equation (the collisional version) can also be easily followed in this picture by varying the mean free path from infinity to less than the channel length The jump to quantum systems involves merely replacing the classical transmission coefficient T by the correct quantum-mechanical version computed for the guide As discussed in detail in section 4 the quantum analogue of the phase-space distribution function f is the Wigner distribution The semi-classical argument works quite well and some justification may be found for a waveguide with a slowly-varying inhomogeneous gutter confinement potential [21] The Wigner function then obeys the classical Liouville equation, although care is needed in handling it near the reservoir regions

The Landauer-Buttiker formalism has been adapted to complicated geometries such as rings and multi-port structures and provided no strong inter-sub-band mixing takes place one may compute the appropriate transmission coefficients (or transfer matrices) as though we had a waveguide or transmission line problem [22,23] and the conductance may be recovered with ease.

4. Quantum model and Wigner functions

4.1. Quantum phase-space distributions

We now digress to show the origins of the classical phase distribution and the conditions under which it is a good approximation to the quantum distribution (see also[11]). Suppose a system is in a general state described by the density matrix ρ. The most direct recipe for computing the quantum-statistical expectation value of an observable Λ for that system involves:

$$\langle\Lambda\rangle = \sum_{r_1,r_2} \langle r_1|\Lambda|r_2\rangle\langle r_2|\rho|r_1\rangle \qquad (16)$$

The matrix elements of Λ and ρ are functions of two positions r_1 and r_2. From these we might construct a representative central position dependence on a vector R and a dependence on a central momentum vector P. A shift of origin to R on a line connecting the two fixed points could be used to locate r_1', r_2'. It follows that r_1, r_2 may be expressed entirely in terms of the relative vector $r = r_1 - r_2$ and the central locator R, where the parameter σ is in the range (0-1).

$$\langle r1|\Lambda|r2\rangle = \Lambda(R,r,\sigma) = \langle R+\sigma r|\Lambda|R-(1-\sigma)r\rangle \qquad (17)$$

$$\langle\Lambda\rangle = \sum_{R,r} \Lambda(R,r,\sigma)\rho(R,-r,\sigma) \qquad (18)$$

The relative vector r is a natural candidate for a transformation to the momentum representation, so let us define the phase-space representation of an operator Λ by

$$\Lambda(R,P,\sigma) = \sum_{r} \Lambda(R,r,\sigma)\exp(-iP.r/\hbar) \qquad (19)$$

$$\Lambda(R,r,\sigma) = (2\pi\hbar)^{-N} \sum_{P} \Lambda(R,P,\sigma)\exp(iP.r/\hbar) \qquad (20)$$

where N is the number of dimensions Then (18) becomes:

$$\langle\Lambda\rangle = \sum_{R,P} \Lambda(R,P,\sigma)f(R,P,\sigma) \qquad (21)$$

$$f(R,P,\sigma) = \rho(R,P,\sigma)\,(2\pi\hbar)^{-N} \qquad (22)$$

This form resembles a classical phase-space average over a distribution $f(R,P,\sigma)$. It includes (or can be used to construct) the majority of quantum distributions $f(R,P,\sigma)$ that have appeared in the literature (the cases σ=0 or 1 involve the product of momentum and direct space wave-functions).

The case σ = 1/2 corresponds to the Wigner-Weyl transformation and we shall use $f(R,P) = f(R,P,\sigma=1/2)$ to represent the *Wigner distribution* function from now on For pure states $\rho = |\psi\rangle\langle\psi|$, or in terms of wavefunctions $\psi(R,r) = \psi^*(R-r/2)\psi(R+r/2)$. The density matrix and hence Wigner function is thus directly calculatable if the wavefunction is known Of the possible

functions (22) , only the Wigner distribution (σ =1/2) is *real*, although real distributions could be built out of linear combinations of functions with different σ parameters

4.2. Quantum ballistic transport

Consider the ballistic transport problem, which we define as the situation when the effects of dissipative collisions are vanishingly small Switching to the centre of mass coordinates (and set $\hbar$ = 1 temporarily for readability) let us take the matrix elements <R-r/2| |R+r/2> of the quantum Liouville equation

$$i\partial_t \rho = [p^2/2m^* + V(r), \rho] \tag{23}$$

$$[\, i\partial_t + (1/m^*)\partial_r \partial_R - V(R + r/2,t) + V(R-r/2,t)]\rho(R,r,t)=0 \tag{24}$$

$$[\ \partial_t + (P/m).\partial_R\ + L]f(R,P,t)=0 \tag{25}$$

where Lf is the "driving field term"

$$Lf(R,P,t)= -\int d^3r\ e^{-iP.r}[\,V(R + r/2,t) - V(R-r/2,t)]\rho(R,r,t) \tag{26}$$

The general case may be put in a more elegant form ,
by defining an effective force $F_{eff}(R,r,t)$ and its Wigner-Weyl transform $F_{eff}(R,P,t)$

$$F_{eff}(R,r,t) .r = [V(R + r/2,t) - V(R-r/2,t)] \tag{27}$$

$$F_{eff}(R,P,t) = \int d^3r\ e^{-iP.r/\hbar} F_{eff}(R,r,t)/(2\pi\hbar)^3 \tag{28}$$

The resulting equation of motion is a *non local* version of the classical Liouville equation or collisionless Boltzmann equation

$$[\ \partial_t + (P/m)\ \partial_R\ ]\ f(R,P,t) + \int d^3P'\ F_{eff}(R,P',t)\ \ \partial_P\ f(R,P-P',t)=0 \tag{29}$$

Planck's constant is hidden away in the quantum force term (28) where we have re exposed it The equation of motion may be formally solved by path-variable techniques along the lines of the Chambers-Rees method but so far only the simplest of one-dimensional problems have been attempted. Convergence problems are very severe, a problem which may be traced to (a) the breakdown of the area-preserving mapping property of classical phase-space, (b) the non-compactness of f

One should note some snags First, the Wigner-Weyl transformation leads to Wigner functions which may assume *negative* as well as positive values, which rules out the interpretation of f(R,P) as a simple probability distribution although it is a perfectly satisfactory statistical distribution function. More seriously , f(R,P) can assume non-zero values at points R where the wave-function is zero, ie in non-physical regions (as indeed can the "A"-density A(R,P)); technically f has non-compact support. Similarly the Wigner function may be non-zero in regions where the momentum wavefunction is zero Finally, we note that f is not a perfect differential except for quadratic Hamiltonians and hence it cannot generally be used to prove the existence of phase-space trajectories We shall return to this point in our lecture on granular nanoelectronics

4.3. The gutter potential

The Wigner function has the extra-ordinary property that for a simple Hamiltonian H=T(p)+V(r) which is a *general quadratic* function of position and momentum the *non-local quantum* equation of motion reduces to the corresponding *local classical* Liouville equation We can easily prove this by writing the potential as the general quadratic form or gutter potential

$$V(R,t)= v_0 + v_1.R + R.V_2\,R \tag{30}$$

where v_0 , v_1, V_2 are arbitrary constant scalars, vector and tensor quantities respectively. The driving field term collapses to a familiar form .

$$[\,V(R + r/2,t) - V_{eff}(R-r/2,t)]=(v_1 + V_2\,R + R.V_2).r= -F.r \tag{31}$$

where F(R), the effective force, is identical to the classical force derivable from (30) In that case, equation (29) may be Wigner transformed to the form,

$$[\ \partial_t + (P/m)\ \partial_R\ + F.\ \partial_P\ ]f(R,P,t)=0 \tag{32}$$

which is just the classical Liouville equation

4.4. Gauge invariant formulation

In the presence of an arbitrary electromagnetic field described by the vector and scalar potentials A(R,t), φ(R,t), the electron velocity is determined by

$$m^*v = p - eA \equiv \pi \tag{33}$$

and the Hamiltonian becomes

$$H = (p - eA)^2/2m^* + V_c + V_f + V + e\varphi \tag{34}$$

The theory is invariant under the gauge transformation

$$A \to A + \nabla\chi, \qquad \varphi \to \varphi - \partial\chi/\partial t \tag{35}$$

provided the Wigner distribution is re-defined as

$$f(R,\pi) = (2\pi\hbar)^{-3} \sum_r \rho(R,r)\exp(\ -i\pi.r/\hbar)\exp(-ie\int_0^1 d\sigma\ r.A(R-r/2\ +\ \sigma r,t)\ ) \tag{36}$$

For a uniform magnetic field and a general *linear* electric field we again obtain a classical transport equation for the Wigner distribution provided the total confining potential is at most quadratic:

$$[\ \partial_t + (\pi/m).\partial_R\ + (eE + e/m^*\pi \times B + F).\ \partial_\pi\ ]f(R,\pi,t) = 0 \tag{37}$$

This remarkable result shows once again that even with a uniform magnetic field and a harmonic potential the electron transport is exactly classical.

4.5. Resolution of the paradox

The results of sections 4 3 and 4.4 show that the equation of motion of the Wigner distribution function is exactly classical when the potential fields are quadratic and the magnetic field is uniform (of whatever strength) It is not surprising then to find that many observations of "quantum" ballistic transport admit a simple classical explanation because a phase-space or configuration space*trajectory* description most certainly exists under the conditions we have just mentioned. Yet *paradoxically*, we know that the eigenstates of an electron in a harmonic potential are quantised and similarly a magnetic field generates well-defined Landau states in the ballistic transport regime *So what has happened to the quantum mechanics?*

The answer lies with the *boundary conditions* to the transport equations. The Wigner distribution cannot take on an arbitrary classical form; for example a delta-function in phase space would violate the uncertainty relations; indeed, the Wigner distribution, constructed directly from wavefunctions, has long been known to satisfy certain inequalities and sum rules which proscribe its behaviour. Only initial distributions which satisfy stringent quantum conditions can be accepted as initial states which then evolve classically in time under quadratic potentials and uniform magnetic fields. We could construct those initial states directly by writing the initial density matrix as a bi-linear function of an appropriate complete set of wavefunctions (eg Harmonic oscillator or Landau states or edge states) and then using the Wigner-Weyl transformation to determine the initial Wigner distribution. Alternatively, we can work entirely with Wigner distributions in an extended formalism which provides the theory with a complete set of stationary Wigner functions. Before elaborating on this point let us summarise what we have learned so far:

(i) for a parabolic kinetic energy, general scalar potential fields and uniform magnetic fields the equation of motion of the Wigner distribution is exactly classical.

(ii) In these circumstances an initial (injected) distribution of carriers will propagate classically (ie via area preserving transformations in phase space : therefore admitting trajectory descriptions exactly). The underlying motion preserves the occupacy of the initial quantum states.

(iii) The initial distribution must be compatible with quantum mechanics.

(iv) The initial distribution for a harmonic potential or for a uniform magnetic field must in particular be a bi-linear function of a superposition of Harmonic oscillator states or Landau states.

It follows that the underlying quantised nature of the states will only be manifest if either : (a) transitions are induced between those states; in particular inelastic transitions most certainly reveal the discrete structure; or (b), the quantisation is revealed through the dependence of the current on the initial distribution, eg via a supply function (this is precisely why we could "derive" the conductance quantisation formulae by semi-classical arguments). Only if the Hamiltonian departs from the conditions of uniform magnetic fields and quadratic potential fields do we pick up the effects of electron diffraction and interference phenomena. As we shall see later even this situation does not necessarily destroy conclusions on say the conductance quantisation phenomena.

