

Quantum Transport in Nanoscale Devices

Authors:

D. Vasileska, D. Mamaluy, I. Knezevic, H. R. Khan, and S. M. Goodnick

Table of Contents

1. Introduction

1.1 Need for Quantum Transport in Nanoscale Devices

- 1.1.1 *Si Based Nanoelectronics*
 - (A) *Device Scaling*
 - (B) *Beyond conventional silicon*
 - (C) *Quantum transport effects in nanoscale devices*
- 1.1.2 *Heterostructure Devices in III-V or II-VI technology*
 - (A) *Modulation doping of AlGaAs/GaAs heterostructures with in-plane transport*
 - (B) *Vertical transport - Resonant Tunneling Devices*
- 1.1.3 *Modeling of nanoscale devices*

1.2 Open Systems

- 1.2.1 *Tunneling Theory*
 - (A) *General Notation*
 - (B) *Stationary States for a Free Particle*
 - (C) *Potential Step*
 - (D) *Tunneling through a single barrier*
- 1.2.2 *Tunneling Through Arbitrary Piecewise-Constant Barrier*
- 1.2.3 *Evaluation of the Current Density*

2. Near-Equilibrium Steady State Transport

2.1 Conductance – The Landauer-Buttiker formula

3. Far-From-Equilibrium Transport

3.1 Mixed States and Distribution Function

- 3.1.1 *Irreversible Processes and MASTER Equations*
- 3.1.2 *The Boltzmann Equation*
 - (A) *Approximations made for the distribution function*
 - (B) *Boltzmann transport equation*
 - (C) *Scattering Processes*
 - (D) *Statistical Averages*
 - (E) *Ensemble Monte Carlo*

3.2 The Wigner Distribution Function

3.3 Green's Functions

4. CBR Method for the Solution of the 3D Green's Function Method as Applied to Modeling 2D/3D FinFET Devices

4.1 Bound States Treatment

4.2 Energy Discretization

- 4.3 Self-Consistent Solution**
- 4.4 Device Hamiltonian, Algorithm and Some Numerical Details**
- 4.5 Simulation Example – 2D Results**
- 4.6 Simulation Example – 3D Results**
 - 4.6.1 DG FinFET: 2D vs. 3D Simulation*
 - 4.6.2 Double-Gate(DG) vs. Tri-Gate(TG) FinFET*
 - 4.6.3 Effects of an Unintentional Dopant : DG vs. TG FinFET*

5. Reduced Density Matrix Formalism and its Application to Modeling RTDs

- 5.1 Partial-trace-free approach to open systems. Equations with memory dressing**
 - 5.1.1 Reduced density matrix (statistical operator)*
 - 5.1.2 Basic definitions*
 - 5.1.3 Projection-operator technique. Conventional time-convolutionless equation of motion*
 - 5.1.4 Eigenproblem of a projection operator. Partial-trace-free approach*
 - 5.1.5 Partial-trace-free time-convolutionless equation of motion for the reduced density matrix*
 - 5.1.6 “Purely system states” and “entangled states”*
 - 5.1.7 Memory dressing and the reduced density matrix*
 - (A) Evaluation of the memory dressing $R(t)$*
 - 5.1.8 Short-time evolution in the case of initially uncorrelated system and environment*
 - 5.1.9 Coarse-grained Markovian evolution*
- 5.2 Decoherence in the active region of a resonant-tunneling diode**
- 5.3 Generalizing Green’s function for a full treatment of dynamically open systems**
 - 5.3.1 Two-time correlation functions for open systems*
 - 5.3.2 Transport in the transient regime*
 - 5.3.3 Transport in a far-from-equilibrium steady state*

6. Conclusions

Acknowledgements

REFERENCES

1. Introduction

Semiconductor device-based electronics industry is the largest industry in the world with global sales of over one trillion dollars since 1998. If current trends continue, the sales volume of the electronics industry will reach three trillion dollars and will constitute about 10% of the gross world product (GWP) by 2010 [1]. The revolution in semiconductor industry, a subset of the electronics industry, began in 1947 (see Figure 1-1) with the fabrication of bipolar devices on slabs of polycrystalline germanium (Ge) [2].

- Bipolar transistor:	1947	- DTL - technology	1962
- Monocrystal germanium:	1950	- TTL - technology	1962
- First good BJT:	1951	- ECL - technology	1962
- Monocrystal silicon:	1951	- MOS integrated circuit	1962
- Oxide mask,		- CMOS	1963
Commercial silicon BJT:	1954	- Linear integrated circuit	1964
- Transistor with diffused		- MSI circuits	1966
base:	1955	- MOS memories	1968
- Integrated circuit:	1958	- LSI circuits	1969
- Planar transistor:	1959	- MOS processor	1970
- Planar integrated circuit:	1959	- Microprocessor	1971
- Epitaxial transistor:	1960	- I ² L	1972
- MOS FET:	1960	- VLSI circuits	1975
- Schottky diode:	1960	- Computers using	
- Commercial integrated		VLSI technology	1977
circuit (RTL):	1961	- ...	

Figure 1-1. Some Historical Dates.

Single-crystalline materials were later proposed and introduced, making possible the fabrication of grown junction transistors. Migration to silicon (Si)-based devices was initially hindered by the stability of the Si/SiO₂ materials system, necessitating a new generation of crystal pullers with improved environmental controls to prevent SiO₂ formation. Later, the stability and low interface-state density of the Si/SiO₂ materials system provided passivation of junctions and eventually the migration from bipolar devices to field-effect devices in 1960. By 1968, both complementary metal-oxide-semiconductor devices (CMOS) and polysilicon gate technology, that allowed self-alignment of the gate to the source/drain of the device, had been developed. These innovations permitted a significant reduction in power dissipation and a reduction of the device overlap capacitance, improving frequency performance and resulting in the essential components of the modern CMOS device. Professor Herbert Kroemer's contributions to heterostructures — from heterostructure bipolar transistors [3] to lasers [4] — culminated in a Nobel Prize in Physics in 2000 and have paved the way for novel heterostructure devices including those in silicon. The unique properties of the variety of semiconductor materials have enabled the development of a wide variety of ingenious devices that have literally changed our world. To date, there are about 60 major devices, with over 100 device variations related to them.

1.1 Need for Quantum Transport in Nanoscale Devices

1.1.1 Si Based Nanoelectronics

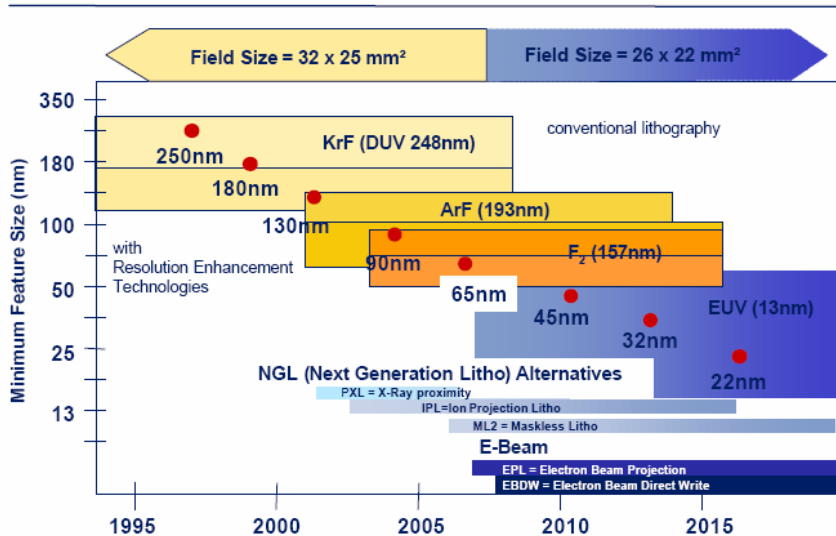
The metal-oxide-semiconductor-field-effect transistor (MOSFET) and related integrated circuits now constitute about 90% of the semiconductor device market. Combining silicon with the elegance of the field-effect transistor (FET) structure has allowed simultaneously making devices smaller, faster, and cheaper—the mantra that has driven the modern semiconductor microelectronics industry. Nowadays, the single factor driving the continuous device improvement is the semiconductor industry's relentless effort to reduce the cost per function on a chip. The way this is done is to put more devices on a chip while either reducing manufacturing costs or holding them constant. This leads to three methods of reducing the cost per function. The first is transistor scaling, which involves reducing the transistor size in accordance with some goal, i.e. keeping the electric field constant from one generation to the next. With smaller transistors, more can fit into a given area than in previous generations. The second method is circuit cleverness, which is

associated with the physical layout of the transistors with respect to each other. If the transistors can be packed into a tighter space, then more devices can fit into a given area than before. The third method is to make the die larger. More devices can be fabricated on a larger die. All the while, the semiconductor industry is constantly looking for technological breakthroughs to decrease the manufacturing cost. All of this effort serves to reduce the cost per function on a chip.

(A) Device Scaling

Device engineers are most concerned with the method of scaling introduced in the previous paragraph. The semiconductor industry has been so successful in providing continued system performance improvement year after year that the Semiconductor Industry Association (SIA) has been publishing roadmaps for semiconductor technology since 1992. These roadmaps represent a consensus outlook of industry trends, taking history as a guide. Recent roadmaps [5] incorporate participation from the global semiconductor industry, including the United States, Europe, Japan, Korea, and Taiwan. They basically affirm the desire of the industry to continue with Moore’s law [6], which is often stated as doubling of transistor performance and quadrupling of the number of devices on a chip every three years. The phenomenal progress signified by Moore’s law has been achieved through scaling of the MOSFET from larger to smaller physical dimensions. Scaling of CMOS technology has progressed relentlessly from a line width of 1 μm to the current 65-nm line width. Two key features characterize this era. First, slavish devotion to scaling by constant improvements in lithography (see Figure 1-2, top panel), as described by Dennard *et al.* [7]. At present, 193 nm lithography steppers are in general use. The active pursuit of advanced lithographic techniques, such as extreme ultraviolet (EUV) lithography currently in use at the Berkeley labs, which makes use of light at a wavelength of 13 nm, illustrates the relentless ardor with which scaling is still being pursued. Secondly, a minimal rate of introduction of substantially new materials and structures. Substantial effort is required to introduce new materials, and great effort is required to ensure that both manufacturable and reliable integration has been attained. Significant efforts that are currently under way include identification for a replacement of silicon dioxide as the gate dielectric for MOSFETs and, recently, announcements regarding the introduction of silicon–germanium in CMOS technology, give further evidence of forces for change.

Technology Nodes below 100nm strongly depend on availability of extended 193nm and 157nm Litho or NGL



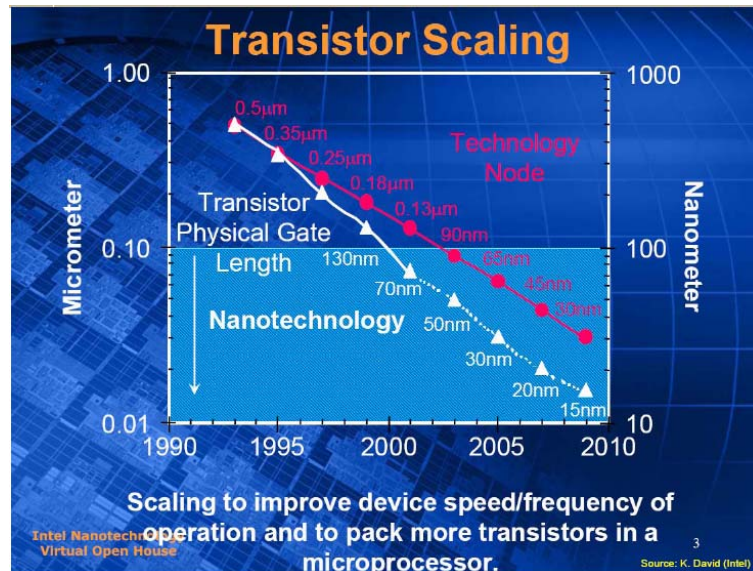


Figure 1-2. Top panel – Needed improvements in lithography. Bottom panel – Transistor scaling as seen by Intel.

Regarding conventional silicon MOSFETs, the device size is scaled in all dimensions (see Figure 1–2 bottom panel), resulting in smaller oxide thickness, junction depth, channel length, channel width, and isolation spacing. Currently, 65 nm (with a physical gate length of 45 nm) is the state-of-the-art process technology, but even smaller dimensions are expected in the near future. The SIA forecasts that this exponential scaling of silicon (or silicon-compatible) FETs and integrated circuits will continue at least until the year 2010, when devices with 10 nm features should become commercially available. The groups from Toshiba and Lucent Bell Labs have fabricated *n*-channel MOSFETs with effective gate lengths below 25 nm [8,9] and thus demonstrated that these feature sizes are feasible. An ultrasmall MOSFET with a channel length of 15 nm has been demonstrated in 2001 [10]. Conventional silicon MOS transistors with physical gate length of 10 nm have been demonstrated by Intel Corporation [11]. These devices can serve as the basis for the most advanced integrated circuit chips containing over one trillion ($> 10^{12}$) devices. Intel has begun making some chips on the new process, with gigabit Ethernet, optical networking, and wireless ICs among the applications. As mentioned, device miniaturization results in reduced unit cost per circuit function. For example, the cost per bit of memory chips has halved every 2 years for successive generations of DRAM circuits. As device dimensions decrease, the intrinsic switching time decreases. Device speed has improved by four orders of magnitude since 1959. Higher speeds lead to expanded IC functional throughput rates. In the future, digital ICs will be able to perform data processing and numerical computation at terabit-per-second rates. As devices become smaller, they also consume less power. Therefore, device miniaturization also reduces the energy used for each switching operation. The energy dissipated per logic gate has decreased by over one million times since 1959.

It is important to point out that the exponential growth in integrated circuit complexity, which has seen a hundred-million-fold increase in transistor count per chip over the past forty years, is finally facing its limits. Limits projected in the past have seemed to melt away before the concerted efforts of researchers and technologists, yet this time the limits seem more real and are already forcing new strategies on the design of future devices. Critical dimensions, such as transistor gate length and oxide thickness, are reaching physical limitations. Maintaining dimensional integrity at the limits of scaling is a challenge. Considering the manufacturing issues, photolithography becomes difficult as the feature sizes approach the wavelength of ultraviolet light. In addition, it is difficult to control the oxide thickness when the oxide is made up of just a few monolayers. Processes will be required that approach atomic-layer precision. Just being able to model future processes to predict geometries and doping concentrations of future devices is a challenge that has not been met. The existing empirical techniques will have to be aided by increasingly sophisticated *ab initio* calculations in order to reduce the experimental parameter space to manageable proportions.

In addition to the processing issues there are also some fundamental device issues. Shrinking the conventional MOSFET beyond the 50-nm-technology node requires innovations to circumvent barriers due to the fundamental physics that constrains the conventional MOSFET. The limits most often cited [12] include: (1) quantum-mechanical tunneling of carriers through the thin gate oxide; (2) quantum-mechanical tunneling of carriers from source to drain, and from drain to the body of the MOSFET; (3) control of the density and location of dopant atoms in the MOSFET channel and source/drain region to provide a high on-off current ratio; (4) control of threshold voltage over the die is another major scaling challenge; (5) voltage-related effects such as subthreshold swing, built-in voltage and minimum logic voltage swing; (6) short-channel effects (SCEs), such as drain-induced barrier lowering (DIBL) that degrade the device performance; (7) Hot carriers that degrade device reliability, and (8) other application-dependent power-dissipation limits. For analog/RF applications, the challenges additionally include sustaining linearity, low noise figure, power-added-efficiency, and transistor matching.

The quickening pace of MOSFET technology scaling is accelerating the introduction of many new technologies to extend CMOS into nanoscale MOSFET structures heretofore not thought possible (see Figure 1-3). A cautious optimism is emerging that these new technologies may extend MOSFETs to the 22 nm node (9-nm physical gate length) by 2016 if not by the end of this decade. These new devices will likely feature several new materials cleverly incorporated into new non-bulk MOSFET structures. They will be ultra fast and dense with a voracious appetite for power. Intrinsic device speeds may be more than 1 THz and integration densities will exceed 1 billion transistors/cm². Excessive power consumption, however, will demand judicious use of these high-performance devices only in those critical paths requiring their superior performance. Two or perhaps three other lower performance, more power-efficient MOSFETs will likely be used to perform less performance-critical functions on the chip to manage the total power consumption.

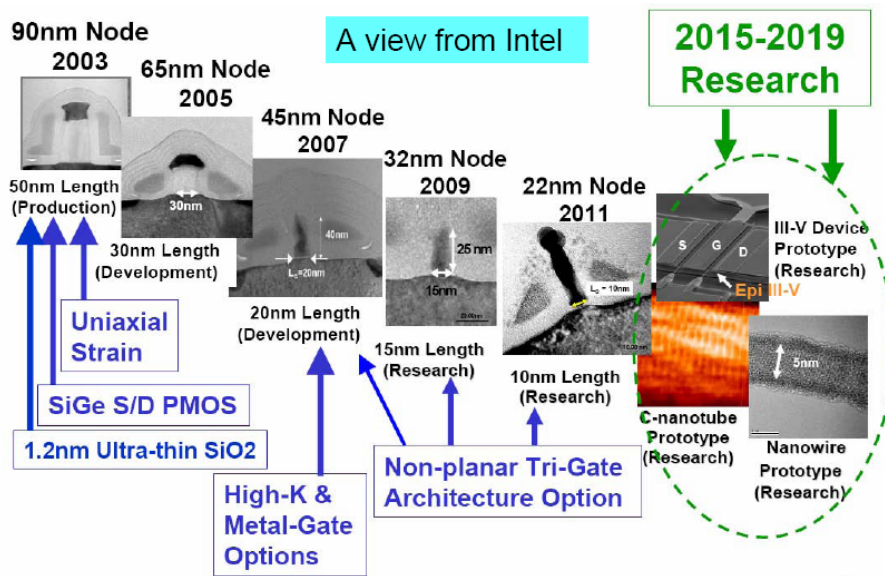


Figure 1-3. A view from Intel on future technology nodes.

(B) Beyond conventional silicon

For digital circuits, a figure of merit for MOSFETs for unloaded circuits is CV/I , where C is the gate capacitance, V is the voltage swing, and I is the current drive of the MOSFET. For loaded circuits, the current drive of the MOSFET is of paramount importance. Keeping in mind both the CV/I metric and the benefits of a large current drive, we note that device performance may be improved [12] by: (1) inducing a larger charge density for a given gate voltage drive; (2) enhancing the carrier transport by improving the mobility, saturation velocity, or ballistic transport; (3) ensuring device scalability to achieve a shorter channel length; and (4) reducing parasitic capacitances and parasitic resistances. For capitalizing these opportunities, the proposed technology options generally fall into two categories: (I) new materials and (II)

new device structures. In many cases, the introduction of a new material requires the use of a new device structure, or vice versa. To fabricate devices beyond current scaling limits, IC companies are simultaneously pushing the planar, bulk silicon CMOS design while exploring alternative gate stack materials (high- k dielectric [13] and metal gates), band engineering methods (using strained Si [14,15,16] or SiGe [5]), and alternative transistor structures. The concept of a band-engineered transistor is to enhance the mobility of electrons and/or holes in the channel by modifying the band structure of silicon in the channel in a way such that the physical structure of the transistor remains substantially unchanged (see Figure 1-4). This enhanced mobility increases the transistor transconductance (g_m) and on-drive current (I_{on}). A SiGe layer or a strained-silicon on relaxed SiGe layer is used as the enhanced-mobility channel layer. It has already been demonstrated experimentally that at $T = 300$ K (room temperature), effective hole enhancement of about 50% can be achieved using the SiGe technology [17]. Intel has adopted strained silicon technology for its 65 nm process [18]. The results were nearly a 20% performance improvement, with only a few additional process steps. Scott Thompson, an Intel fellow, said Intel believes it can get another performance boost by increasing the germanium content at the 45 nm node.

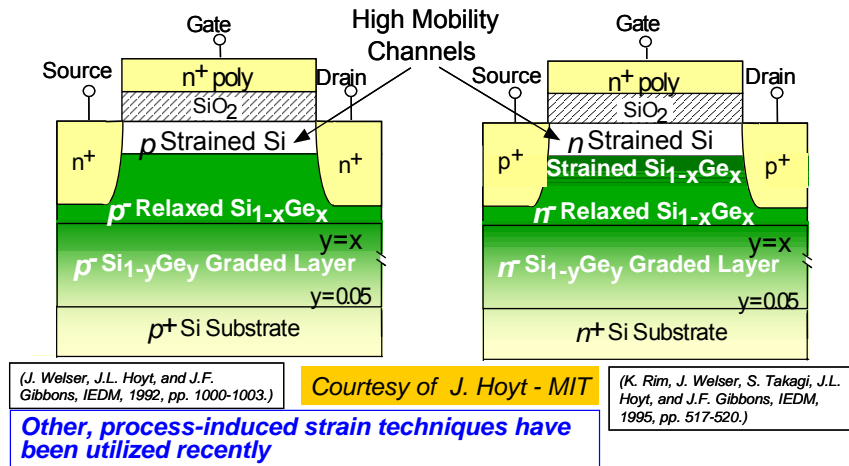


Figure 1-4. Method I for improving device performance – Introduction of new materials that lead to globally induced strain. Other methods that lead to locally induced strain have been recently pursued by Intel Corporation.

The challenge in identifying suitable high- k dielectrics and metal gates for both conventional PMOS (p -channel MOS) and NMOS (n -channel MOS) transistors has led to early adoption of alternative transistor designs (see Figure 1-5). These include primarily partially-depleted (PD) and fully-depleted (FD) silicon-on-insulator (SOI) devices. Today there is also an extensive research in double-gate (DG) structures, and FinFET transistors [19], which have better electrostatic integrity and theoretically have better transport properties than single-gated FETs. A FinFET is a form of a double gate transistor having surface conduction channels on two opposite vertical surfaces and having current flow in the horizontal direction. The channel length is given by the horizontal separation between source and drain and is usually determined by a lithographic step combined with a side-wall spacer etch process. Many innovative structures, involving structural challenges such as fabrication on nanometer-scale fins and nanometer-scale planarization over an entire wafer, are currently under investigation. In conclusion, the semiconductor industry is approaching the end of an era of scaling gains by rote shrinkage of device dimensions, and entering a post-scaling era, a new phase of CMOS evolution in which innovation is demanded simply to compete. The trends in benefits to density, performance, and power will be continued through such innovations. Rather than coming to a close, a new era of CMOS technology is just beginning. Table 1-1 [20] summarizes the advantages and challenges of some of the above-mentioned device structures.

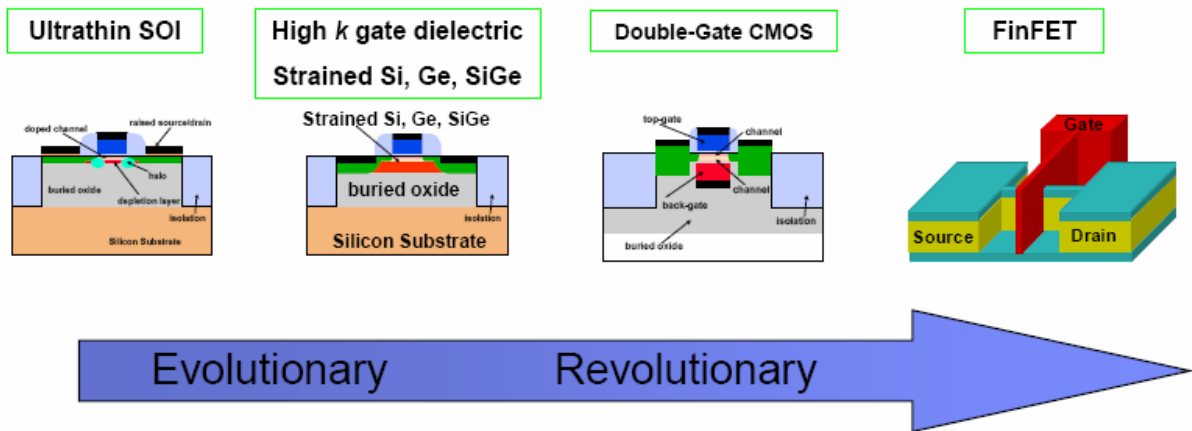


Figure 1-5. Method II for improving device performance – Introduction of new device structures.

Table 1-1. Non-classical CMOS devices.