4.6. Stationary Wigner distributions

The initial state for an evolving quantum distribution must be a linear superposition of *stationary* Wigner functions made up from the eigenstates of the appropriate Hamiltonian [11]. The Wigner equation for ballistic motion in a quadratic potential propagates an initial distribution classically. This shows that the Wigner equation of motion gives an incomplete picture of the physics- it cannot describe stationary states. A similar problem arises with the density matrix equation

$$ih\,\partial_t \rho = [H,\rho] \tag{38}$$

whose Wigner-Wey transform generates the Wigner equation of motion directly; it is not completely equivalent to the Schrödinger equation as can be seen by by trying to set up an eigenvalue equation. In fact the stationary states have to be obtained from an adjunct equation for the stationary density matrix [12],

$$\varepsilon\rho = \{H,\rho\}/2 \tag{39}$$

which involves the anti-commutator $\{\ ,\ \}$. The Wigner-Weyl transform of this equation yields

$$(\varepsilon_{mn}/2)f_{mn}(R,P) = [P^2/2m-(h^2/8m)\,\partial^2_R\,]f_{mn}(R,P)+$$

$$\int d^3r\int d^3P'\, e^{-iP'.r/h}\,\{V(R + r/2) + V(R-r/2)\}\, f_{mn}(R,P-P')/2(2\pi h)^3 \tag{40}$$

where Planck's constant is firmly re-instated. Here the $\varepsilon_{mn} = \varepsilon_m + \varepsilon_n$, where the ε_n are the usual eigenvalues of H. For m = n we obtain the usual stationary Wigner functions, they are real valued. The case m≠n gives complex functions which relate to the other eigenfunctions of the super-operator [H,]. The entire set of $f_{mn}(R,P)$ form a complete orthonormal set for all Wigner functions, stationary or otherwise. Any initial Wigner function should be projected onto this space if the correct boundary conditions for time-dependent transport are required. All the special sum rules of f such as $|f| \le 1/(h)^3$ are then taken care of.

5. Coupled-mode wave theory

The present interest in quantum ballistic electron transport in semiconductor structures has generated a requirement for accurate modelling of transport in single and multiply-connected spatially-extended electron waveguide structures which may be inhomogeneous. Target devices are for example, the Aharonov-Bohm ring device [13] and the squeezed-channel or throttle device [14, 10]; Until recently these have only been modelled as one-dimensional structures (we may quote the elegant work by Buttiker[15], Gefen, Imry and Azbel[16] as examples) with a few numerical calculations based on the 2D Schrödinger equation [3,4, 18]. Frohne and Datta[17] have described an approximate numerical technique based on wavefunction matching to calculate the scattering matrix for electron transfer between 2D channel regions with different confining potentials in the transverse direction. Kirczenow[19] has described a wavefunction matching technique for the abrupt uniform channel case to model the quantised resistance. A general formalism has been developed in [3,4, 20, 21] and has interesting analogies with electromagnetic tapered waveguide theory [22]. In the following we develop the general wave-mechanical formalism for an arbitrary shaped electron waveguide following the ideas in [3,4].

5.1. Basic formulation for curved guides

Let the wavefunction $\Psi(r,t) = \langle r|\Psi(t)\rangle$ at a location r = (x,y,z) = (R,z) be represented as a superposition of eigenstates $\varphi_m[a_\mu(z), R]$ of H_c corresponding to the confinement potential in the 2-D region formed by the x-y plane passing through the point z. Referring to the generic waveguide illustrated in figure 5 we use different transverse states at different cross-sectional slices in the guide. Introducing an amplitude factor $\psi_m(z,t)$ we have

$$\Psi = \sum_m \psi_m(z,t)\,\varphi_m[a_\mu(z), R] \tag{41}$$

The set $\{\varphi_m[a_\mu(z), R];\ \mu=1,2,...N\}$ is a complete orthonormal set parameterised by N parameters $a_\mu(z)$ which will generally depend on the z-spatial coordinate and which relate to the geometry of the confinement potential V_c (for example, we might have a tapered 3D channel, channel radius $a_1=a(z)$ with perfectly reflecting walls). By substituting expression (2) into the time-dependent Schrödinger equation and projecting onto the z-domain by forming the partial scalar product on the complete set $\{\varphi_m[a_\mu(z), R]\}$ we find that the amplitudes $\psi_m(z)$ obey coupled 1-D equations of the form:

$$ih\partial\psi_m(z)/\partial t = [(-h^2/2m^*)\partial^2/\partial z^2 + E_m(z)]\psi_m + \sum_n \{A_{mn}(z)\,\psi_n + B_{mn}(z)\,\partial\psi_n/\partial z\} \tag{42}$$

Here, $E_m(z)$ is the eigenvalue of state $\varphi_m[a_\mu(z), R]$. The coupling coefficients $A_{mn}(z)$, $B_{mn}(z)$ are known functions of : the local guide eigenstates φ_m, the guide parameters a_μ and their derivatives and the partial matrix elements $V_{mn}(z)$ of the internal potential V(R,z).

$$A_{mn} = \int d^2R\ \varphi_m\{\ V(R,y)\varphi_n + (-h^2/2m^*)\{\sum_{\mu,\nu}\partial^2\varphi_n/\partial a_\mu\partial a_\nu.[\partial a_\mu/\partial z][\partial a_\nu/\partial z] + \sum_\mu \partial\varphi_n/\partial a_\mu.\partial^2 a_\mu/\partial z^2\ \}\}$$

$$= V_{mn}(z) + \sum_{\mu,\nu} W^{mn}_{\mu\nu}(z) \qquad (43)$$

$$B_{mn} = \int d^2R\ \varphi_m\{ (-\hbar^2/2m^*)^2 \Sigma \partial\varphi_n/\partial a_\mu [\partial a_\mu/\partial z]\} = \sum_\mu U^{mn}_\mu(z) \qquad (44)$$

$$V_{mn}(z) = \int d^2R\ \varphi_m\{V(R,z)\}\varphi_n \quad ; \quad D_{mn} = A_{mn} - V_{mn} \qquad (45)$$

If the confinement parameters a_μ are time-dependent we should include an additional term

$$V^t_{mn} = i\hbar \int d^2R\ \varphi_m \Sigma \partial\varphi_n/\partial a_\mu [\partial a_\mu/\partial t] \qquad (46)$$

in the expression for A_{mn} E_m, A_{mn} and B_{mn} then become functions of time.

Equations (42) are a set of coupled one-dimensional equations for the mode amplitudes In solving the stationary-state version of these equations it is convenient separate the contributions from forward propagating, backward propagating and evanescent states by the ansatz

$$\psi_m(z) = \psi^{(+)}_m(z)\exp(ik_m z) + \psi^{(-)}_m(z)\exp(-ik_m z) \qquad (47)$$

where $k_m^2 = 2m^*(E-E_m)/\hbar^2$, E is the total energy and the case $k_m^2 < 0$ describes evanescent states

5.2. Adiabatic limit

In the case of the extreme quantum limit (lowest mode occupied) or the case of weak inter-mode scattering (slowly varying parameters)we obtain the *adiabatic approximation* in which a carrier will remain in the same mode n throughout the channel. Then expression (42) assumes an interesting structure if we group the coupling terms A, B into a manifestly hermitian Hamiltonian form using the operator $p_z = -i\hbar\partial/\partial z$ and preserving only diagonal terms m=n:

$$i\hbar\partial\psi_m(z,t)/\partial t = \{[p_z + \langle m|p_z|m\rangle]^2/2m^* - [\langle m|p_z|m\rangle]^2/2m^* - [\langle p_z m|p_z m\rangle]/2m^* + E_m(z) + V_{mm}(z)\}\psi_m(z,t) \qquad (48)$$

$$\langle m|p_z|m\rangle = \int d^2R\ \varphi_m\{ (-i\hbar)\Sigma\partial\varphi_m/\partial a_\mu [\partial a_\mu/\partial z]\} = (-i\hbar/2)\Sigma[\partial a_\mu/\partial z]\partial/\partial a_\mu \int d^2R\ \varphi_m\varphi_m = 0$$

$$\langle p_z m|p_z m\rangle = \int d^2R\ (-i\hbar)2\Sigma\Sigma\partial\varphi_m/\partial a_\nu . \partial\varphi_m/\partial a_\mu . [\partial a_\mu/\partial z][\partial a_\nu/\partial z]\} = -T_{mm}$$

We first observe that the diagonal momentum dependent coupling terms ($\sim \partial\psi_m(z,t)/\partial z$) vanish identically in this limit ($\langle m|p_z|n\rangle$ is generally non-zero), secondly, the term T_{mm}, which is positive, adds to the zero-point energy of the mode provided some of the gradients $[\partial a_\mu/\partial z]$ are non-zero:

$$i\hbar\partial\psi_m(z,t)/\partial t = \{[p_z]^2/2m^* + T_{mm}(z) + E_m(z) + V_{mm}(z)\}\psi_m(z,t) \qquad (49)$$

The terms T_{mm} and E_m play the role of one-dimensional quasi-potential fields whose gradients correspond to the effective force on a carrier due to the size quantization and the longitudinal interaction with the confinement potential.

A straightforward Wigner transform on the wavefunction $\psi_m(z,t)$ (following Barker, 1989a) allows us to construct a relatively simple equation of motion for the Wigner distribution function $f_m(z,p_z,t)$: a non-local version of the collisionless Boltzmann equation:

$$[\partial/\partial t + (p_z/m^*)\partial/\partial z]\, f_m(z,p_z,t) + \int dP'\, F_{eff}(z,P',t) \cdot \partial f_m(z,p_z-P',t)/\partial P = 0 \qquad (50)$$

$$F_m(z,r,t) = \{V_m(z + r/2,t) - V_m(z-r/2,t)\}/r; \qquad V_m = T_{mm}(z) + E_m(z) + V_{mm}(z)$$

$$F_m(z,P,t) = \int dr\ e^{-iP.r/\hbar}\, F_m(z,r,t)/(2\pi\hbar)$$

In the case of slowly-varying applied and quasi-fields (ie very little quantum reflection, resonances or tunnelling) we obtain the local collisionless Boltzmann kinetic equation:

$$[\partial/\partial t + (p_z/m^*).\partial/\partial z]\, f_m(z,p_z,t) + F_m(z,t).\partial f_m(z,p_z,t)/\partial p_z = 0;\ F_m(z,t) \to -\partial/\partial z V_m(z) \qquad (51)$$

This equation is exact for the case that F_{eff} is a gutter potential.

5.3. Coupled kinetic equations

It is relatively easy to re-introduce the inter-mode coupling if the previous conditions for the local kinetic equation are met. For weak inter-mode coupling the standard projection super-operator calculus may be used in analogy with the case of longitudinal magnetotransport [23]to obtain an inter-mode scattering integral within the Golden Rule approximation for the scattering rates R(mn,z). The latter are straightforwardly related to matrix elements compounded from the coefficients A_{mn} and B_{mn} [3] and we find:

$$[\partial/\partial t + (p_z/m^*).\partial/\partial z] f_m(z,p_z,t) + F_m(z,t).\partial f_m(z,p_z,t)/\partial p_z = -\Sigma\{f_m(z,p_z,t)R(mn,z) - f_n(z,p_z,t)R(nm,z)\} \qquad (52)$$

Interesting selection rules apply depending on the choice of confinement potential.

A completely wave-mechanical description of the general coupled-mode theory has been developed recently by Barker and Laughton ([24]to be published elsewhere).

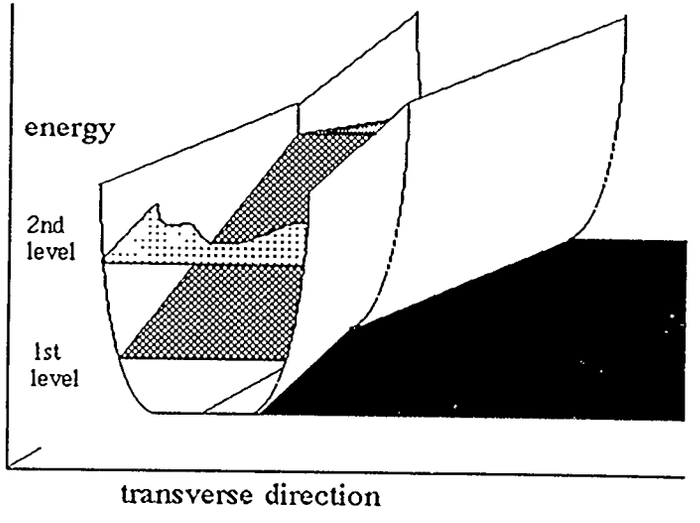


Figure 10 The transverse energies vary with position along the axis of a non-uniform guide.

5.4. Effective fields

The quasi-fields $F_n(z)$ may be engineered by profiling the walls of the quantum wire. Thus a linearly-tapered section of a hard-walled wire gives rise to an effective constant force field (or linear potential); a succession of abruptly-connected uniform wires of different widths corresponds to a succession of step potentials. Figures 3 and 11 illustrate typical potentials that can ideally be achieved for different lateral patterning. It would appear that a number of architectures which have been difficult to achieve with vertical transport through heterostructures, such as the staircase potentials, are realisable with laterally patterned nanostructures. There is however the question of how strong we can make the quasi-fields or equivalently how energetic can we make the quasi-potential barriers. A broad measure of scale is given by the zero point energy, which for a 20 nm width hard-walled wire in GaAs gives $E_1 \sim 14$ meV; similarly for the parabolic potentials described by Kumar, Laux and Stern [2] we have $\hbar\omega \sim 2.6$ meV corresponding to a squeezed gate channel of width 0.4 μm and gate voltage -1.0 V. Figure 10 sketches the variation of the zero point energy with distance along a pinched quantum wire. It is clear therefore that hot electron effects induced by such fields will only be manifest at low temperatures. It is tantalising to note that with a suitable squeezed gate geometry on the limiting scales of 10nm defined channel widths it should be possible to achieve electron energies equivalent to 1000 K.

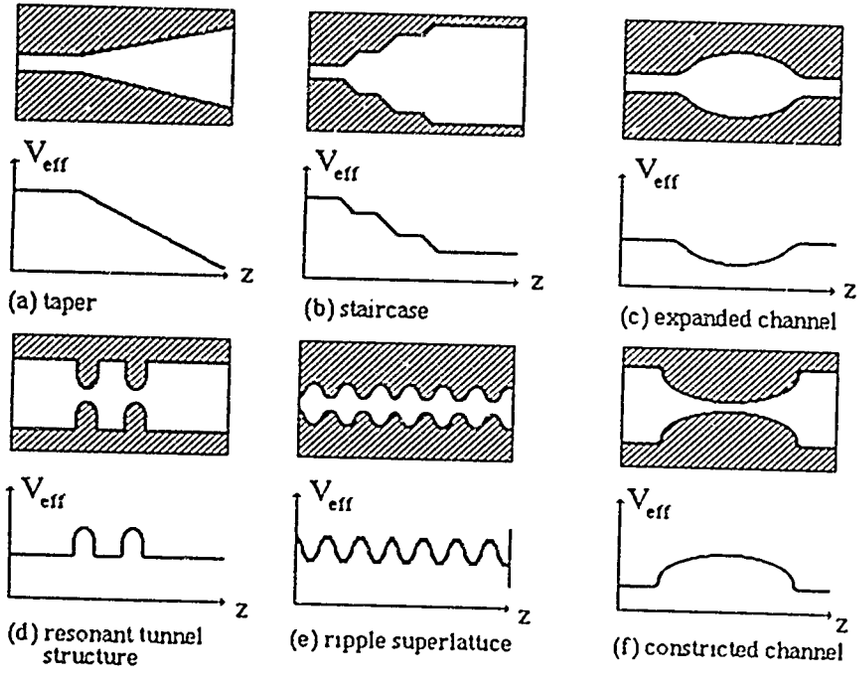


Figure 11 Electron waveguide squeezed gate geometries showing the effective potential for adiabatic motion in one of the transverse sub-bands- ideal confinement potential.

5.5. Numerical methods

A number of computational techniques have been developed recently [3,4, 20, 21, 24] to address these issues and are discussed in detail elsewhere. The most physical [24] uses the a coupled-mode theory of transport in electronwaveguides which is suitable for describing the influence of realistic confinement potentials including the effects of fluctuations. Numerical modelling here is complicated by the need to include the evanescent and travelling wave modes the number of which vary along the channels in extreme cases such as a ripple superlattice or an Aharonov-Bohm ring device. The standard Thomson-Haskell propagator matrix methods are unstable here and a number of methods to remove the problem by decoupling algorithms have been discovered recently[24].This method may be complemented by direct finite-difference solutions to the 2D time dependent effective mass Schrödinger equation for the structure using a modified ADI algorithm [3,4,20]. Finite element methods[18] have been also been reported very recently for studying resistance quantisation in pinched channels. A further method utilises a network of one-dimensional lines to span the appropriate waveguide geometry and this method is a natural extension of traditional 1-D theory [21,25]. Baranger and co-workers have used standard Green function propagator methods to numerically study electron waveguides .

6. Interference phenomena and the Aharonov-Bohm effect

In multiply-connected geometries such as a waveguide ring, there have been extensive observations of Aharonov-Bohm effects due to alteration of the phase-difference between a split coherent electron flow (a detailed theory is described in [3,4, 20]).

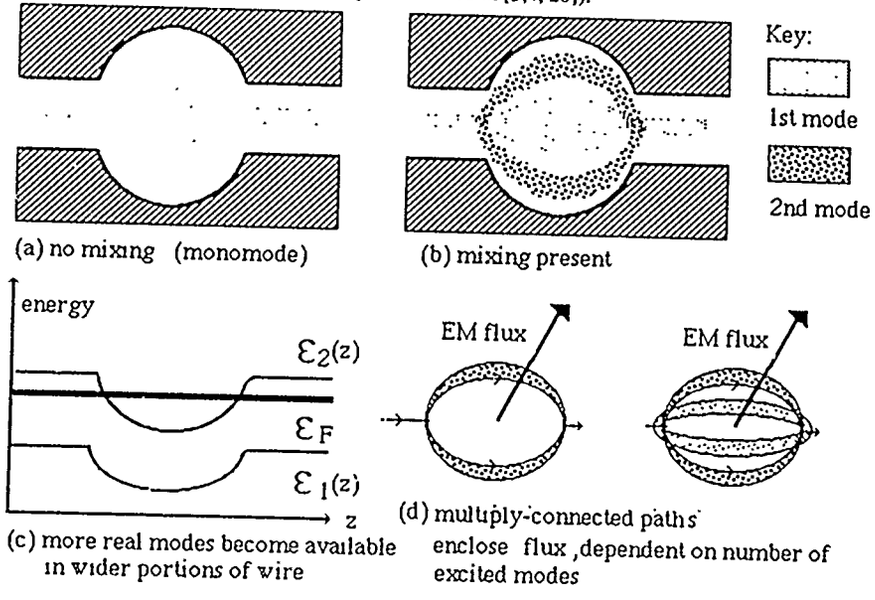


Figure 12 Mode mixing can lead to multiply-connected "electron trajectories" which are manifest in phenomena such as the Aharonov-Bohm effect.

The Aharonov-Bohm effects, like most of the interference based phenomena have small amplitudes of modulation (typically 20-30%) and are observable only at very low temperatures The amplitude dependence can be partly explained by the difficulty in achieving monomode conduction in the waveguide [3, 20]. The general theory of inhomogeneous electron waveguides [3, 24] shows that intermode scattering by the curved boundaries will occur easily except at very low energies. Aharonov-Bohm effects are also weakly present in extended waveguides due to multiply-connected electron paths arising from mode-mixing [4] (see figure 12). The effect of inhomogeneous boundaries may be split into two parts: (i) the variation of sub-band zero point energy with distance along the guide gives rise to an effective potential on the guided motion (see figure 10); (ii) the spatial derivatives of guide boundary parameters leads to inter-sub band scattering. For 2D models conformal mapping can be used to derive effective potentials; a bend in a waveguide leads to quasi-bound states for example

7. Influence of the fluctuation potential

All the mentioned properties are features of electron waveguides describable by smooth confinement potentials (eg harmonic cross section). However, it can be shown that the confinement potential which is assumed to derive ultimately from the mean field of the donors and various image charges must be supplemented by a fluctuation potential which arises because of the discrete spatial distributions of donors [5,6]. Depending on the amount of screening, the rms fluctuation potential can be as large as 30 meV.The effect is to distort the electron gas into a random structure which is illustrated in figure 8 (b). Detailed calculations [26,27] for split-gate waveguides show that classically random pinch-off generally occurs well before a scale at which monomode conduction would be established quantum mechanically - at typically 60 - 80 nm for 1 μm long wires with 0.4 μm split gates. The basic effect is sketched in Figure 13.

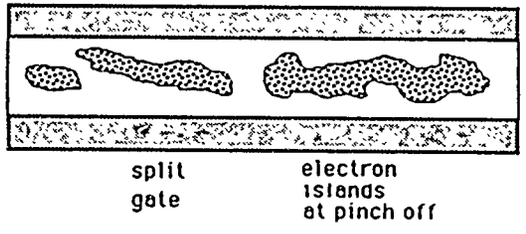


Figure 13 The quasi- 1 D electron gas breaks into disconnected puddles due to the fluctuation potential in a squeezed gate wire.

These semi-classical calculations neglect tunnelling and a quantum theory of the image force remains to be developed so they are probably pessimistic results; but they do indicate a serious problem for waveguide design.

Very recently, the full coupled mode theory has been applied to the problem of modal transport through a squeezed gate quantum point contact using realistic potential profiles in the presence of the fluctuation potential [24,28]. The fluctuation potential was computed self-consistently from the semi-classical Thomas-Fermi approximation. The calculation includes the contribution from the donors, assumed to be fully ionised and distributed at random in a δ-doped layer (following the earlier study by Nixon, Davies and Baranger [27] which used recursive Green function techniques). The confinement and fluctuation potentials are very device specific because of the random nature of the donor layer.

Both the curvature of the confinement potential and the inhomogeneities introduced by the fluctuation potential lead to mode coupling. The detailed transport is quite complex and includes in-

channel resonances and indirect processes which lead to strong back-scattering.The latter cannot be treated by Born approximation arguments for the mobility where the assumption of independent scatterers leads to contributions to back-scattering from direct processes only. The Born approximation thus seriously over-estimates the mobility. The *indirect* back-scattering is part of the breakdown of independent scattering which ultimately leads to exponential localisation in one dimensional systems.

For realistic device potentials the transport is found to be *non-adiabatic* and yet good quantised conductance is still predicted. This surprising feature is attributed to "compensated scattering" an idea originally due to Payne[29] but not thought to be significant in the region following the narrowest constriction of the device. Basically, flux scattering forward out of the lowest modes is compensated by scattering-in from the higher modes which are also occupied. The result is that the effective one-dimensionality of the transport is maintained and this, following the arguments of section 3, sustains the conductance quantisation effect: a one-dimensional channel has a maximum conductance of $2e^2/h$. The compensated scattering leads additionally to "mode scrambling" which reduces the coherence of the flux. This mechanism may be significant in reducing the amplitudes of interference proceses in ring structures.

8. Charging effects

Structures where the electrons of a two-dimensional electron gas are ideally confined to disconnected regions could be fabricated by the use of an appropriate gate geometry as sketched in figures 14, 15. Such local confinement also occurs due to the fluctuation potential (see figure 13). Transport between the conducting islands may take place by tunnelling but will be strongly influenced by the charging energy $E_c = e^2/2C$ associated with the transfer of a single electron where C is the *small* capacitance between the islands.

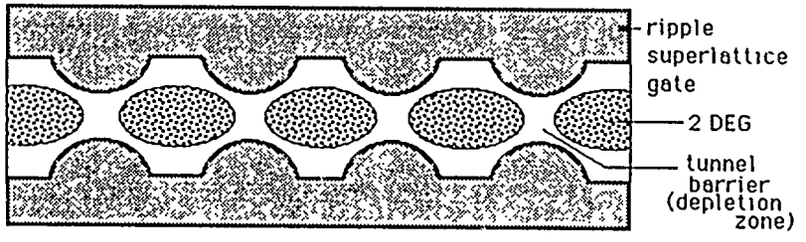


Figure 14 Charging effects are expected in the tunnelling of electrons between confining pockets in a ripple superlattice.

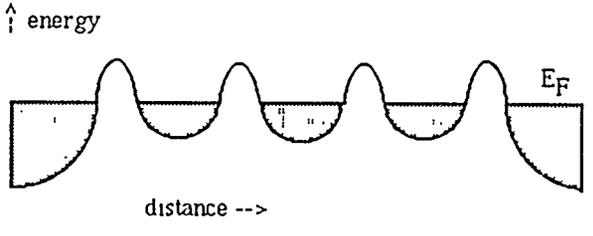


Figure 15 Effective potential for a short ripple super-lattice.