Device	Ultrathin Body (UTB) SOI	Band-Engineered Transistor	Vertical Transistor	FinFET	Double-Gate
Concept	Fully-depleted SOI	SiGe or Strained Si; bulk Si or SOI	Double-gate or surround-gate structure		
Application/Driver	Higher performance, higher transistor density, lower power dissipation				
Advantages	Improved subthreshold slope; V_T controllability	Higher drive current; compatible with bulk Si and SOI	Higher drive current; lithography independent gate length	Higher drive current; Improved subthreshold slope; improved short-channel effect (SCE)	
Scaling Issues	Si film thickness, gate stack; worse SCE than bulk CMOS	High mobility film thickness (SOI); gate stack; integratability	Si film thickness; gate stack; integratability; process complexity; accurate TCAD	Gate alignment; Si film thickness; gate stack; integratability; process complexity; accurate TCAD	
Design Challenges	Device characterization; compact model and parameter extraction	Device characterization	Device characterization; PD versus FD; compact model and parameter extraction; applicability to mixed signal applications		

(C) Quantum transport effects in nanoscale devices

Semiconductor transport in the nanoscale region has approached the regime of quantum transport. This is suggested by two trends: (1) within the effective-mass approximation, the thermal de Broglie wavelength for electrons in semiconductors is on the order of the gate length of nano-scale MOSFETs, thereby

encroaching on the *physical optics* limit of wave mechanics; (2) the time of flight for electrons traversing the channel with velocity well in excess of 10^7 cm/sec is in the 10^{-15} to 10^{-12} sec region—a time scale which equals, if not being less than the momentum and energy relaxation times in semiconductors which precludes the validity of the Fermi's Golden Rule that is used to calculate scattering rate out of initial state \mathbf{k} [21].

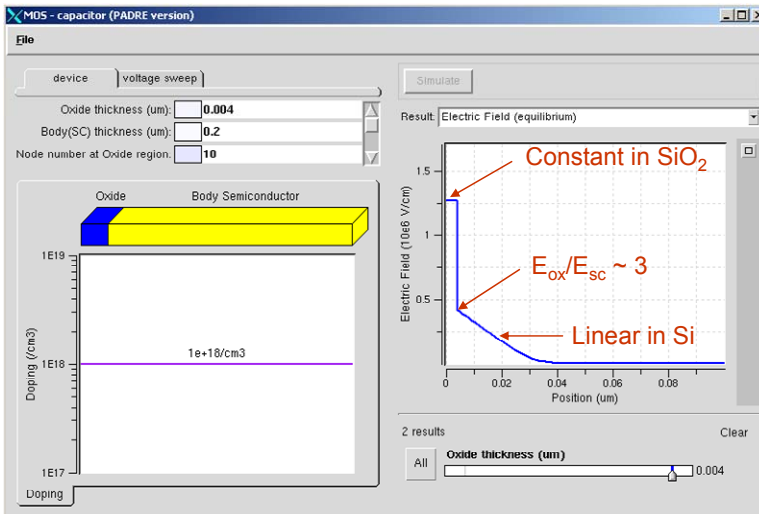
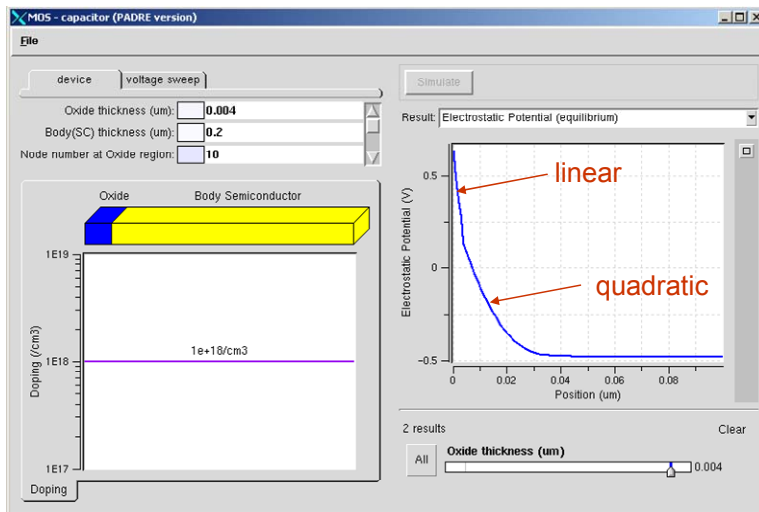
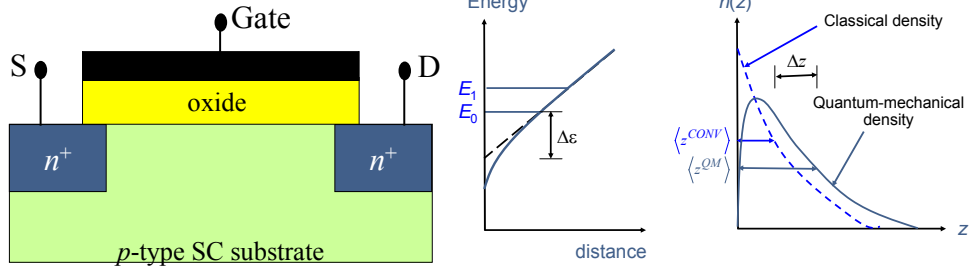
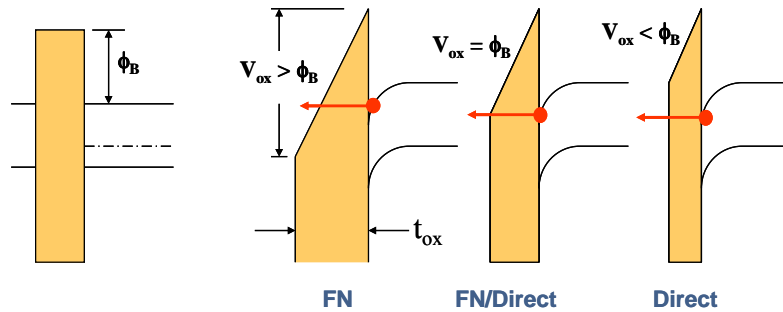


Figure 1-6. Top left panel – prototypical description of MOSFET device. Top right panel – classical vs. quantum charge description in a triangular potential well. Middle and bottom panel – potential and electric fields as obtained by the MOSCap tool designed and deployed on the nanoHUB by Dragica Vasileska. It utilizes PADRE as a background simulator.

The static quantum effects, such as tunneling through the gate oxide and the energy quantization in the inversion layer of a MOSFET are also significant in nanoscale devices (see Figure 1-6). The current generation of MOS devices has oxide thicknesses of roughly 15-20Å and is expected that, with device scaling deeper into the nanoscale regime, oxides with 8-10Å thickness will be needed. The most obvious quantum mechanical effect, seen in the very thinnest oxides, is gate leakage via direct tunneling through the oxide (see Figure 1-7). The exponential turn-on of this effect sets the minimum practical oxide thickness (~10Å). A second effect due to spatial/size-quantization in the device channel region is also expected to play significant role in the operation of nanoscale devices. To understand this issue, one has to consider the operation of a MOSFET device based on two fundamental aspects: (1) the channel charge induced by the gate at the surface of the substrate, and (2) the carrier transport from source to drain along the channel. Quantum effects in the surface potential will have a profound impact on both, the amount of charge which can be induced by the gate electrode through the gate oxide, and the profile of the channel charge in the direction perpendicular to the surface (the transverse direction). The critical parameter in this direction is the gate-oxide thickness, which for a nanoscale MOSFET device is, as noted earlier, on the order of 1 nm.



- For $t_{ox} \geq 40 \text{ \AA}$, Fowler-Nordheim (FN) tunneling dominates
- For $t_{ox} < 40 \text{ \AA}$, direct tunneling becomes important
- $I_{dir} > I_{FN}$ at a given V_{ox} when direct tunneling active
- For given electric field: I_{FN} independent of oxide thickness, I_{dir} depends on oxide thickness

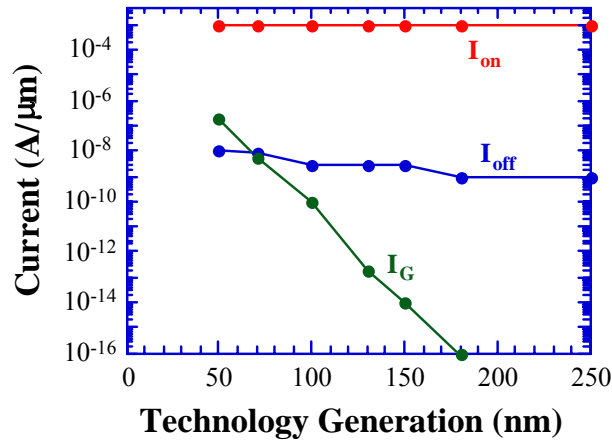


Figure 1-7. Top panel – Fowler-Nordheim vs. Direct tunneling. Bottom panel – Currents vs. technology generation.

Another aspect, which determines device characteristics, is the carrier transport along the channel (lateral direction). Because of the two-dimensional (2D), and/or one-dimensional (1D) in the case of narrow-width devices, confinement of carriers in the channel, the mobility (or microscopically speaking, the carrier scattering) will be different from the three-dimensional (3D) case. Theoretically speaking, the 2D/1D mobility should be larger than its 3D counterpart due to reduced density of states function, i.e. reduced number of final states the carriers can scatter into, which can lead to device performance enhancement. A well known approach that takes this effect into consideration is based on the self-consistent solution of the 2D Poisson–1D Schrödinger–2D Monte Carlo, and requires enormous computational resources as it requires storage of position dependent scattering tables that describe carrier transition between various subbands [22]. More importantly, these scattering tables have to be re-evaluated at each iteration step as the Hartree potential (the confinement) is a dynamical function and slowly adjusts to its steady-state value [23]. It is important to note, however, that in the smallest size devices, carriers experience very little or no scattering at all (ballistic limit), which makes this second issue less critical when modeling nanoscale devices. Ballistic transport in 2D/3D FinFETs is elaborated in great details in section 4 of this review article.

On the other hand, the dynamical quantum effects in nanoscale MOSFETs, associated with energy dissipating scattering in electron transport are physically much more involved [24]. There are several fundamental problems one must overcome in this regard. For example, since ultrasmall devices, in which quantum effects are expected to be significant, are inherently three-dimensional (3D) one must solve the 3D Schrödinger equation. In addition, the device region (channel) is always connected to the classical reservoirs (source and drain) from which the *macroscopic* currents are extracted. In other words, the entire device is intrinsically an open-system and the quantum region and the reservoirs must be treated on the same physical ground [25]. This is, of course, one of the most difficult problems to solve in quantum physics and will be addressed in section 5 of this review article.

There is another fundamental problem associated with quantum transport. Since one is mainly concerned with devices operated at room temperature, phase-breaking inelastic scattering is inevitable. One would like to stress that this is true even under quasi-ballistic as well as diffusive transport regime. One is, therefore, in a somewhat controversial situation. The phase coherence should be preserved because of the small device size (see Figure 1-8), whereas phase breaking scattering has to be included because of the relatively high operating temperature. However, the treatment of the phase-breaking scattering in quantum transport is not quite clear.

Intel - 2004

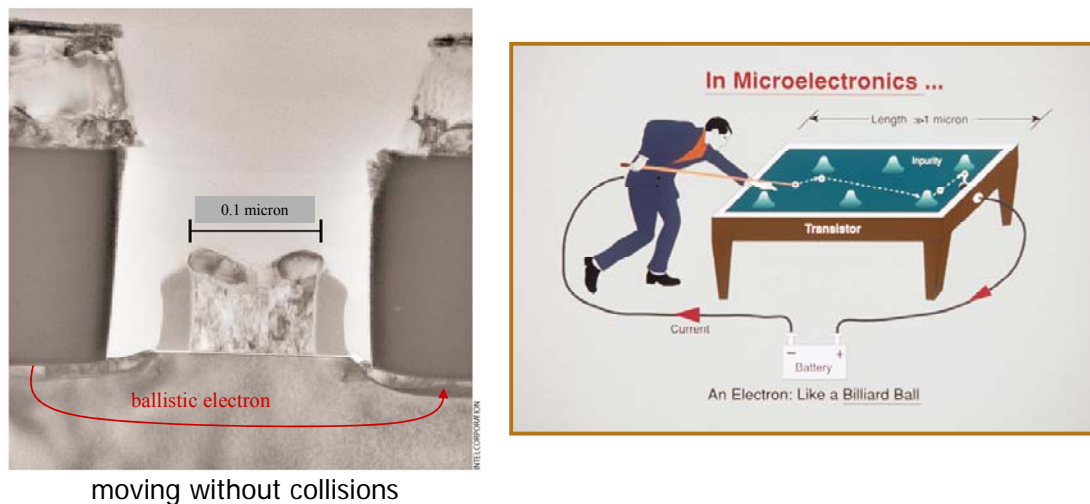


Figure 1-8. Ballistic Motion in transistors – left panel. An electron like a billiard ball moving through potential barriers due to impurities.