For low temperatures ($k_BT << E_c$) the *Coulomb blockade* occurs, an effect already revealed in observations of the of tunnelling of single electrons (normal materials) and Cooper pairs (superconducting materials) in very small metal-insulator tunnel junctions (areas ~ 0.1 x 0.1 μm) at low temperatures[30-33] . A related effect, the time-correlated tunnelling of single electrons in charge-soliton states is also anticipated.These charging effects require the existence of very small cacitances. Tpically a $(100\ nm)^2$ junction area gives a capacitance of the order of 10^{-15} F leading to $e^2/2Ck_B \sim 1$ K. Evidently the very much smaller areas possible in electron waveguide structures raises the possibility of observing and using charging effects at very much higher temperatures. Additionally the existence of extended charge-soliton states in arrays of small tunnel junctions provides a possible mechanism to overcome some of the limitations imposed on devices by quantum fluctuations. In the structure sketched in figures 14, 15 (Barker and Selbmann, 1989)both the tunnelling probability and the charging energy could be controlled by the variation of the confinement potential with gate voltage. It is emphasised that the correlated nature of the transport mechanism is determined by charging effects only and thus nearly independent from electron scattering and potential fluctuations over a wide range of parameters. A detailed theory of charging effects remains to be developed for electron waveguide structures, particularly in the few electron limit.

9. The few electron limit - granular electronics

With certain important exceptions the new phenomena discovered and exploited in electron waveguide structures have been observed via measurements on large ensembles of electrons. The granularity of electric current manifest in the discreteness of the electron has largely been hidden. Charge quantisation has of course been directly observed in the conductance quantisation of quantum wires (via the sub-band filling factors), the universal conductance fluctuations in disordered systems and the Aharonov-Bohm effects in multiply-connected materials (a consequence of gauge invariance). Significant effects due to individual electrons were first reliably reported in the measurements of gated quantum wire structures in silicon structures[34] in which the trapping and de-trapping effects of single electrons could be related to the characteristic telegraph noise signal in the conductance of the wire (the single electron events essentially bottleneck the flow of a macroscopic current). A similar result is found in the current-voltage characteristics of scanning tunnelling microscopy of defects on the surface of semiconductors.

Many existing electron waveguide structures, quantum wires, rings, quantum dots, quantum point contacts are already close to the regime where very few electrons occupy the active region of the structure. Important challenges arise here for both instrumentation and device applications. Considerable experimental and theoretical work will be required to understand and control electron waveguides particularly close to the few electron limit. For example, it is not clear what limits observed interference phenomena to very low temperatures, nor what is the nature and origin of de-phasing processes. The coupling of coherent electrons to the electrical environment is not understood and indeed brings about problems of interpretation in quantum measurement theory. These issues are discussed in more detail in my lecture on granular nanoelectronics.

10.References

1. Laux, S and Stern, F Appl.Phys.Lett. 49 91 (1986).
2. Kumar, A., Laux, S and Stern, F. Channel sensitivity to gate roughness in a split-gate GaAs-AlGaAs heterostructure-preprint (1989).
3. Barker, J. R. in *Physics and Fabrication of Nanostructures*, ed. M.A. Reed and W P Kirk (Academic Press:New York), 253 (1989).
4. Barker, J. R, Pepin, J, Finch, M and Laughton, M.Solid St. Electronics 32 1155 (1989).
5. Davies, J.H and Nixon, J.A. Phys.Rev B 39 3423 (1989).
6. Nixon, J.A, Davies, J.H and Barker, J.R. in *Physics and Fabrication of Nanostructures*, ed. M.A. Reed and W.P Kirk (Academic Press:New York), 123 (1989).
7. Landauer, R. Phil.Mag. 21 863 (1970).
8. Stone, A.D and Szafer, A IBM J. Res.Dev 32 384 (1988)

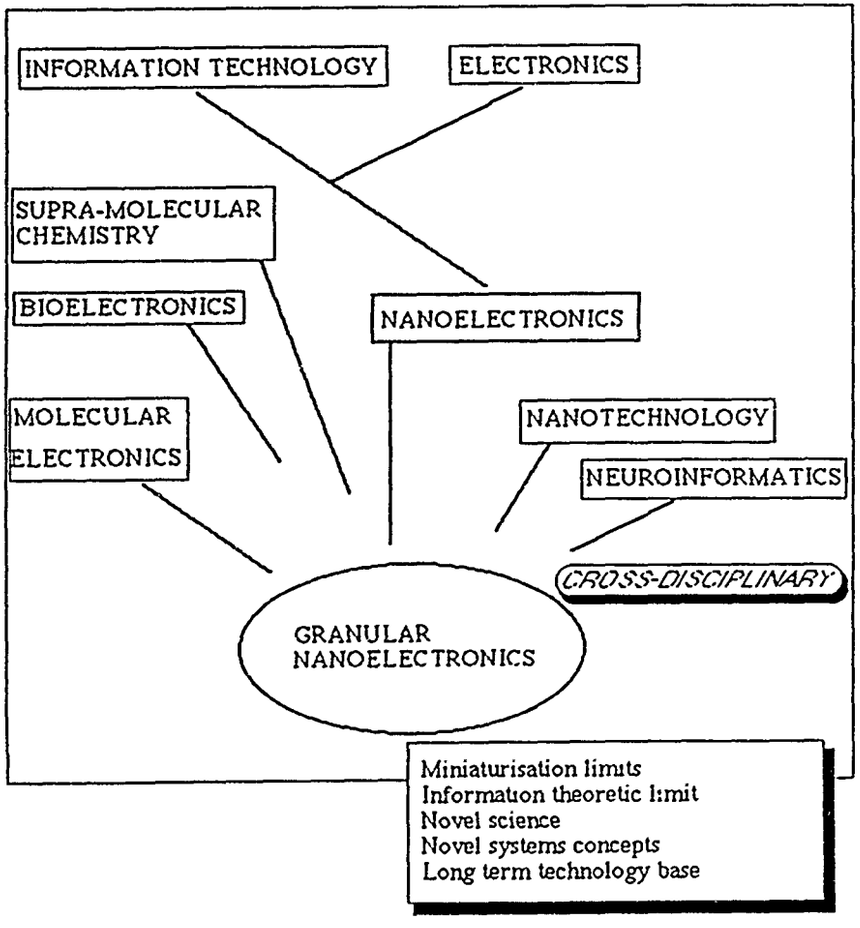
9. Barker, J.R. Symposium on New Phenomena in Mesoscopic Structures, Keauhou-Kona, Hawaii (1989) - unpublished
10. van Houten, H, Beenakker, C.J.W and van Wees, B.J. Quantum Point Contacts in a volume in Semiconductors and Semi-metals ed. M.A. Reed (Academic Press:New York) , (1990).
11. Barker, J.R. Chapter 13, Semiconductor Device Modelling, ed. C.M. Snowden, (Springer-Verlag:London), 207-226 (1989).
12. Carruthers, P and Zachareisen, F. Rev.Mod.Phys. 55 245 (1983).
13. Ford, C.J.B, Thornton, T.J, Newbury, R, Pepper, M, Ahmed, H, Foxon, C.T, Harris, J.J and Roberts, C. J.Phys C 21 L325 (1988).
14. Wharam, D.A, Thornton, T.J, Newbury, R, Pepper, M, Ahmed, H, Frost, J.E.F, Hasko, D.G, Peacock, D.C, Ritchie, D.Aand Jones, G.A.C. J.Phys.C 21 L209 (1988).
15. Buttiker, M. Phys.Rev.Letters 57 1761 (1986).
16. Gefen, Y, Imry, Y and Azbel, M.Ya Phys.Rev.Letters 52 129 (1984).
17. Frohne , R and Datta, S. J.Appl.Phys. 64 4086 (1988).
18. Lent, C, Sivaprakasam and Kikner, D.J. in *Physics a . Fabrication of Nanostructures*, ed. M.A. Reed and W.P Kirk (Academic Press:New Y rk), 279 (1989).
19. Kirczenow, G. Phys.Rev B 38 10958 (1988).
20. Finch, M. PhD Thesis University of Glasgow (1989).
21. Pepin, J. PhD Thesis, University of Glasgow, (1990).
22. Sporleder, F and Unger, H.G. Waveguide Tapers Transitions and Couplers (Peter Peregrinus Ltd, IEE London and New York) (1979).
23. Barker, J.R. Chapter 13, *Handbook of Semiconductors* vol 1, ed. W. Paul (North-Holland:New York) 617, (1982).
24. Barker, J.R and Laughton, M-to be published.
25. Barker, J.R and Pepin, J. to be published.
26. Nixon, J.A and Davies, J.H. preprint (1990).
27. Nixon, J.A, Davies, J.H and Baranger, H. preprint (1990).
28. Laughton, M, Barker, J.R, Nixon, J.A and Davies, J.H. submitted for publication (1990).
29. Payne, M.C. J.Phys.Condens.Matt. 1, 4939 (1989).
30. Likhaerev, K.K. IBM J. Res. Develop.32, 144 (1988)
31. Likhaerev, K.K, Bakhvalov,N.S, Kazacha,G.S, and Serdyukpva, S.I. IEEE Trans. Magn. 25, 1436, (1989)
32. Delsing, P, Likhaerev, K.K, Kuzmin, L.S, Claeson, T. Phys.Rev.Lett. 63, 1861, (1989).
33. Geerligs, L.J, Anderegg, V.F, Holweg, P.A.M, Mooij,J.E, Pothier, H, Esteve, D. Urbina, C. Devoret,M.H.pre-print Phys.Rev.Letters (1990).
34. Skocpol, W. in *Physics and Fabrication of Microstructures and Microdevices*, ed M.J. Kelly,(Springer-Verlag:London) 255 (1986).

Granular nanoelectronicss

J. R. Barker
Nanoelectronics Research Centre
Department of Electronics and Electrical Engineering
University of Glasgow
Glasgow G12 8QQ, Scotland, UK.

OUTLINE

1. Introduction

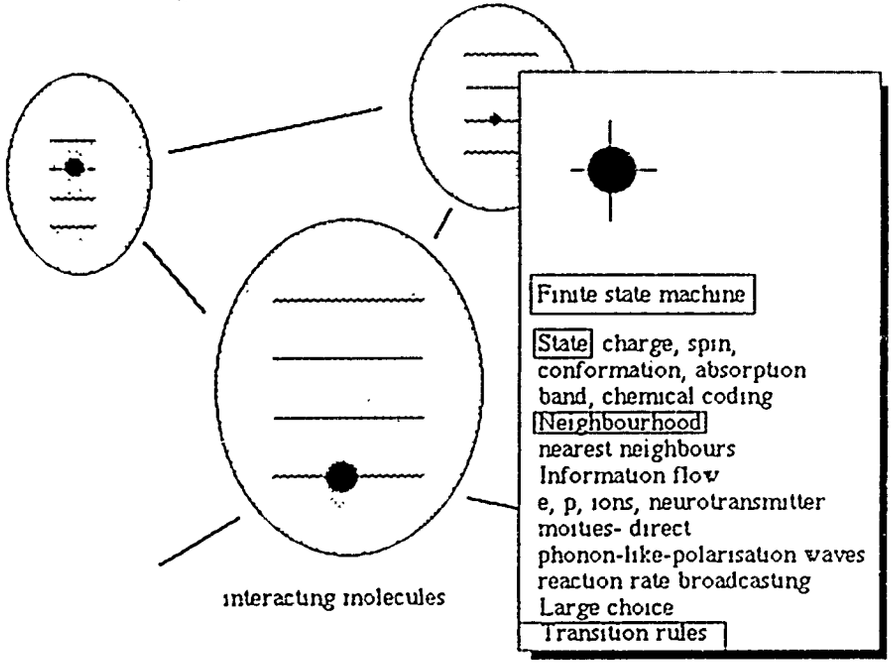


Interfacing to biological and molecular structures

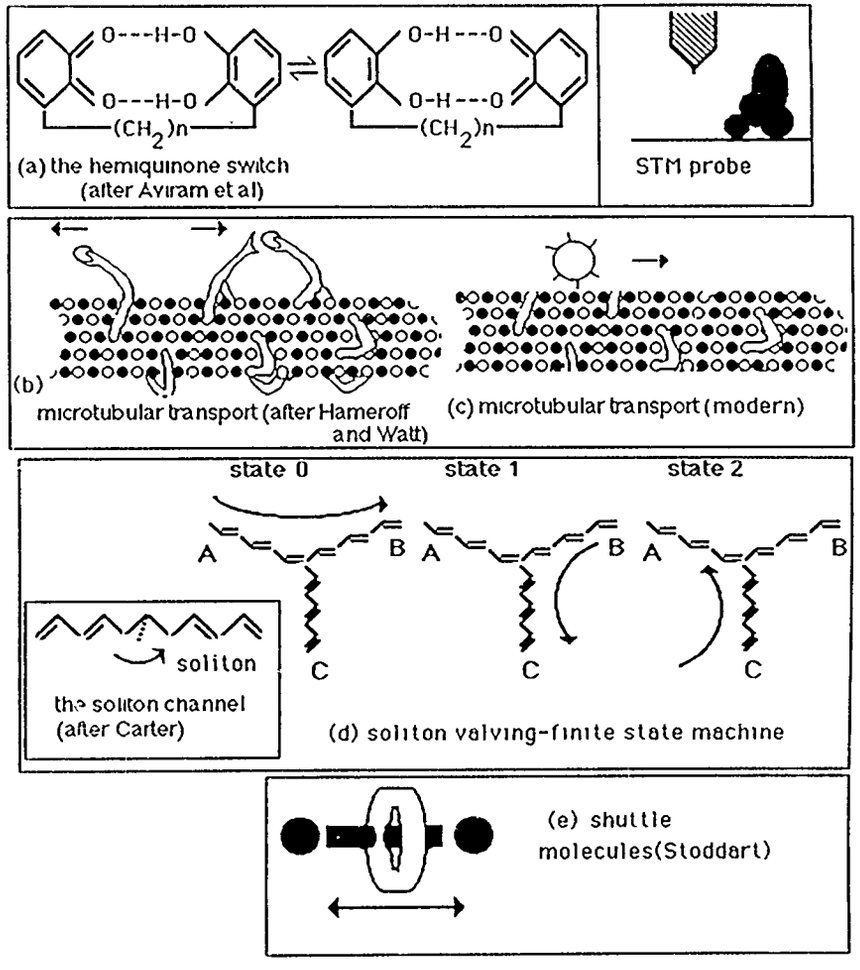
J R Barker, P C Connolly and G Moores
Molecular Electronics Group
Department of Electronics and Electrical Engineering
University of Glasgow
Glasgow G12 8QQ, Scotland, UK

OUTLINE OF TALK

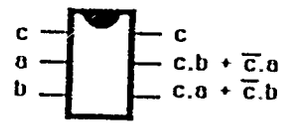
1. Introduction
2. Some lessons from biology
3. Molecular electronics
4. System targets



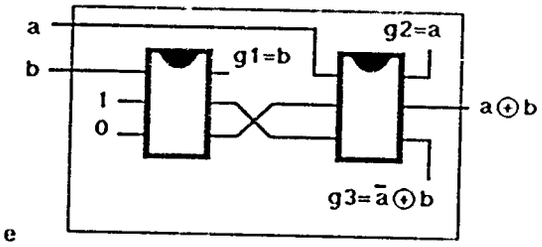
5. Candidates for the granular limit



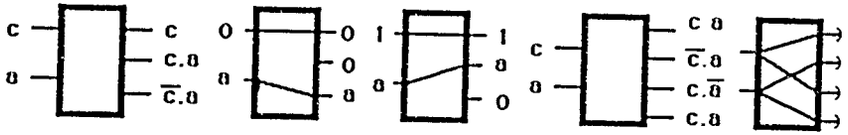
6. Physical representation of data and logic



(a)
a binary conservative logic gate-Fredkin gate



(b) a XOR gate from Fredkin gates
note the garbage lines g1-g3
and data lines 0, 1.

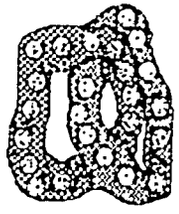


(c) a unary conservative logic gate-Priese gate

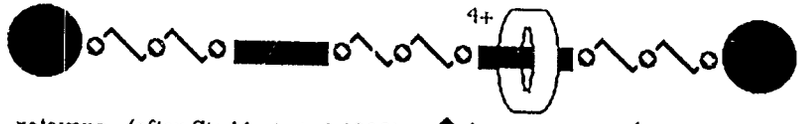
(d) Interaction gate

7. Criteria for workable molecular electronics

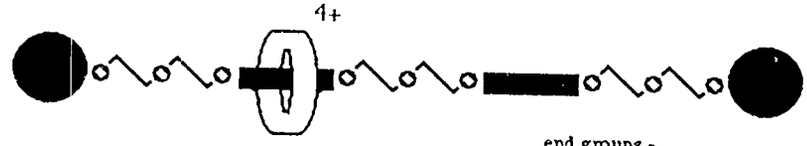
8. The quasi-mechanical option



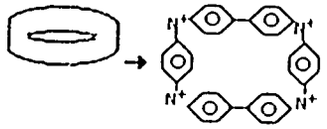
[2] catenane



rotaxane (after Stoddart et al 1990) ↑↓ 13 kcal mol^{-1}



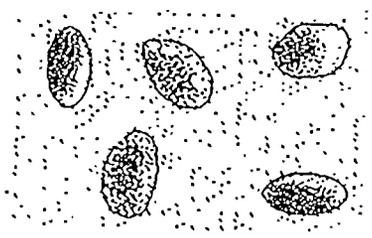
end groups -
large tri-isopropyl silyl units



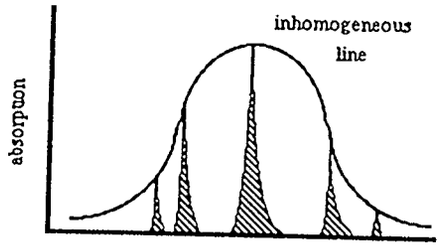
tetracationic cyclophane

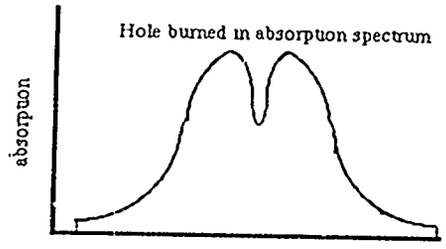
= 2 bi-pyridinium units
linked by para-phenylene
rings

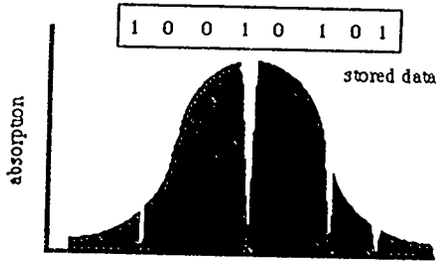
9. Addressing individual molecules
Scanning tunnelling probes
Spectral hole burning



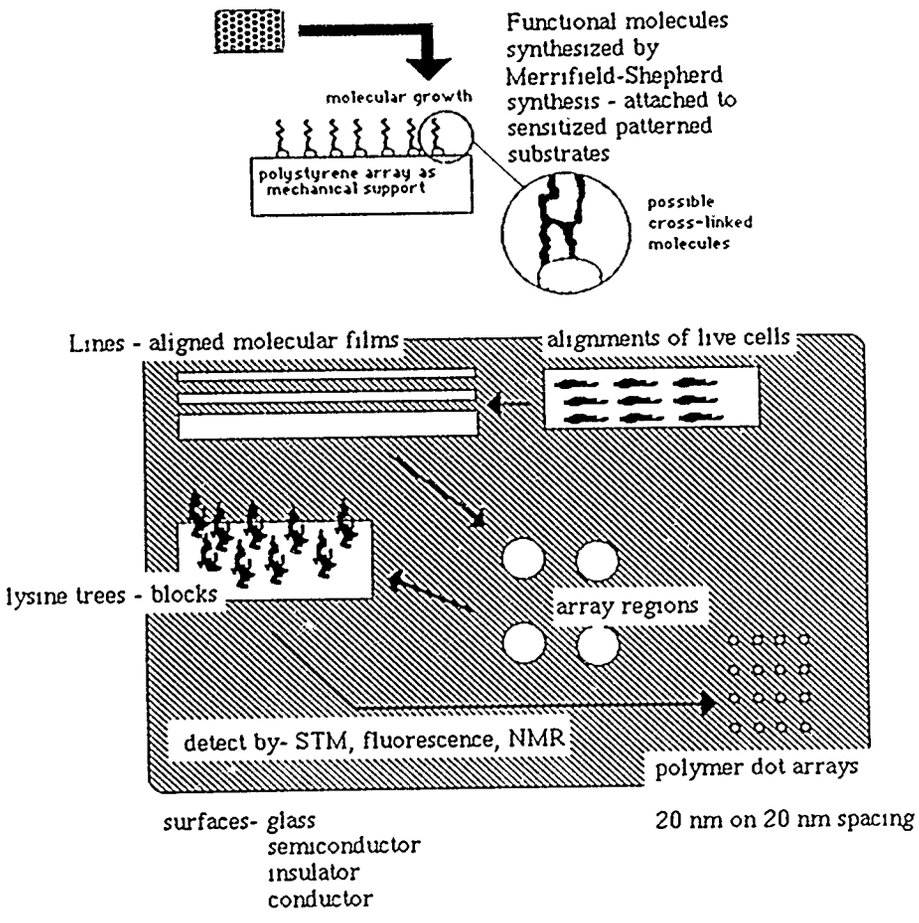
molecules in amorphous lattice



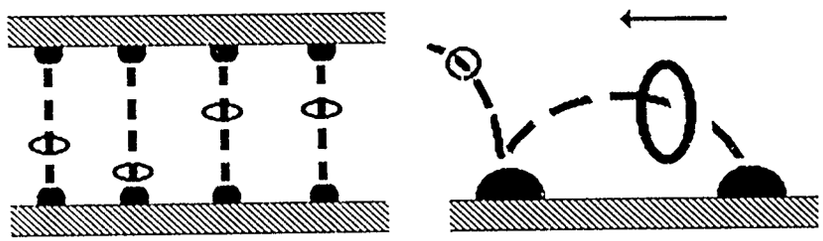




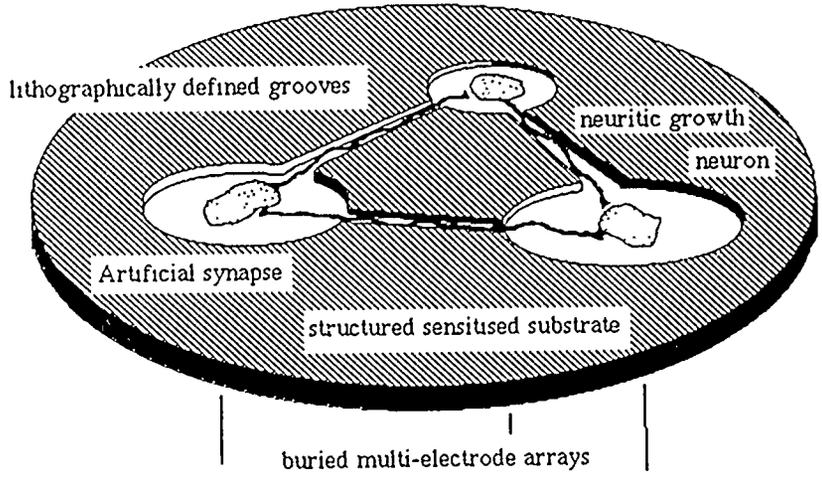
10. An experimental programme for molecular interfacing