Another question that becomes important in nanoscale devices is the treatment of the scattering process itself. Within the Born approximation, the scattering processes are treated as independent and instantaneous

events. It is, however, a nontrivial question to ask whether such an approximation is actually satisfactory under high temperature, in which the electron strongly couples with the environment (such as phonons and other carriers). In fact, many dynamical quantum effects, such as the collisional broadening of the states or the intra-collisional field effect, are a direct consequence of the approximation employed for the scattering kernel in the quantum kinetic equation. Depending on the orders of the perturbation series in the scattering kernel, the magnitude of the quantum effects could be largely changed. Many of these issues relevant to quantum transport in semiconductors are highlighted in Table 1-2. Note that at present there is no consensus as to what can be the unified approach to quantum transport in semiconductors. Density matrices, and the associated Wigner function approach, Green's functions, and Feynman path integrals all have their application strengths and weaknesses.

Table 1-2. Quantum Effects.

1.	Static Quantum Effects
	<ul style="list-style-type: none"> • Periodic crystal potential and band structure effects • Scattering from defects, phonons • Strong electric and magnetic field • Inhomogeneous electric field • Tunneling–gate oxide tunneling and source-to-drain tunneling • Quantum wells and band-engineered barriers
<hr/>	
2.	Dynamical Quantum Effects
	<ul style="list-style-type: none"> • Collisional broadening • Intra-collisional field effects • Temperature dependence • Electron-electron scattering • Dynamical screening • Many-body effects • Pauli exclusion principle

1.1.2 Heterostructure Devices in III-V or II-VI technology

Innovations in materials growth technologies have been the key to the investigation of new materials, new physical concepts and their application in novel electronic and optical devices. The invention of semiconductor lasers [26] and metal semiconductor field effect transistors (MESFETs) [27] were important technological breakthroughs that occurred in GaAs and determined the directions of its future research to overcome the shortcomings in the then existing GaAs materials technology. The first breakthrough was the development of liquid phase epitaxy (LPE) for GaAs and other related III-V compounds [28]. The advantages of LPE included reduced background impurity, native defect concentrations, and the realization of alloy material systems and new structures by combining different materials (heteroepitaxy and heterojunctions) which resulted in its widespread use. These attributes resulted in advances in microwave, high speed digital, and optoelectronic devices based upon two factors, firstly, the improvement in the materials properties of GaAs and, secondly, the application of AlGaAs/GaAs heterostructures. Improvement in the purity of the materials reduced the non-radiative recombination rates, resulting in longer minority carrier lifetimes and lower trap-related noise levels. Though LPE led to the introduction of heterojunctions it had a lot of shortcomings in controlling layer thicknesses, surface and interface flatness and interface abruptness.

The development of Molecular Beam Epitaxy (MBE) [29] (see Figure 1-9) has been pushed by device technology to achieve structures with atomic layer dimensions and this has led to an entirely new area of condensed matter physics and investigation of structures exhibiting strong quantum size effects. MBE has played a key role in the discovery of phenomena like two dimensional electron and hole gases, quantum Hall effect [30] and new structures like quantum wires and quantum dots, etc. The continued miniaturization of solid state devices is leading to the point where quantization-induced phenomena

become more and more important. These phenomena have shown that the role of material purity, native defects and interface quality are very critical to the device performance. Modulation doping is employed to achieve adequate carrier densities in one region of the device which is physically separated from the source of the carriers, the ionized impurities.

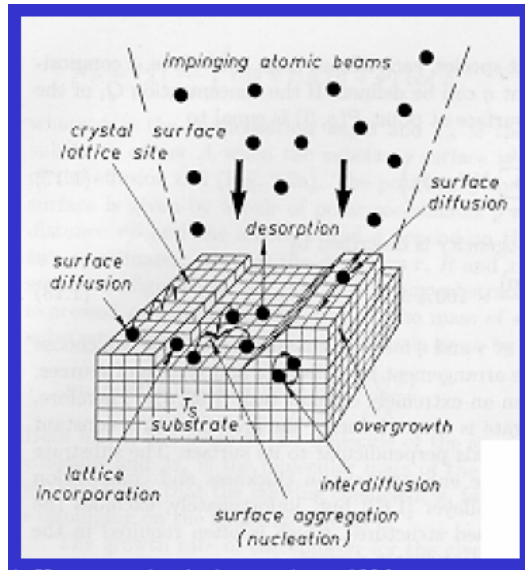


Figure 1-9. Molecular beam epitaxy process explained. After A. Herman.

Since many devices have to maintain the phase coherence of the electron wavefunction over the entire length of the device, there can be no inelastic scattering of the electrons. Thus, long mean free paths are crucial to the operation of such devices. The scattering of electrons by, for example, high background impurity or defect densities or rough interfaces would nullify the quantum phenomena. The evolution of high-purity MBE material has been the result of improvements in four major areas: (1) technologies for achieving ultra high vacuum; (2) application of superior materials for high temperature MBE system components; (3) identification and development of the optimum substrate preparation and epitaxial growth conditions, and (4) improvement in the purity of the substrate, source and crucible materials. The development of high purity GaAs/AlGaAs materials has been closely linked to the identification of residual impurities in these materials. The Hall mobility is a very sensitive qualitative measure of material purity at low temperatures where impurity scattering is dominant. As noted earlier, the approach to GaAs changed in 1978, when Dingle and co-workers demonstrated that very high mobilities could be achieved in modulation doped structures grown by MBE.

(A) Modulation doping of AlGaAs/GaAs heterostructures with in-plane transport

The low temperature mobility of modulation doped GaAs/AlGaAs heterostructures with in-plane transport is a good measure of the GaAs/AlGaAs quality. This depends very strongly on the epitaxial structure, particularly the placement and quantity of dopant impurities. The two-dimensional electron gas (2DEG) that exists at the interface between GaAs and the wider band gap AlGaAs exhibits a very high mobility at low temperatures. Even at room temperatures, the mobility is larger than that of bulk GaAs. Two factors contribute to this higher mobility, both arising from the selective doping of AlGaAs buffer layers rather than the GaAs layers in which the carriers reside. The first is the natural separation between the donor atoms in the AlGaAs and the electrons in the GaAs. The second is the inclusion of an undoped AlGaAs spacer layer in the structure. Such structures are quite complicated but can be easily fabricated using MBE techniques. A typical heterostructure begins with the bulk GaAs wafer upon which a GaAs buffer layer or super lattice is grown. The latter is used to act as a barrier to the out-diffusion of impurities and defects from the substrate. It also consists of a GaAs cap layer and alternating layers of AlGaAs and GaAs. The common practice is to use a doping for the AlGaAs layers (see Fig. 1-10 – top panel) in the active region but nowadays undoped AlGaAs layers are used and a delta doped layer is included (see Fig.

1-10 – bottom panel). This delta doped layer along with the growth of superlattices restricts the formation of defects, known as D – X centers [31], to a minimum. There are two important AlGaAs layers on either side of the δ -doped layer and they are called buffer and spacer layer, respectively. The spacer layer is closer to the GaAs quantum well and is of high purity to prevent scattering of the channel carriers by the ionized impurities. A usual practice is to use undoped AlGaAs layers to have very good confinement of the charge carriers in the well.

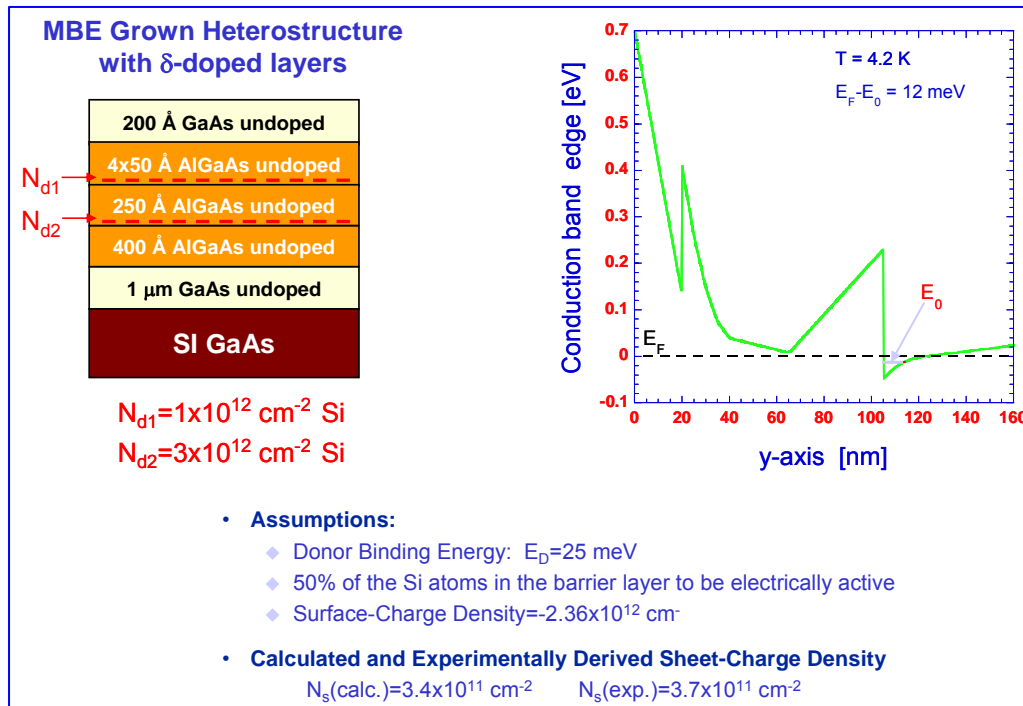
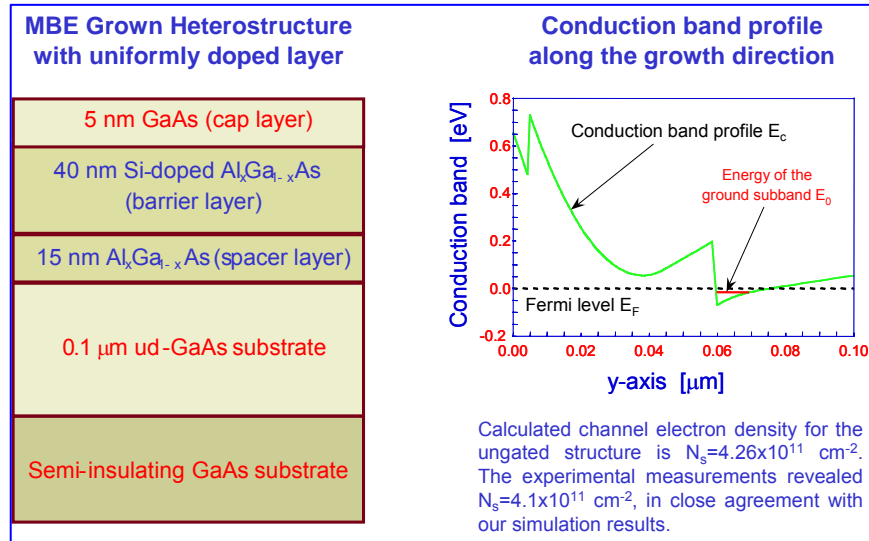


Figure 1-10. Top panel – uniformly doped barrier layer. Bottom panel – delta doped barrier layer.

Other device parameters that have to be considered are the composition of the Aluminum in AlGaAs. There is a compromise in the value chosen for x : if x is smaller than 0.2 then the band discontinuity will be too small to properly confine carriers in the well; if x is too large then defects, termed as D-X centers, tend

to appear in the AlGaAs. To overcome this problem Aluminum content is limited to about 20% and other variations like δ -doping layer and growth of superlattices are introduced into the MBE techniques. Examples of devices that utilize modulation doping are high-electron mobility transistor HEMT [32] (see Figure 1-11) in which size-quantization effects must be taken into account for proper description of size quantization. For very short devices this necessitates the use of some of the quantum transport approaches discussed in Section 3 of this review article.