11. The molecular abacus



12. Artificial biological neural networks



13. Interfacing to sub-cellular level bio-systems

14. Summary

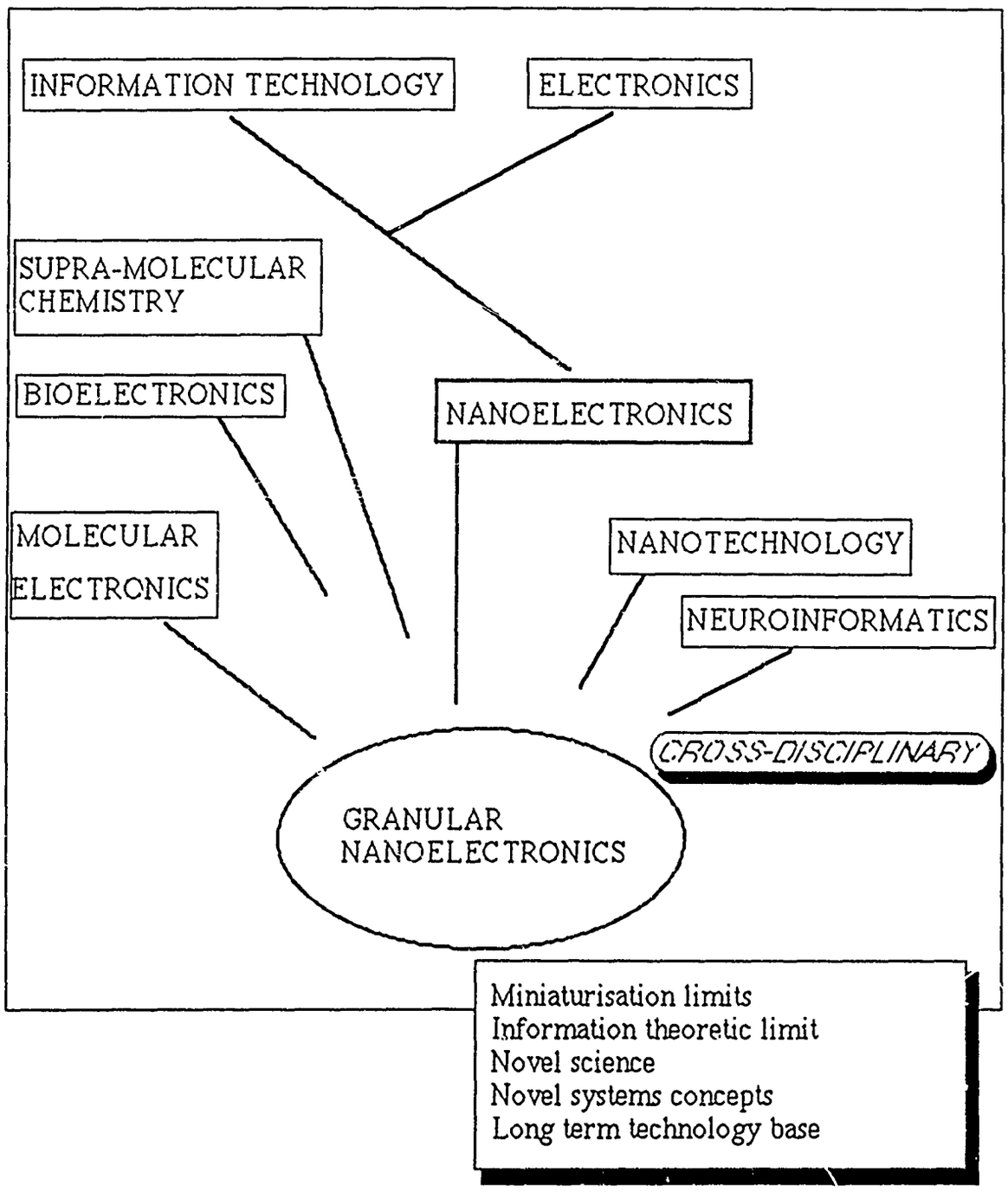
15. References

Granular nanoelectronicss

J. R. Barker
Nanoelectronics Research Centre
Department of Electronics and Electrical Engineering
University of Glasgow
Glasgow G12 8QQ, Scotland, UK.

OUTLINE

1. Introduction



ASI - Granular Nanoelectronics

MOLECULAR ELECTRONICS

J. P. Launay, Molecular Electronics Group,
CEMES / LOE, 29 rue Jeanne Marvig
31055 TOULOUSE CEDEX France

Abstract. A review is presented of "Molecular Devices" which have been already realized. This includes molecular wires, rectifiers, photodiodes, switches and bistable elements. Then some basic problems encountered in Molecular Electronics are evoked, in particular the need to associate a large number of components, the energy problem, and the place of Molecular Electronics in information processing systems. Several important research themes are listed, including nonlinear behaviours, logical functions at the molecular scale, the need to cascade components and to investigate the behaviour of a Quantum Computer. Finally some possible short term realizations are indicated.

INTRODUCTION

Molecular Electronics can be broadly defined as the processing of electric, magnetic, or optic signals with molecule based devices. After the discovery of conducting polymers in the 70's, the idea that molecules could be used to process signals was introduced by Aviram and Ratner (Aviram, 1974), who introduced the molecular rectifier concept, and later by the (late) F.L. Carter. From the beginming of the 80's, in a series of fascinating multidisciplinar workshops (Carter, 1982, 1987, 1988), he launched definitively the concept of Molecular Electronic Devices (MED). Of course a number of ideas evoked there were (and still are) highly speculative. But this had the immense advantage to trigger an increasing activity in chemical synthesis. There are now several example of "Molecular Devices" emerging from the laboratories. Most of them are based on simple and often naive analogies with actual Electronic Devices. But their fundamental study will certainly be highly instructive and full of surprises. The question now is to try to foresee the behaviour of large assemblies of such devices. This problem could have strong relations with the concept of Granular Nanoelectronics and both approaches (one from the above, the other from the bottom) are destinated to meet somewhere.

The present paper is organized as follows. In (I), we present the state of the art about "Molecular Devices" which are classified according to simple concepts coming from electrical analogy. They constitute thus the "spare part box" for future engineers. In part (II) we address the basic problems encountered in Molecular Electronics, which all come frome the extreme size reduction. Part (III) will suggest several research tracks which (we believe) could be crucial to transform Molecular Devices from simple Laboratory curiosities to actual useful systems. Most of these tracks involve long term research ; however in part (IV) it will be seen that short term

realizations and applications are nevertheless possible and will be extremely useful to reinforce credibility for this topic

1. A REVIEW OF EXISTING "MOLECULAR DEVICES"

A large number of molecular electronic devices have been suggested in the literature. Here we shall restrict to systems which have been actually synthesized, or which appear around the corner from a chemical point of view. We furthermore restrict on molecules based on electron transfer. Composents based on the propagation of other particles are conceivable but either the chemistry is extremely difficult (case of ions for instance) or the transfer process is not conservative (case of energy transfer for instance). In addition electron transfer is a universal process, already extensively used in conventional electronic devices. It can in principle survive to extreme size reduction and there is no theoretical impossibility to perform single molecule addressing by electrical means.

1.1. Wire

This is the most simple element the fuction of which is to transfer a bit of information (represented by an electron). However at the molecular scale it is not possible to study independently the behaviour of the different parts of a quantum system. Thus the "molecular wire" must encompass not only the "conducting" part, but also the terminal sites the function of which is interfacing with the outside world or with another component.

A first realization makes use of a binuclear coordination complex M-L-M (M = metal site, L = bridging ligand) in which the two M sites are able to exist in two oxidation states (e. g. Ru $(NH_3)_5^{2+/3+}$) and L is a π conjugated system (Woitellier et al 1989). When the complex is in the "mixed valence" state, i.e. one of the metal site is in oxidation state 2+, and the other in oxidation state 3+, there is a possibility of electron transfer between the two remote sites. This gives rise to a special electronic transition, the intervalence band, which usually occurs in the near IR. This band gives some information about the degree of through bond coupling between the two sites. Thus one has indirectly some information about the rate of the thermal process :

$$Ru^{2+} - L - Ru^{3+} \rightarrow Ru^{3+} - L - Ru^{2+}$$

Figure 1. A model of molecular wire

It should be noted that this process occurs more or less rapidly through any type of organic ligand. At this scale, there is no perfect insulator and the electron transfer occurs through tunneling without an actual oxidation or reduction of the ligand. To use semi-conductor language, $Ru^{2+/3+}$ sites can be considered as impurities for which the energy levels fall in the gap between the valence band and the conduction band. The electron transfer is due to an indirect coupling between these levels, without creation of charge carriers in the conduction or

valence bands. Recent results obtained in our group show that for the family of complexes bridged by α,ω bipyridylpolyenes, the electronic coupling decreases very slowly with distance (Joachim et al, 1990). The longest systems of this type have been described by J.M. Lehn and coworkers, for instance a "caroviologen" molecule in which 2 terminal pyridinium sites are linked by a chain of 11 conjugated double bonds (Arrhenius et al, 1986). The total length is ca 34 Å and this molecule can span a vesicle membrane and perform electron transfer between the outside and the inside (Blanchard-Desce, 1989). Recently, Effenberger has described an analogous system but dedicated to energy transfer (Effenberger et al, 1988).

1.2. Rectifier

In the previous case, the two terminal sites were identical. If now one introduces some asymmetry by associating a donor group (D) on one end and an acceptor group (A) on the other end, the electron transfer will be favored in one way and rectification properties will appear. It should be noted however that the term "molecular rectifier" or "molecular diode" is frequently used to designate two different types of molecules.

The first type corresponds to the original concept by Aviram and Ratner of a molecule with topology D-σ-A, where σ is a bridge containing only σ bonds, in order to limit the mixing of electronic levels of A and D. This molecule is destinated to be sandwiched between two metallic conductors so that it could be addressed by a permanent electron flux. Several molecules of this type have been prepared (Metzger et al, 1986) They were designed for incorporation in Langmuir Blodgett (LB) films with the objective of measuring the electrical resistance through a single monolayer placed between a Sn O_2 and a mercury electrode. This experimental set up suffers from two drawbacks : the frequent non perpendicular disposition of molecules in the film and the existence of microscopic holes in the films. Some attempts have been made to detect rectification with the tip of a Scanning Tunneling Microscope (STM) but they are unconclusive at the present time (Aviram et al 1988, 1989). Finally very recently, Sambles and coworkers have claimed that a LB monolayer of a potential molecular rectifier sandwiched between Pt and Mg electrodes exhibited rectification (Geddes et al, 1990). This seems to be the first successful report.

Figure 2. Examples of molecular rectifiers

The second type of molecular rectifier uses a π link between D and A sites. This provides a strong hyperpolarizability (β factor) and such molecules are currently used to devise organic materials for non linear optics (second harmonic generation). In these molecules, contrarly to the previous case, the electron transfer is virtual because the D^+-A^- configuration is never reached. The effect of the electric field is to alter the degree of mixing between the preponderant D-A configuration and the D^+-A^- one (Chemla and Zyss, 1987).

1.3. Photodiode.

Molecular photodiodes have been extensively studied in the last year with the prospect of achieving artificial photosynthesis. The general principe is to start from a photosensitizer (PS) linked to donor and acceptor groups. After excitation, the species PS * has an electron in a high energy orbital and a hole in a low lying orbital (cf electron hole creation in conduction and valence bands of a semi conductor). By a proper design of the molecule one can expect an electron transfer towards A and a hole transfer towards D, so that a vectorial charge separation is achieved :

$$D\text{-}PS\text{-}A \xrightarrow{h\nu} D\text{-}PS^*\text{-}A \longrightarrow D^+\text{-}PS\text{-}A^-$$

Thus with PS = porphyrin, A = quinone and D = carotene (see Figure 3), the exited state has a lifetime of 3 µsec (Moore, Gust et al 1984). With a more complex system, i.e. D-PS-PS-A-A, it has been possible to reach a lifetime of 340 µsec (Gust, Moore, 1990), which means that the back reaction is strongly slowed. Thus one can hope to be able to use the stored chemical energy in later steps.

Figure 3. The carotene-porphyrin-quinone triad

But for electronic applications, it is necessary to incorporate these molecules in organized structures so that a macroscopic current could be detected. Photocurrents have been observed with such molecules embedded in lipid bilayers (Seta et al, 1985). Another interesting system has been described with A = viologen, D = ferrocene, PS = pyrene (Fujihira et al, 1985). A

photocurrent was also observed with this molecule deposited as a L B monolayer on a semitransparent electrode and in contact with an electrolyte.

1.4. Switch.

If we use as bridging ligand a molecule with some "structural mobility", it is conceivable to control the ligand geometry by some external perturbation, and thus to control the electron transfer process. The simplest idea is to use a photochemical reaction. Thus several donor-acceptor molecules exhibit the so called TICT effect (Twisted Internal Charge Transfer) in which upon excitation there is a twisting motion around an essentially single bond with decoupling of the π systems (Grabowski et al, 1979). Thus the electronic interaction is "shut off" but only during the excited state lifetime. A bridging TICT molecule has been prepared (Launay et al, 1989). However, it should certainly be more useful to have a bistable switch by combining electron transfer ability and photoisomerization. But the synthetic work appears much more difficult.