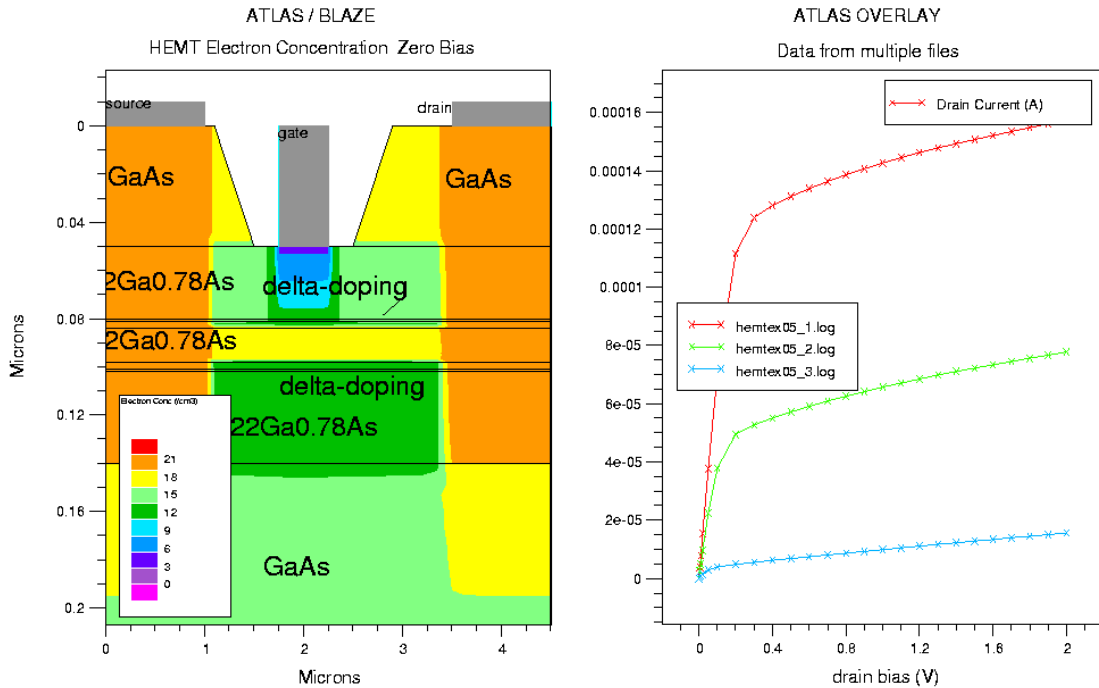


Figure 1-11. Left panel – pseudomorphic HEMT. Right panel – Typical IV characteristics of a HEMT device.

(B) Vertical transport - Resonant Tunneling Devices

Over the past three decades, resonant tunneling diodes (RTD's) have received a great deal of attention following the pioneering work by Esaki and Tsu [33]. Significant accomplishments have been achieved in terms of RTD device physics (see Figure 1-12), modeling, fabrication technology, and circuit design and applications. The RTD has been widely studied, and well over a thousand research papers have been written on various aspects of this seemingly simple device. Yet, whether RTD's will find their way into mainstream electronics in the future remains inconclusive. The research is ongoing and, in some areas, very active.

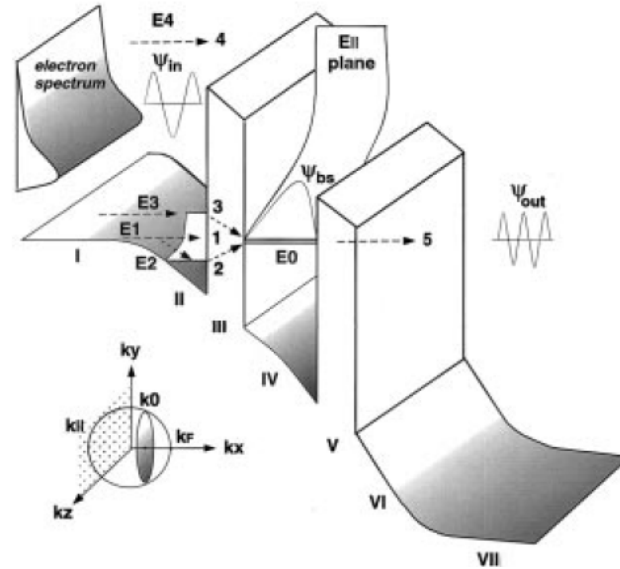


Figure 1-12. Schematics of the conduction band profile and the physical processes occurring in resonant tunneling diode (double barrier structure).

It is well documented that today's advanced information technology is mainly attributed to the electronic representation and processing of information in a low-cost, high speed, very compact, and highly reliable fashion, and that the quest and accomplishments of continual miniaturization and integration of solid-state electronics have been the key to the success of the computer industry and computer applications. The advanced multimedia infrastructure and services in the future will demand further reduction in chip size. Chip density, represented by memory technology, has been following Moore's law and has roughly doubled every other year over the last three decades. The trend remains strong and definite, at least for the foreseeable future. For example, a $0.15 \mu\text{m}$ process technology has been implemented in the first 4-Gb dynamic random access memory (DRAM) unveiled in 1997, and the feature size of DRAM transistors is projected to be $0.10 \mu\text{m}$ (16 Gb) in 2007, and $0.07 \mu\text{m}$ (64 Gb) by 2010 [3]. A natural and realistic question, then, is whether this desired trend will continue indefinitely. While an ultimate limit on the downscaling of conventional transistors and integrated circuits (IC's) will eventually be reached, device physicists and IC engineers have pondered answers, both evolutionary and revolutionary, to the challenge. While the downscaling of conventional transistors enjoys an exceptional, rapid evolution, revolutionary device concepts have been actively sought, particularly in the two related areas known as nanoelectronics and single electronics [34].

The idea of nanoelectronics was popularized in the mid-1980's, when pioneering work on resonant tunneling and bandgap engineering in low-dimensional semiconductor quantum wells and superlattices grew and was championed by several groups for the exploration of new opportunities for circumventing the limit on the downscaling of conventional transistors and IC's. For example, in the early 1980's Bob Bate analyzed the trends of the semiconductor industry and concluded that new device concepts based upon quantum effects would be required to keep the exponential growth trends going into the 21st century. At about this time, Gerry Sollner and co-workers [35] at MIT-Lincoln Lab reported the first quantum-well resonant-tunneling diode (RTD) with respectable performance. The RTD was quickly adopted as the prototypical quantum semiconductor device. Since then, the RTD, and its several variations, has become a research focus in nanoelectronics for its promise as a primary nanoelectronic device for both analog and digital applications.

For device realization, nanofabrication technology has made impressive advances during the last decade by routinely producing artificial semiconductor structures using molecular-beam epitaxy, metal-organic chemical vapor deposition, and chemical-beam epitaxy. Accurately controlled feature sizes as small as monolayers of atoms in the growth direction for dissimilar semiconductor materials, or heterostructure systems, have been achieved. Nanoscale lithography and patterning by electron-beam lithography have also been highly developed in the direction perpendicular to the growth direction. Although further improvements in this area call for more precise control, better resolution, and improved

interfaces, recent advances in nanofabrication technology have brought quantum effect device concepts to reality and have presented a great challenge for device physicists in the theoretical analysis of nanoelectronic devices.

Continuing effort in quantum transport modeling of vertical transport in RTD's is motivated by the need to understand device operation and to provide a primary test for developing theoretical tools for nanoelectronic devices. Not surprisingly, this is very different from traditional device modeling. Moreover, it provides valuable knowledge of the quantum aspects of electron transport in mesoscopic systems. Since the useful device properties, e.g., fast switching operation between ON and OFF states, are a consequence of the desired and controlled electron motion in the device, it is essential for device designers to understand and quantify the transport processes. Among the numerous nanoelectronic devices proposed and demonstrated, the RTD is perhaps the most promising candidate for digital circuit applications due to its negative differential resistance (NDR) characteristic, structural simplicity, relative ease of fabrication, inherent high speed, flexible design freedom, and versatile circuit functionality. There is a good practical reason to believe that RTD's may be the next device based on quantum confined heterostructures to make the transition from the world of research into practical application. Progress in epitaxial growth has improved the peak-to-valley current ratio at room temperature even beyond that required for many circuit applications. This temperature requirement is the single most important feature that any new technology must satisfy. It is what distinguishes the RTD from other interesting quantum device concepts that have been proposed but that show weak, if any, desired phenomena at room temperature.

A variety of circuit functions has already been demonstrated, providing proof-of-concept of proposed applications. The main issue at present is not, in fact, the RTD performance itself but the monolithic integration of RTD's with transistors [high electron mobility transistors (HEMT's) or heterojunction bipolar transistors (HBT's)] into integrated circuits with useful numbers and density of devices. Major challenges include the variation in the current-voltage characteristic of the RTD's across a wafer and from wafer to wafer, fabrication-dependent parasitic impedances, and edge effects as the RTD mesa area is decreased in order to reduce the intrinsic parasitic impedances and to achieve higher integration levels. Recently developed techniques for providing feedback during epitaxial growth via optical and photoemission probes have greatly improved the situation as far as uniformity of growth is concerned. It is for these reasons that RTD research has been sustained for more than two decades and may now be rapidly approaching the stage of technology implementation.

1.1.3 Modeling of nanoscale devices

Standard sequence that one follows when modeling device structures of interest involves (1) process simulation step that is followed by a (2) device simulation and is finalized with a (3) circuit simulation step. In this regard, device simulation is the process of using computers to calculate the behavior of electronic devices, i.e. of calculating the current-voltage (I - V) curves of a device in general. The devices are defined mathematically in terms of their dimension, material composition, and other relevant physical information, all of which is obtained from the process simulation step.

There are two issues that make simulation important. Product cycles are getting shorter with each generation, and the demand for production wafers shadows development efforts in the factory. Consider the product cycle issue first. In order for companies to maintain their competitive edge, products have to be taken from design to production in less than 18 months. As a result, the development phase of the cycle is getting shorter. Contrast this requirement with the fact that it takes 2-3 months to run a wafer lot through a factory, depending on its complexity. The specifications for experiments run through the factory must be near the final solution. While simulations may not be completely predictive, they provide a good initial guess. This can ultimately reduce the number of iterations during the device development phase. The second issue that reinforces the need for simulation is the production pressures that factories face. In order to meet customer demand, development factories are making way for production space. It is also expensive to run experiments through a production facility. The resources could have otherwise been used to produce sellable product. Again, device simulation can be used to decrease the number of experiments run through a factory. Device simulation can be used as a tool to guide manufacturing down the right path, thereby decreasing the development time and costs (see Figure 1-13). Besides offering the possibility to test hypothetical devices which have not (or could not) yet been manufactured, device simulation offers unique insight into device behavior by allowing the observation of phenomena that can not be measured on real devices. It is related to, but usually separate from process simulation, which deals with various physical processes such as material growth, oxidation, impurity diffusion, etching, and metal deposition inherent in

device fabrication leading to integrated circuits. Device simulation is distinct from another important aspect of computer-aided design (CAD), device modeling, which deals with compact behavioral models for devices and sub-circuits relevant for circuit simulation in commercial packages such as SPICE.

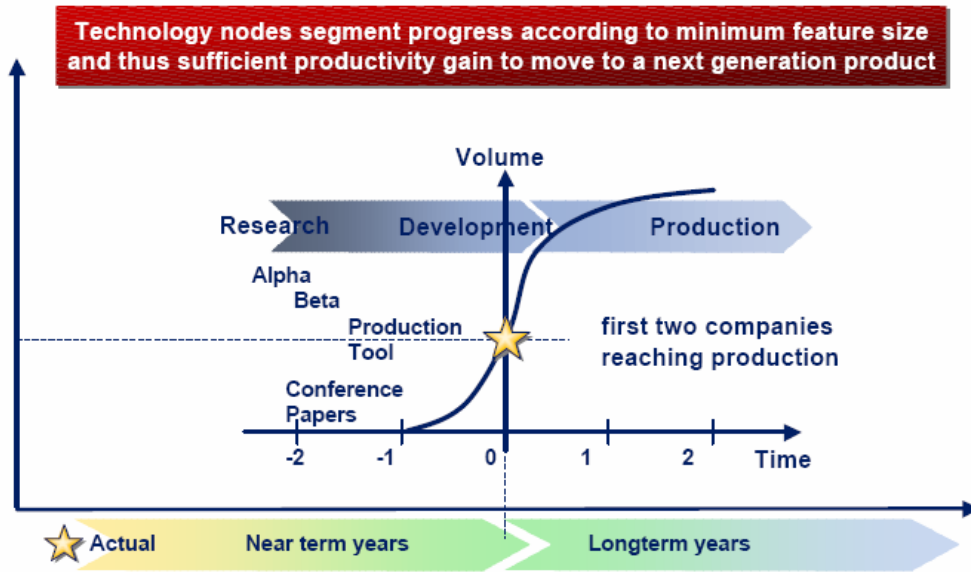


Figure 1-13. ITRS Technology node and related timeline.