Another possibility to build a molecular switch is to use a complexation / decomplexation reaction to change the conformation of a flexible ligand. Such a system has been recently obtained and is presently under study in the author's laboratory (Gourdon et al, 1990). Finally some protonation / deprotonation reactions could be used to alter the electronic structure of the bridging ligand (Launay et al, 1990).

1.5. Bistable element.

Bistable elements are already avaible if we consider photochromic molecules. Efficient systems have been described such as the "fulgides" (Heller, 1983) and they allow repeated cycling between two forms with different absorption spectra. The absorption spectra having very few overlap, it is possible to convert quantitatively one form into the other and to come back with a different wavelength (Figure 4). A special case of photochromic systems is represented by transition metal complexes exhibiting the spin crossover phenomenon. In some case they can also be switched between two states with different optical and magnetic properties (Hauser, 1986). This constitutes the LIESST effect (Light Induced Excited Spin State Trapping) which is formally analogous to photochromism with the added dimension of magnetic effects.

Figure 4. A bistable molecule of the"fulgide" family (Heller, 1983)

1.6. Others.

Many other molecular devices have been proposed but have not been synthesized. One can cite soliton devices (Carter, F. L., 1987), a memory shift register (Hopfield et al, 1988), a logical gate (Aviram, 1988) A special mention should be made about molecules with easily accessible excited states, because they constitute potential systems to realize logical functions (see below).

2. SOME BASIC PROBLEMS IN MOLECULAR ELECTRONICS

Although the molecular device concept has triggered a noticeable research effort, it presently does not imply a revolutionary kind of experiments. Rather the synthesis and present characterization of these devices is made by standard chemical and physico-chemical methods. But for really using these molecules, it is necessary to go beyond and consider several problems.

2.1. Testing components or associating components ?

The will to devise truly unimolecular devices has led to sophisticated methods to perform molecular addressing, i.e to be able to interrogate a single molecule. This technological "tour de force" is presently attacked by several means. Thus an optical method to detect a single molecule has been described recently by Moerner (Moerner et al, 1989) and also by Orrit (Bernard and Orrit, 1990). Electrical addressing is already realized in STM experiments when imaging single molecules is achieved (Quate et al, 1989). Another electrical approach, not yet realized, is based on a combination of advanced nanolithographic techniques and chemical grafting.

Although these ingenious methods will help the characterization of Molecular Devices and will pave the way for Intramolecular Physics, they cannot be considered as the ultimate goal in Molecular Electronics. The true challenge for building an information processing system is in fact to associate a large number of components and to predict the behaviour of a large population. This aspect has been ignored until now because it raises formidable difficulties. The main one comes from the lack of a measurable intermediate quantity which would be the output of a device and at the same time the input of another one. In traditional electronics, this quantity is either the current or the potential. In Molecular Electronics the notion of electrical potential is not immediately apparent and one has to rely on concepts such as "quasi chemical potential" (Joachim, 1990, Buttiker, 1989). Another problem is to achieve the directionality in communication, i.e that the output of device 1 is the input of device 2 and not the reverse. Finally, since the very process of measurement introduces a quantum perturbation, there is no simple step-by-step procedure to build a molecular circuit. One has to concieve the overall system with macroscopic input and output, and give up the idea of measuring intermediate quantities without completely modifying the system's behaviour.

2.2. Near or far from equilibrium (The energy source problem) ?

Most work devoted to electron transfer has been performed with low driving forces, i.e near to equilibrium. This is particulary true for studies on mixed valence complexes (Woitellier et al, 1989), and for the case of model switches embedded in a conducting chain (Sautet and Joachim, 1988). As a result there is few (if any) energy dissipation in the device itself, but rather in the interfacing sites or the electron reservoirs.

Theoretically this is not rhedibitory for an information processing system, because it has been shown that a computation can be performed in principle without the dissipation of energy (Bennett and Landauer, 1985). But it appears very difficult to put this general result in practice and to build an actual computing system (Keyes, 1988). As all information processing systems known to date proceed with energy dissipation, it is clear that we lack fundamental studies on the behaviour of molecular devices in the high driving force regime. Such systems would constitute simple 2-terminal devices, but their study is not a simple task because, if we consider electrically addressed molecules, a high driving force would mean an extremely high electric field, several orders of magnitude greater than usual values.

As far as 3-terminal devices are concerned, the problem is to control a large flow of energy by a small one in order to get amplification. This aspect has not been considered until now, because in most cases the energy source is not clearly defined in the prototypes of molecular devices. In particular, it is likely that the prototypes of switches evoked in 1.4 do not realize amplification .

2.3 Where could Molecular Electronics compete in information processing systems ?

Once elaborate molecules playing the role of Molecular Devices have been obtained, it is tempting to imagine how to assemble them in order to realize complex functions. This line of thought parallels the classical development of silicium electronics from simple components (transistors) to more complex ones (adders, logical gates, ...), then microprocessors and so on. The fact that trying to mimic closely silicium would be foolish has been recognized very soon, thus alternate computer architectures more adapted to molecular electronics have been evoked, such as cellular automata, or neural networks. However these are not miracle solutions because present and foreseed molecular devices can perform only very simple and rudimentary functions. On the other hand, they can be assembled in really great numbers, of the order of Avogadro number.

To help clarify ideas, the diagram of Figure 5 can be used. Each information treatment system can be broadly characterized by three qualitative parameters : (i) the degree of complexity of a processing unit or more properly of the task it can perform (ii) the number of such processing units and (iii) the complexity in the connectivity between units. Thus a sequential computer (one very complex unit) can be represented by point SQ (see Figure 5). In parallel machines, one associates several processing units, but each one is still a microprocessor, only slightly simplified. Neural networks are presently realized with a limited number of neurons (typically 10 -100). Their performances come from a subtle interplay between processors and synapses, so that the connectivity is rather complex and plays an essential rôle. Finally increasing considerably the number of processors and simplifying both their task and the connectivity leads to the cellular automaton concept. Here all connexions are identical and local.

Figure 5. Diagram showing the place of the different information processing systems
SQ : Sequential computer (1 processor) ; NN : Neural network ; CA : Cellular automata ;
ME : Molecular Electronics

In this diagram, Molecular Electronics (as we can imagine it to day) would lie in the area ME beyond cellular automata. The complexity of the task which can be performed by a molecule is very low, much lower than the one required for a cellular automaton for instance. Also the connectivity is very simple since chemists can only master near neighbour interactions. Only the number can be very great.

Thus a computer architecture based on Molecular Electronics cannot be easily imagined at the present time. It does not fit in existing architectures because they have not been devised for this kind of components. With respect to cellular automata, the present molecules are too simple; with respect to neural networks, the connectivities are too rudimentary. There is clearly a need to "start from the bottom" , i. e. to begin to think what can be done from the point of view of architecture with present molecular devices.

3. A PROGRAM FOR THE FUTURE

3.1. The need for nonlinear behaviours

Nonlinearity appears to be a necessary crossing point for an information processing system. With a sigmoidal shape such as Figure 6 (a), one can perform threshold or logical functions. A shape such as shown on Figure 6 (b) yields amplification or oscillation. Both shapes can provide bistability if a suitable amount of feedback is introduced. Table I summarizes some nonlinear behaviours which can be encountered either with molecules or at the material level. For electrical properties, very few is known as only the behaviour at weak polarization has been approached (Joachim, 1990). For high polarizations one could expect resonant tunneling as in quantum well structures, but this is still a speculation. It is clear that experimental data on systems far from thermal equilibrium would be extremely desirable as they could lead to pattern formation (Haken, 1985). With optical properties, non linear effects are now well documented (Chemla and Zyss, 1987) As far as chemical effects are concerned, nonlinearities can be observed for the dependences of concentration with time. This could be the basis of a chemical parallel computer (Haken, 1987). A demonstrator showing rudimentary image processing by a

Figure 6. Examples of nonlinear behaviours

chemical system has been built (Kuhnert, 1986). All these studies should be developed in a more systematic way.

3.2. Logical functions at the molecular scale

It is possible to realize logical functions with molecules. The simplest way for that is to make use of the properties of photochemical excited states. Experiments in which the absorption spectrum of an excited state is recorded could be the basis for a logical function (AND because they realize a double excitation). Recent experiments on photon gated hole burning can be considered as the use of this AND function (Carter, T. P., Moerner et al, 1987) (see below). More complex functions are of course highly welcome but the realization of say a NAND gate would require at least three optical beams to test the system (Birge et al, 1989). Using modified organic materials for non linear optics would perhaps be a realistic way to build a logical gate. It would be necessary to find an optical $\chi^{(2)}$ or $\chi^{(3)}$ material which could be perturbed by a photochemical excitation.

property	single molecule	material or assembly of molecules
electrical	strongly polarized molecules '	switching materials (Potember,1982, Denisevich, 1981)
optical	β hyperpolarizability	χ(2) materials
chemical concentration		chemical processor (Kuhnert, 1986)

Table I. Summary of non linear effects.

3.3. Cascading components

It is vital for the future development of Molecular Electronics to study the basic association in which the output of a given device controls the imput of another. But for Molecular Devices, generally a single electron event is studied and this constitutes the signal. Thus it is necessary to be able to detect a single electron. With present technology it is much easier to detect a flux of electrons. This means that an intermediate goal would be to show how a permanent electron flux could be controlled by a single electron (In a later step the single electron event should control another single electron event). This process is already known as "Coulomb blockade" and occurs in ultrasmall structures when the charging energy of a single electron exceeds the characteristic energy of thermal fluctuations in a properly designed system (Fulton and Dolan, 1987). The observation of this effect with a molecular device will certainly be attempted in the future. The use of large electron transfer molecules has been proposed (Gilmanshin and Lazarev, 1988) because it would theoretically facilitate the observation of single electron effects.

3.4. The Quantum Computer

The analogy between spin glasses and neural networks leads to the idea of a quantum computer. This raises however considerable problems because most structures that chemists can build are static ones, i.e. the interactions between the different constitutive elements are fixed, while the essence of a neural network is to use adaptative connexions (synapses). Thus it is vital to investigate the mutual influences of molecular devices and to learn how to modify them in a dynamic way. In a similar way, the theory of Quantum Cellular Automata is just beginning to appear. A model has been proposed where the interactions between the different elements are strictly local and are represented by a simple Hamiltonian. The time evolution has been computed, but only for small values of t (Grossing and Zeilinger, 1988). Thus much remains to be discovered.

4. SOME POSSIBLE SHORT TERM REALIZATIONS

As seen above, the development of Molecular Electronics still need a considerable effort in basic research before a complete integrated information processing sytem can be built. In the meantime however, there are some less ambitious but more realistic goals which are worth to be considered. They all have in common to be compatible with known architectures.

4.1. Synaptic materials

The implementation of neural networks suffers from the lack of a simple solution to mimic synapses. It would be interesting to have a material exhibiting a programmable electrical resistance (Hopfield et al, 1986). This effect could perhaps be realized with molecular materials, because there are already examples of compounds for which the conductivity can be altered by an electrical perturbation (Potember et al, 1982, Denisevich et al, 1981).

4.2. Optical readout of electrical potentials

To overcome the "memory contention" problem which occurs in shared memory architectures, it has been proposed to use an optical readout of electrical potentials (Kowel et al, 1987). This could be performed with a thin film of molecules exhibiting strong quadratic nonlinearities, coated on a CCD matrix. These molecules would then modulate the intensity of a laser beam according to the value of the charge stored on the matrix element (electro-optic effect).

4.3. New materials with non linear optical cubic susceptibility

Molecular materials with high cubic susceptibilities can in principle be used in optical devices showing bistability or four wave mixing, from which a variety of applications in optical treatment of signals can be conceived (Chemla and Zyss, 1987).

4.4. Optical memories

High density optical memories could be obtained by the hole burning technique which uses spatial and frequency addressing (Carter, T. P., Moerner et al, 1987). The variant of photon gated hole burning mentioned above represents a further improvement because writing an information necessitates two beams, so that reading can be nondestructive. Few (if any) molecules have been specifically designed for this kind of function, so that there is a great potential for improvement.