The main components of semiconductor device simulation at any level of approximation are illustrated in Figure 1-14 [36]. There are two main kernels, which must be solved self-consistently with one another, the *transport equations* governing charge flow, and the *fields* driving charge flow. Both are coupled strongly to one another, and hence must be solved simultaneously. The fields arise from external sources, as well as the charge and current densities which act as sources for the time varying electric and magnetic fields obtained from the solution of Maxwell's equations. Under appropriate conditions, only the quasi-static electric fields arising from the solution of Poisson's equation are necessary. The fields, in turn, are driving forces for charge transport as illustrated in Figure 1-15 for the various levels of approximation within a hierarchical structure ranging from compact modeling at the top to an exact quantum mechanical description at the bottom.

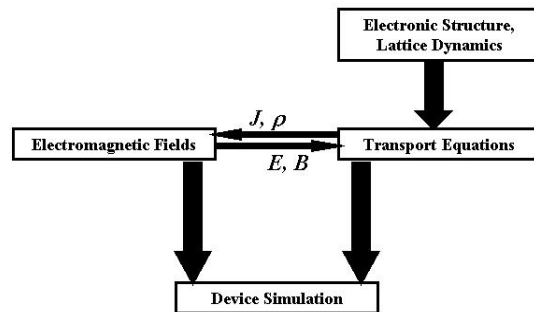


Figure 1-14. A schematic description of the device simulation sequence.

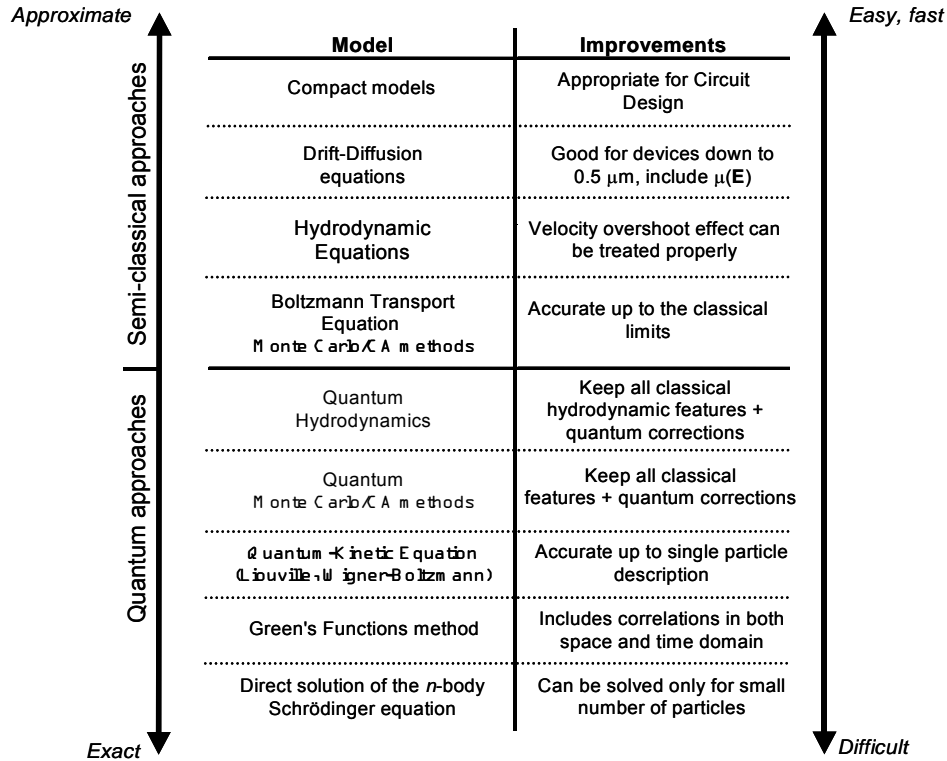


Figure 1-15. Illustration of the hierarchy of transport models.

Note that semiclassical Boltzmann transport has been the mainstay of the semiconductor technology from its early development. Up until now, most device simulations including the full-band Monte-Carlo (FBMC) method are based on the solution of the Boltzmann transport equation (BTE) and its simplifications, the hydrodynamic (HD) transport equations and the drift-diffusion (DD) model. But in the last decade, as semiconductor technology has continued to pursue the down scaling of device dimensions into the nanoscale regime, many new and interesting questions have emerged concerning the physics of small devices. Ref. [36] highlights some of the basic physical effects that are viewed as important in nanoelectronics research.

Table 1-3. Important effects in ultrasmall electronics.

1.	Transport Effects
	(a) Drift
	<ul style="list-style-type: none"> • Velocity overshoot • Ballistic transport • Oxide polar optical phonons decreasing channel mobility • Hot-electron effects (scattering in high electric field, injection into oxide) • Hot-phonon effects
	(b) Diffusion
	<ul style="list-style-type: none"> • Hot-electron diffusion (invalidation of Einstein relation) • Anisotropy of diffusion • Diffusion and reduced dimensionality
<hr/>	
2.	Size Effects
	<ul style="list-style-type: none"> • Spatial quantization (one- and two-dimensional electron) • Quantum resonances—surface plasmons, phonons • Interfaces, surfaces, metal boundaries (influence of these boundaries on important semiconductor parameters)

-
3. Environmental Effects
 - Low-level radiation effects (α -particles from IC packages, cosmic rays)
 - Synergetic effects
 - Remote polar scattering
 - Parasitic and interconnect factors, model contacts

 4. Generation-Recombination Effects
 - Hot-electron thermionic emissions
 - G-R noise for non-stationary transport
 - Impact ionization effects
-
5. Solid State Physics/Electronics
 - Nonlinear response theory
 - Reexamine effective mass theory
 - Statistical mechanics of the finite Fermi systems
 - Electron-phonon interactions with confined phonons
 - Long-range Coulomb potential
 - Quantum transport
 - Interface physics modeling
 - Low-dimensional effects
-

Note that the semiclassical device models become invalid in the nanometer regime since the classical transport concepts treat electrons (and holes) as particles, and the transport parameters are defined by taking ensemble averages of the particle motion. In nanostructures, a device theory that can properly treat quantum transport phenomena between macroscopic and microscopic scales is therefore needed. Work in this field has invoked mesoscopic physical principles and hierarchical quantum device models for formulating and quantifying nanoelectronic transport and device analysis. However, the available nanoelectronic device models have not been well established to the standards of conventional device models. The goal of this review article is to represent the latest in modeling quantum transport in a variety of device structures including RTD's and FinFETs using most suitable approaches for the problem at hand.

1.2 Open Systems

A general feature of electron devices is that they are of use only when connected to a circuit, and to be so connected any device must possess at least two terminals, contacts, or leads. As a consequence, every device is an open system with respect to carrier flow [37]. This is the overriding fact that determines which theoretical models and techniques may be appropriately applied to the study of quantum devices. For example, the quantum mechanics of pure, normalizable states, such as those employed in atomic physics, does not contribute significantly to an understanding of devices, because such states describe closed systems.

To understand devices, one must consider the un-normalizable scattering states, and/or describe the state of the device in terms of statistically mixed states, which casts the problem in terms of quantum kinetic theory. As a practical matter of fact, a device is of use only when its state is driven far from thermodynamic equilibrium by the action of the external circuit. The non-equilibrium state is characterized by the conduction of significant current through the device and/or the appearance of a non-negligible voltage drop across the device.

In classical transport theory, the openness of the device is addressed by the definition of appropriate boundary conditions for the differential (or integro-differential) transport equations. Such boundary conditions are formulated so as to approximate the behavior of the physical contacts to the device, typically Ohmic or Schottky contacts [38]. In the traditional treatments of quantum transport theories, the role of boundary conditions is often taken for granted, as the models are constructed upon an unbounded spatial domain. The proper formulation and interpretation of the boundary conditions remains an issue, however, and will be examined in the present work. It should be understood that, unless otherwise specified, all models to be considered here are based upon a single-band, effective-mass Schrodinger equation.

1.2.1 Tunneling Theory

The simplest model of quantum transport in devices is to describe the problem in terms of the scattering of the electron wavefunction by a spatially varying potential. One assumes that this potential is situated between two electron reservoirs, each of which emits particles with an equilibrium distribution into the scattering region. The reservoirs will, in general, have different chemical potentials, their difference representing an applied bias voltage. The net flux of electrons passing between the reservoirs constitutes the electrical current conducted by the device. A single-particle Schrodinger equation can only describe a situation in which the electrons move perfectly coherently throughout the device. Any loss of coherence due to inelastic collisions requires a higher-level description. Nevertheless, the solutions of Schrodinger's equation remain one of the fundamental tools available to understand and predict the behavior of quantum-scale devices.

In this section we will first show that all particles with $E < V_0$ incident on a potential step of height V_0 are reflected. In other words, although the quantum-mechanical treatment predicts penetration of the wavefunction into the classically-forbidden region (the probability density in this region is non-zero), we can not demonstrate this phenomenon in a laboratory. However, one can demonstrate the penetration into a classical-forbidden region if we chop-off the potential step, changing it into a barrier of width L . Then, according to quantum physics, particles with $E < V_0$ incident on the barrier from the left, will penetrate the potential and will be transmitted into the classically-allowed region where we can detect them. This wholly non-classical phenomenon is called TUNNELING (see Figure 1-16) [21]. The two remarkable applications of tunneling are:

- (a) Resonant tunneling diodes, which are used as switching units in fast electronic circuits.
- (b) Scanning tunneling microscope (STM), based on the penetration of electrons near the surface of a solid sample through the barrier at the surface (see Figure 1-17). These electrons form a "cloud" of probability outside the sample. Although the probability of detecting one of these electrons decays exponentially with distance (from the surface), one can induce and measure a current of these electrons and attain a magnification factor of 100 million - large enough to permit resolution of a few hundredths the size of an atom. Gerd Binnig and Heinrich Rohrer won the Noble Prize in Physics in 1986 for the invention of the STM [39].

Quantum Tunneling

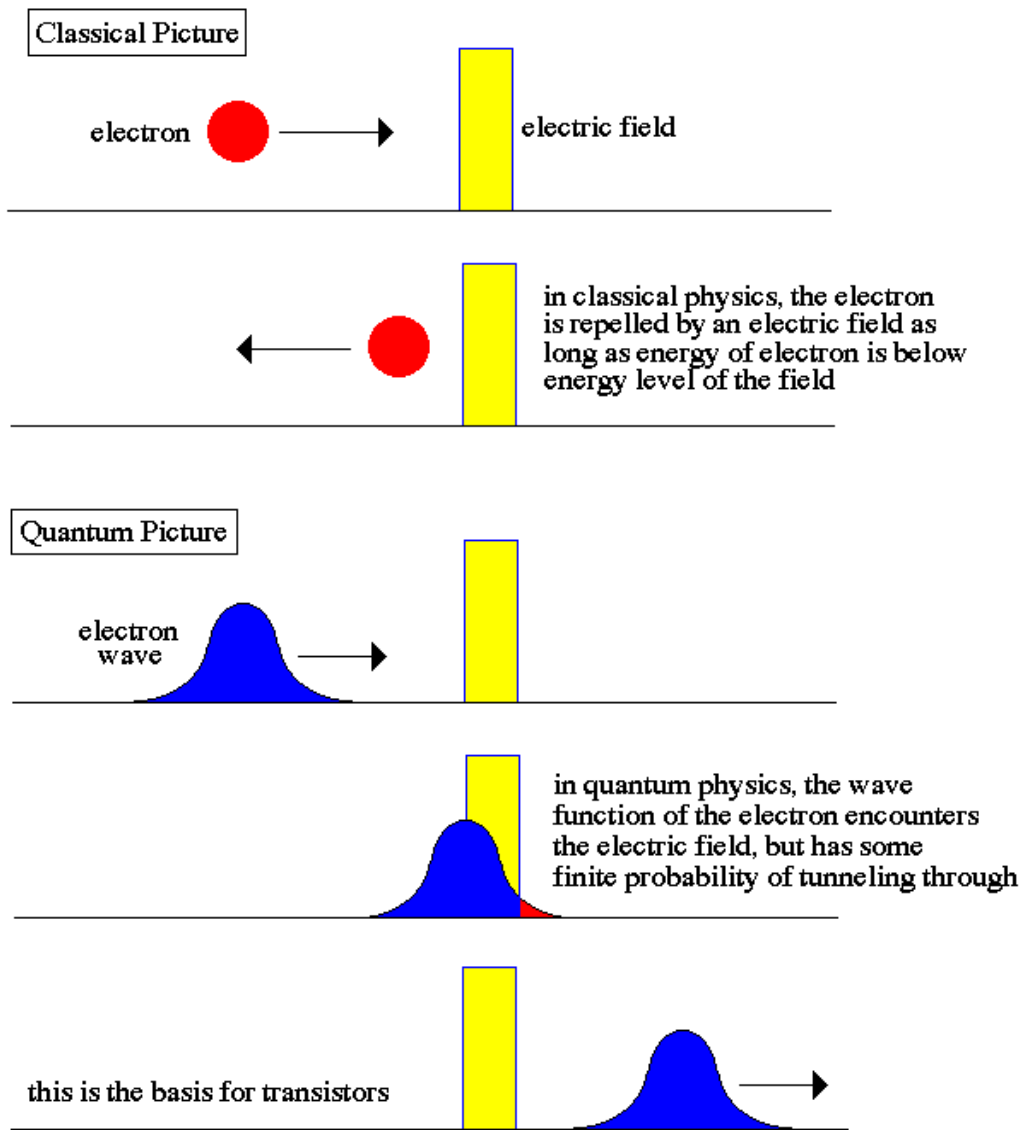


Figure 1-16. Quantum-mechanical tunneling.

revolution of tunnelling: **Scanning Tunnelling Microscope**

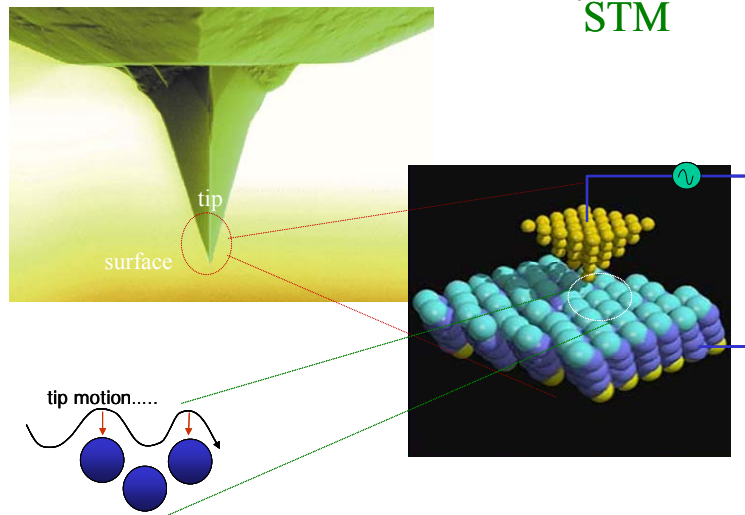


Figure 1-17. Scanning tunneling microscope operation based on quantum-mechanical tunneling.

In the rest of this section, we will first describe the tunneling phenomenon through the example of a single barrier. Then, we will talk about multiple barrier case formulation, the double barrier being a special case which we have in resonant tunneling diodes.

(A) General Notation

Quantum physicists are interested in all kinds of physical systems (photons, conduction electrons in metals and semiconductors, atoms, etc.). States of these rather diverse systems are represented by the same type of functions, the state functions [40]. The **first postulate of quantum mechanics** states that *every physically-realizable state of the system is described in quantum mechanics by a state function ψ that contains all accessible physical information about the system in that state*. Physically realizable states are states that can be studied in laboratory, accessible information is the information we can extract from the wavefunction, and the state function is function of position, momentum and energy that is spatially localized. Also, if ψ_1 and ψ_2 represent two physically-realizable states of the system, then the linear combination

$$\Psi = c_1\psi_1 + c_2\psi_2, \tag{1-1}$$

where c_1 and c_2 are arbitrary complex constants, represents a third physically realizable state of the system.

Quantum mechanics describes the outcome of an ensemble of measurements, where an ensemble of measurements consists of a very large number of identical experiments performed on identical non-interacting systems, all of which have been identically prepared so as to be in the same state. This brings us to the **second postulate of quantum mechanics** that states: *If a system is in a quantum state represented by a wavefunction ψ , then*

$$PdV = |\psi|^2 dV \tag{1-2}$$

is the probability that in a position measurement at time t the particle will be detected in the infinitesimal volume dV . Note that $|\psi(x, t)|^2$ is the position and time probability density. The importance of normalization follows from the Born interpretation of the state function as a position probability amplitude. According to the second postulate of quantum mechanics, the integrated probability density can be interpreted as a probability that in a position measurement at time t , we will find the particle anywhere in space. Therefore, the normalization condition for the wavefunction is:

$$\int P dV = \int |\psi(x, y, z)|^2 dV = \int \psi^*(x, y, z)\psi(x, y, z)dV = 1 \quad (1-3)$$

There are several limitations on the wavefunction that arise from this normalization condition: (1) Only normalizable functions can represent a quantum state and these are called physically admissible functions. (2) State function must be continuous and single valued function. (3) State function must be a smoothly-varying function (continuous derivative).

The Born interpretation of quantum mechanics enables us to determine from a wavefunction, the probabilistic information. For example, we can answer the following question: In an ensemble measurement of position at time t , what is the probability that a member of the ensemble will exhibit a value in the range from x to $x+dx$? To characterize the results of an experiment, we use two statistical quantities: ensemble average $\langle x \rangle$ and standard deviation Δx . In quantum theory, the ensemble average of an observable for a particular state of the system is called the expectation value of that observable that is calculated using

$$\langle x \rangle = \int xP(x, t)dx = \int \psi^*(x, t)x\psi(x, t)dx = (\psi, x\psi) \quad (1-4)$$

Note that the expectation value can be time dependent, i.e. $\langle x \rangle = \langle x(t) \rangle$. Also, the expectation value depends upon the state of the system. Different states, represented by different state functions have different ensemble averages. Let's denote a generic observable $Q(x)$ that depends only upon position. The expectation value of this observable is given by:

$$\langle Q(t) \rangle = \int \psi^*(x, t)Q(x)\psi(x, t)dx \quad (1-5)$$

The other statistical quantity that one uses in quantum physics is the standard deviation of an observable - otherwise known as uncertainty. For a position measurement, the uncertainty in x answers the following question: In an ensemble measurement at time t of the position of a particle in a state $\psi(x, t)$, what is the spread of the individual results around the expectation value $\langle x \rangle$? To answer this question, one needs to calculate the dispersion:

$$\Delta x^2 = \int \psi^*(x, t)[x - \langle x \rangle]^2 \psi(x, t)dx = \langle x^2 \rangle - \langle x \rangle^2 \quad (1-6)$$

The uncertainty, or the standard deviation is given by:

$$\Delta x = \sqrt{\langle x^2 \rangle - \langle x \rangle^2}, \quad (1-7)$$

or the uncertainty equals the square-root of the dispersion. In general, the uncertainty in the measurement of the observable $Q(x)$ is given by:

$$\Delta Q = \sqrt{\langle Q^2 \rangle - \langle Q \rangle^2} \quad (1-8)$$

We have already stated that $\psi(x, t)$ is the state function of a system in the position representation. It must be normalizable since it describes a localized particle. Therefore, one can define a Fourier transform of this function:

$$\psi(x, 0) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{\infty} dk e^{ikx} \phi(k), \quad (1-9)$$

where $\phi(k)$ is the Fourier coefficient, or in this particular case it represents the momentum wavefunction also known as the amplitude function. The description of the particle in momentum state is achieved using these momentum wavefunctions. In summary, the position probability amplitude $\psi(x, t)$ and the momentum state function $\phi(k)$ are state descriptors for a microscopic system. The second class of basic elements are the observables, which represent the physical attributes of a system that can be measured in a laboratory. Examples of observables are position $x(t)$, momentum $p(t)$ and energy $E(x(t))$. The problem of applying a classical definition of observables for quantum states is that it is impossible to measure the properties of a microscopic system without altering its state. **The third postulate of quantum mechanics** states that: *Every observable in quantum mechanics is represented by an operator which is used to obtain physical information about the observable from the state function.* For an observable that is represented in classical physics by a function $Q(x, p)$, the corresponding operator is $Q(\hat{x}, \hat{p})$ [41].

Table 1-4. Most important operators in quantum Physics.

Observable	Operator
Position	\hat{x}
Momentum	$\hat{p} = \frac{\hbar}{i} \frac{\partial}{\partial x}$
Energy	$E = \frac{\hat{p}^2}{2m} + V(\hat{x}) = -\frac{\hbar^2}{2m} \frac{\partial^2}{\partial x^2} + V(x)$

An operator (see Table 1-4) is an instruction, a symbol which tells us to perform one or more mathematical acts on a function, say $f(x)$. The essential point is that they act on a function. Operators act on everything to the right, unless the action is constrained by brackets. The addition and subtraction rule for operators reads:

$$(\hat{Q}_1 \pm \hat{Q}_2)f(x) = \hat{Q}_1 f(x) \pm \hat{Q}_2 f(x) \quad (1-10)$$

The product of two operators implies successive operation:

$$\hat{Q}_1 \hat{Q}_2 f(x) = \hat{Q}_1 [\hat{Q}_2 f(x)] \quad (1-11)$$

The product of two operators is a third operator:

$$\hat{Q}_3 = \hat{Q}_1 \hat{Q}_2 \quad (1-12)$$

Two operators commute if they obey the simple operator expression:

$$[\hat{Q}_1, \hat{Q}_2] = \hat{Q}_1 \hat{Q}_2 - \hat{Q}_2 \hat{Q}_1 = 0 \Rightarrow \hat{Q}_1 \hat{Q}_2 = \hat{Q}_2 \hat{Q}_1 \quad (1-13)$$

The requirement for two operators to be commuting operators is a very important one in quantum mechanics and it means that we can simultaneously measure the observables represented with these two operators. The non-commutivity of the position and the momentum operators (the inability to simultaneously determine particles position and its momentum) is represented with the Heisenberg uncertainty principle, which in mathematical form is expressed as:

$$\Delta x \cdot \Delta p \geq \frac{\hbar}{2} = \frac{1}{2} |[\hat{x}, \hat{p}]|, \quad (1-14)$$

and can be generalized for any pair of observables.

In 1926 Erwin Schrödinger [42] proposed an equation that describes the evolution of a quantum-mechanical system (which represents quantum equations of motion), and is of the form:

$$-\frac{\hbar^2}{2m} \frac{\partial^2 \psi}{\partial x^2} + V(x)\psi(x, t) = \left[-\frac{\hbar^2}{2m} \frac{\partial^2}{\partial x^2} + V(x) \right] \psi(x, t) = i\hbar \frac{\partial \psi}{\partial t} \quad (1-15)$$

This work of Schrödinger was stimulated by a 1925 paper by Einstein on the quantum theory of ideal gas [43], and the de Broglie theory of matter waves [44]. Examining the time-dependent SWE, one can also define the following operator for the total energy:

$$\hat{E} = i\hbar \frac{\partial}{\partial t} \quad (1-16)$$

The introduction of the Schrodinger equation brings us to the **fourth fundamental postulate of quantum mechanics** which states that *the time development of the state functions of an isolated quantum system is governed by the time-dependent SWE $\hat{H}\psi = i\hbar\partial\psi / \partial t$, where $\hat{H} = \hat{T} + \hat{V}$ is the Hamiltonian of the*

system. The time-dependent Schrödinger wave equation (TDSWE) describes the evolution of a state provided that no observations are made. An observation alters the state of the observed system, and as it is, the TDSWE can not describe such changes.