References

Arrhenius, T. S., Blanchard-Desce, M., Dvolaitzky, M., Lehn, J. M., Malthete, J., 1986, **Proc. Natl. Acad. Sci. USA**, 83 : 5355

Aviram, A., Ratner, M., 1974, **Chem. Phys. Lett.**, 29 : 277

Aviram, A., 1988, **J. Am. Chem. Soc.** 110 : 5687

Aviram, A. , Joachim, C., Pomerantz, M., 1988, **Chem. Phys. Lett.** 146 : 490

Aviram, A. , Joachim, C., Pomerantz, M., 1989, **Chem. Phys. Lett.** 162 : 416

Bennett, C. H., Landauer, R., 1985, **Scientific American**, 253 : 48

Bernard, J., Orrit, M., 1990, **Compt. Rend. Acad. Sc.** in the press

Birge, R. R., Ware, B. R., Dowben, P. A., Lawrence, A. F., 1989, in "**Molecular Electronics, Science and technology**", A. Aviram, Ed, Engineering Foundation, New York, p. 275

Blanchard-Desce, M. 1989, Thesis, Université Pierre et Marie Curie, Paris

Buttiker, M., 1989, **Phys. Rev. B**, 40 : 3409

Carter, F. L. (Ed), 1982, "**Molecular Electronic Devices**", M. Dekker, New York

Carter, F. L. (Ed), 1987, "**Molecular Electronic Devices II**", M. Dekker, New York

Carter, F. L., Siatkowski, R. E., Wohltjen, H. (Eds), 1988, "**Molecular Electronic Devices**", North Holland, Amsterdam.

Carter, T. P., Brauchle, C., Lee, V. Y., Manavi, M., Moerner, W. E., 1987, **J. Phys. Chem.** 91 : 3998

Chemla, D., Zyss, J. (Eds), 1987, "**Nonlinear optical properties of organic molecules and crystals**", Acad. Press, Orlando

Denisevich, P., Willman, K. W., Murray, R. W., 1981, **J. Am. Chem. Soc.** 103 : 4727

Effenberger, F., Schlosser, H., Bauerle, P., Maier, S., Port, H., Wolf, H. C., 1988, **Angew. Chem. Int. Ed. Engl.** 27 : 281

Fujihira, M., Nishiyama, K., Yamada, H., 1985, **Thin Solid Films**, 132 : 77

Fulton, T. A., Dolan, G. J., 1987, **Phys. Rev. Lett.**, 59 : 109

Geddes, N. J., Sambles, J. R., Jarvis, D. J., Parker, W. G., Sandman, D. J., 1990, **Appl. Phys. Lett.**, 56 : 1916

Gilmanshin, R. I., Lazarev, P. I., 1988, **J. Molec. Electronics**, 4 : S 83

Gourdon, A. Launay, J. P., 1990, Work in progress

Grabowski, Z. R., Rotkiewicz, K., Siemiarczuk, A., Cowley, D. J., Baumann, W., 1979, **Nouv. Journ. Chim.** 3 : 443

Grossing, G., Zeilinger, A., 1988, **Physica D**, 31 : 70

Gust, J. D., Moore, T. A., 1990, **Science**, 248 : 199

Haken, H., 1985, **Physica Scripta**, 32 : 274

Haken, H., 1987, **J. Chim. Phys. (Paris)**, 84 : 1289

Hauser, A., 1986, **Chem. Phys. Lett.** 124 : 543

Heller, H. G., 1983, **IEE Proceedings**, 130 : Pt 1 : 209

Hopfield, J. J., Tank, D. W., 1986, **Science**, 233 : 625

Hopfield, J. J., Onuchic, J. N., Beratan, D. N., 1988, **Science**, 241 : 817

Joachim, C., Launay, J. P., Woitellier, S., 1990, **Chem. Phys.** In press

Joachim, C., 1990, **New. Journ. Chem.** in the press

Keyes, R. W., 1988, in **Advances in Electronics and Electron Physics**, Acad. Press, 70 : 159

Kowel, S. T., Selfridge, R., Eldering, C., Matloff, N., Stroeve, P., Higgins, B. G., Srinivasan, M. P., Coleman, L. B., 1987, **Thin Solid Films**, 152 : 377

Kuhnert, L., 1986, **Nature**, 319 : 393

Launay, J. P., Sowinska, M., Leydier, L., Gourdon, A., Amouyal, E., Boillot, M. L., Heisel, F., Miehé, J. A., 1989, **Chem. Phys. Lett.** 160 : 89

Launay, J. P., Tourrel-Pagis, M., Lipskier, J. F., Marvaud, V., Joachim, C., 1990, submitted

Metzger, R. M., Panetta, C. A., Heimer, N. E., Bhatti, A. M., Torres, E., Blackburn, G. F., Tripathy, S., Samuelson, L. A., 1986, **J. Molec. Electronics**, 2 : 119

Moerner, W. E., Kador, L., 1989, **Phys. Rev. Lett.**, 62 : 2535

Moore, T. A., Gust, D., Mathis, P., Mialocq, J. C., Chachaty, C., Bensasson, R. V., Land, E. J., Doizi, D., Liddell, P. A., Lehman, W. R., Nemeth, G. A., Moore, A. L., 1984, **Nature**, 307 : 5952

Potember, R. S., Poehler, T. O., Cowan, D. O., Carter, F. L., Brant, P., 1982, in **"Molecular Electronic Devices"**, M. Dekker, New York p. 73

Quate, C. F., Lang, C. A., 1989, in **"Molecular Electronics, Science and technology"**, A. Aviram, Ed, Engineering Foundation, New York, p. 79

Seta, P., Bienvenue, E., Moore, A. L., Mathis, P., Bensasson, R. V., Liddell, P., Pessiki, P. J., Joy, A., Moore, T. A., Gust, D., 1985, **Nature**, 316 : 653

Sautet, P., Joachim, C., 1988, **J. Physics C.**, 21 : 3939

Woitellier, S., Launay, J. P., Spangler, C. W., 1989, **Inorg. Chem.** , 28 : 758

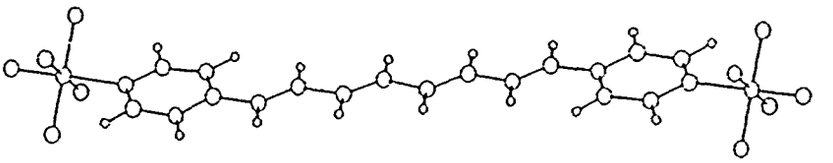
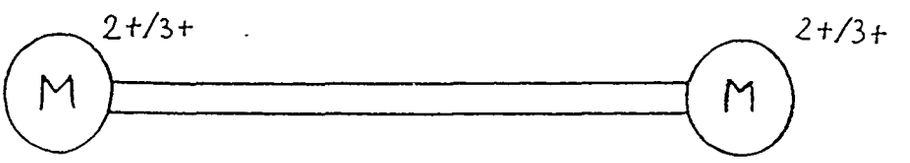


Fig 1



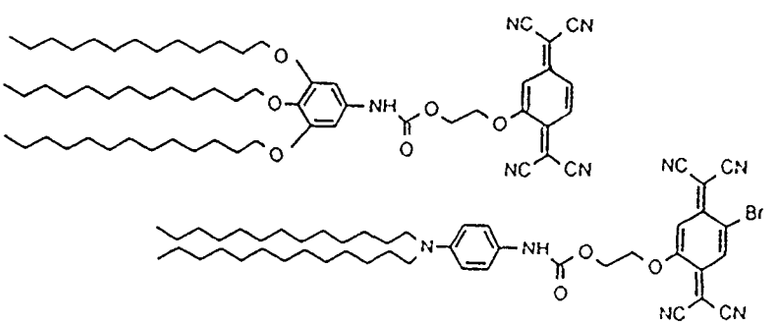


Fig 2

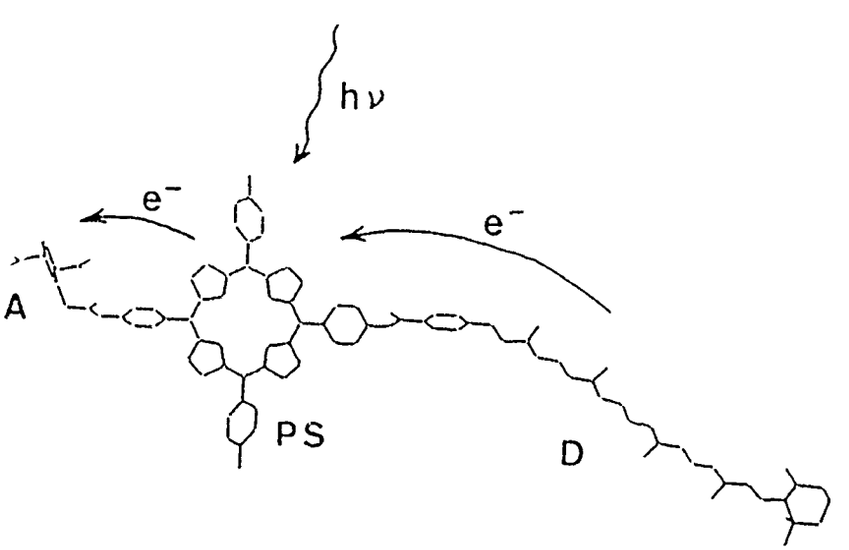


Fig 3

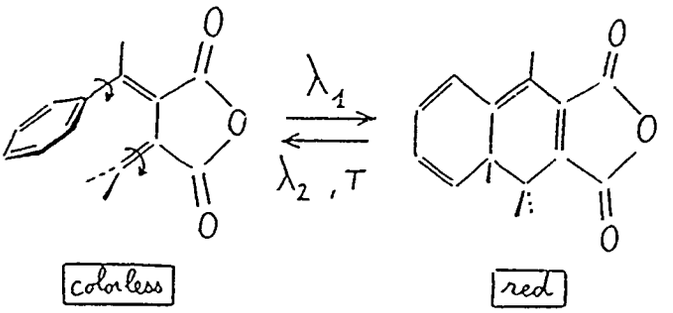


Fig 4

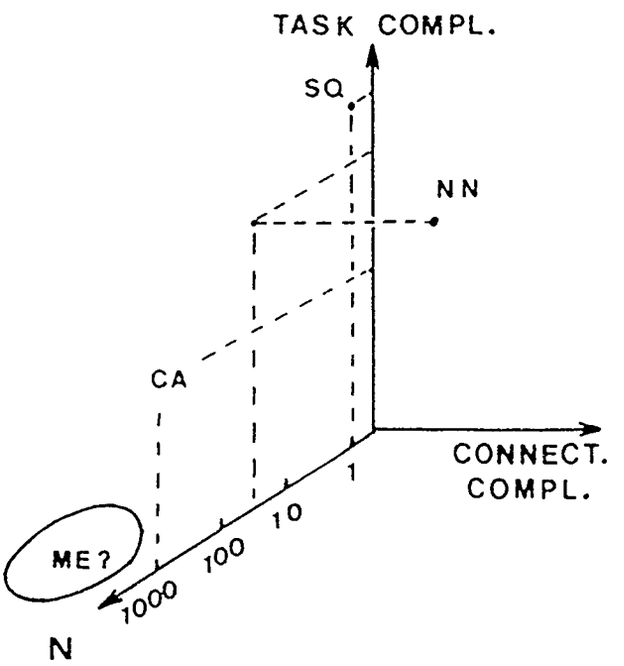


Fig 5

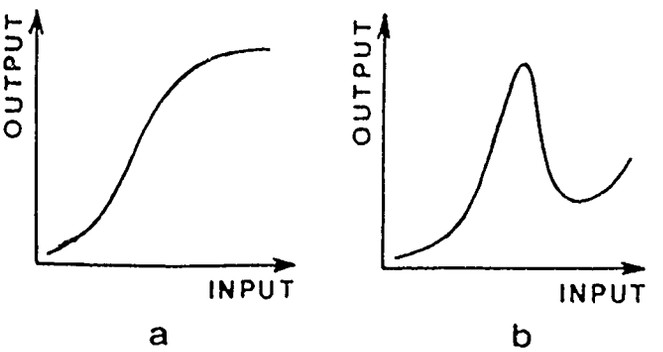


Fig 6