The solution of the TDSE is a rather formidable problem even in 1D. The underlying problem is not just that it is a partial differential equation of second order in x and first order in t , but that we must consider both of these variables at once. As a consequence, the space and time dependence of the wavefunction may be very complicated. One way to solve the TDSE is to seek solutions that have a particularly simple form, i.e.

$$\Psi(x, t) = \psi(x)\xi(t) \quad (1-17)$$

These product functions are called separable solutions of the partial differential equation (PDE). In Quantum Mechanics, they are called stationary-state wavefunctions [45]. One can find these stationary-state wavefunctions using a method called separation of variables. For example, substituting the above expression for $\Psi(x, t)$ into the TDSE, one gets:

$$\xi(t) \left[-\frac{\hbar^2}{2m} \frac{\partial^2 \psi(x)}{\partial x^2} + V(x, t)\psi(x) \right] = \psi(x) i\hbar \frac{\partial \xi(t)}{\partial t} \quad (1-18)$$

or:

$$\frac{1}{\psi(x)} \left[-\frac{\hbar^2}{2m} \frac{\partial^2 \psi(x)}{\partial x^2} + V(x, t)\psi(x) \right] = i\hbar \frac{1}{\xi(t)} \frac{\partial \xi(t)}{\partial t} \quad (1-19)$$

The last equation is almost separable, since, in general, the potential energy can depend on both x and t . If we assume that $V(x, t) = V(x)$, i.e. the potential energy is time-independent, then the LHS is only a function of x and the RHS is only a function of t . Therefore, the two sides can be equal if they are constant. This gives us two equations:

$$\begin{cases} \frac{1}{\psi(x)} \left[-\frac{\hbar^2}{2m} \frac{\partial^2 \psi(x)}{\partial x^2} + V(x)\psi(x) \right] = \alpha \rightarrow -\frac{\hbar^2}{2m} \frac{\partial^2 \psi(x)}{\partial x^2} + V(x)\psi(x) = \alpha\psi(x) \\ i\hbar \frac{1}{\xi(t)} \frac{\partial \xi(t)}{\partial t} = \alpha \rightarrow i\hbar \frac{\partial \xi(t)}{\partial t} = \alpha\xi(t) \end{cases} \quad (1-20)$$

The fact that we can obtain these two equations at all, provided that V does not depend on time, proves that stationary state wavefunctions exist for systems with a time-independent potential energy. A system whose potential energy is time-independent is said to be conservative. Hence, stationary states exist for conservative systems.

Now, the solution of the second equation [for $\xi(t)$] is very simple, and is given by:

$$\xi(t) = \xi(0)e^{-i\alpha t/\hbar} \quad (1-21)$$

The Born interpretation of $\Psi(x, t)$ [46] provides a clue to the physical meaning of the separation constant α . The wavefunction of a stationary state oscillates at a frequency $\omega = \alpha/\hbar$. From the de Broglie-Einstein relation, one has that $E = \hbar\omega = \alpha$. Hence, we can represent α as the total energy of the particle in the state represented with this wavefunction, i.e.

$$\xi(t) = \xi(0)e^{-iEt/\hbar} \quad (1-22)$$

The first equation then becomes:

$$\left[-\frac{\hbar^2}{2m} \frac{\partial^2}{\partial x^2} + V(x) \right] \psi_E(x) = E\psi_E(x) \quad (1-23)$$

This is the time-independent SE (TISE), which is an equation of a very special form. There is a differential operator representing the total energy of the system, which operates on the unknown function $\psi_E(x)$ to produce the same function $\psi_E(x)$ multiplied by a parameter E . Mathematicians call an equation of this form an eigenvalue equation where $\psi_E(x)$ is the eigenfunction and E is the corresponding eigenvalue.

Thus, the results presented in this section can be summarized as follows: (1) If a microscopic system is conservative, then there exist special quantum states of the system, called stationary states, in which the energy is sharp. (2) Even if the number of these eigenstates is infinite, the energies of the bound states form a discrete list. (3) If there is a one-to-one correspondence between the quantized energies of a quantum system and its bound state, or stationary-state wavefunctions, then the bound state energy is non-degenerate. If there are stationary states for which there correspond more than one distinct spatial functions, such bound states are called degenerate.

(B) Stationary States for a Free Particle

Let's try to solve the TISE for a free particle, for which $V(x) = 0$, i.e.

$$-\frac{\hbar^2}{2m} \frac{\partial^2 \psi(x)}{\partial x^2} = E\psi(x) \rightarrow \frac{\hbar^2}{2m} \frac{\partial^2 \psi(x)}{\partial x^2} + E\psi(x) = 0 \quad (1-24)$$

This is a homogeneous, second-order partial differential equation with constant coefficients, and the solution of this equation is of the form:

$$\psi(x) = Ae^{ikx} + Be^{-ikx}, \quad k = \sqrt{\frac{2mE}{\hbar^2}} \quad (1-25)$$

This function is not normalizable, since it does not decay for $|x| \rightarrow \infty$. Two properties follow from the impossibility of normalizing this function: (1) The energies are not quantized, i.e. all values $E > 0$ are allowed. (2) The energies are degenerate ($+k$ and $-k$).

(C) Potential Step

Our next task is to solve the TISE for one-dimensional single-particle system whose potential energy is piecewise constant [47]. A piecewise constant potential is one that is constant for all values of x except at a finite number of discontinuities-points, where it changes from one constant value to another. One of the simplest piecewise-constant potentials is the potential step

$$V(x) = \begin{cases} 0, & x < 0 \\ V_0, & x \geq 0 \end{cases} \rightarrow V(x) = V_0 \Theta(x) \quad (1-26)$$

shown graphically in Figure 1-18.

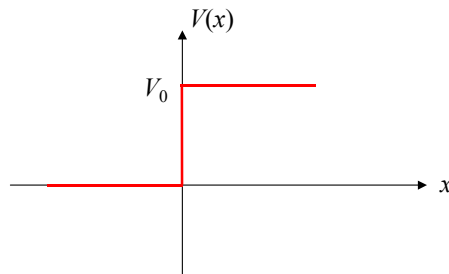


Figure 1-18. Potential step (an example of a piecewise constant potential).

Our task is to solve the following problem: We assume that we have a beam of particles incident from the left that encounters an impulsive force at $x=0$. We will consider two cases: Case (a) when the energy E of the particle is less than V_0 , and Case (b) when the energy of the particle is greater than V_0 .

Case (a): $E < V_0$

For this particular case, we need to distinguish between two separate regions:

- On the left from the potential step, $E > V(x) \rightarrow$ classically-allowed region
- On the right from the potential step, $E < V(x) \rightarrow$ classically-forbidden region
- $x=0$ is called a turning point \rightarrow point that separates the classically-allowed from the classically forbidden regions (see the figure below).

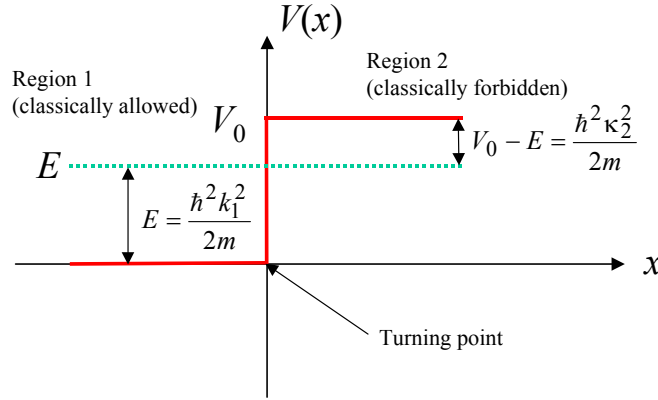


Figure 1-19 Description of the various regions for case (a).

The solution procedure for this type of problems is the following one:

- Write down the TISE for each region in which the potential energy is constant. Solve the TISE up to arbitrary constants.
- If necessary, apply the asymptotic condition to get physically admissible state function.
- Match the wavefunctions and the derivatives of the wavefunctions at each of the turning points.

Following the above-outlined procedure, we get the following general expressions for the wavefunctions in region 1 ($x < 0$) and region 2 ($x > 0$):

$$\psi(x) = \begin{cases} A^{(1)} e^{ik_1 x} + B^{(1)} e^{-ik_1 x} = \psi_1(x), & k_1 = \sqrt{\frac{2mE}{\hbar^2}} \\ A^{(2)} e^{-\kappa_2 x} = \psi_2(x), & \kappa_2 = \sqrt{\frac{2m(V_0 - E)}{\hbar^2}} \end{cases} \quad (1-27)$$

The unknown coefficients that appear in the above expression are calculated from the requirement that $\psi(x)$ be continuous and smooth, which implies that it must also have continuous first derivative. Therefore, using the boundary conditions:

$$\begin{cases} \psi_1(0) = \psi_2(0) \\ \left. \frac{d\psi_1(x)}{dx} \right|_{x=0} = \left. \frac{d\psi_2(x)}{dx} \right|_{x=0} \end{cases} \quad (1-28)$$

we get:

$$\begin{cases} B^{(1)} = \frac{k_1 - i\kappa_2}{k_1 + i\kappa_2} A^{(1)} \\ A^{(2)} = \frac{2k_1}{k_1 + i\kappa_2} A^{(1)} \end{cases} \quad (1-29)$$

Some important conclusions that can be derived from the above equations are:

- $|B^{(1)}| = |A^{(1)}|$, which means that in region 1, two waves of equal amplitude travel in the opposite direction and with the same phase velocity. Because of this, the wavefunction in region 1 represents a standing wave.
- The above observation suggests that the probability of reflection $R(E)=1$ and that of transmission $T(E)=0$. This result is just what we would expect were we applying classical physics to the problem.
- In region 2, the wavefunction represents an evanescent wave, whose amplitude equals to

$$|\psi_2(x)| = \frac{2k_1}{\sqrt{k_1^2 + \kappa_2^2}} |A^{(1)}| e^{-\kappa_2 x} \quad (1-30)$$

This is consistent with the previous observation that all incident particles with $E < V_0$ are reflected back. Therefore, we might say that no probability flow is associated with evanescent waves.

Case (b): $E > V_0$

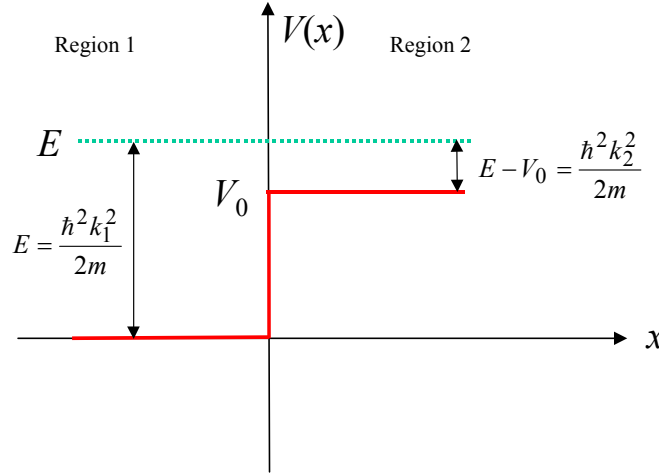


Figure 1-20. Graphical description of the variables used for this case.

Following the procedure described above, we get the following general expressions for the wavefunctions in region 1 and region 2:

$$\psi(x) = \begin{cases} A^{(1)} e^{ik_1 x} + B^{(1)} e^{-ik_1 x} = \psi_1(x), & k_1 = \sqrt{\frac{2mE}{\hbar^2}} \\ A^{(2)} e^{ik_2 x} = \psi_2(x), & k_2 = \sqrt{\frac{2m(E - V_0)}{\hbar^2}} \end{cases} \quad (1-31)$$

The unknown coefficients that appear in the above expression are again calculated from the requirement that $\psi(x)$ be continuous and smooth, which implies that it must also have continuous first derivative. Therefore, using the boundary conditions:

$$\begin{cases} \psi_1(0) = \psi_2(0) \\ \left. \frac{d\psi_1(x)}{dx} \right|_{x=0} = \left. \frac{d\psi_2(x)}{dx} \right|_{x=0} \end{cases} \quad (1-32)$$

in this particular case we obtain:

$$\begin{cases} B^{(1)} = \frac{k_1 - k_2}{k_1 + k_2} A^{(1)} \\ A^{(2)} = \frac{2k_1}{k_1 + k_2} A^{(1)} \end{cases} \quad (1-33)$$

Some important conclusions that can be derived from the above equation are:

- Since all coefficients are real and their magnitude is in general not equal to one, we might conclude that the wavefunction for continuum stationary states is of the following general form:

$$\psi(x) = \begin{cases} \text{incident wave} + \text{reflected wave} & \rightarrow \text{source} \\ \text{transmitted wave} & \rightarrow \text{detector} \end{cases}$$

We want to define the following two quantities:

- $\rho(E) = B^{(1)} / A^{(1)} \rightarrow$ reflection probability amplitude
- $\tau(E) = A^{(2)} / A^{(1)} \rightarrow$ transmission probability amplitude

Then, we can write:

$$\psi(x) = \begin{cases} A^{(1)} e^{ik_1 x} + \rho(E) A^{(1)} e^{-ik_1 x} = \psi_1(x), & x < 0 \\ \tau(E) A^{(1)} e^{ik_2 x} = \psi_2(x), & x \geq 0 \end{cases} \quad (1-34)$$

From the probability amplitudes, one can calculate the reflection and transmission coefficients. To do so, we return to the probability current densities - the fluxes - for the incident, reflected and transmitted waves. If we use the expression for the probability current, of the form

$$J(x) = -\frac{i\hbar}{2m} \left(\frac{\partial \psi^*}{\partial x} \psi - \psi^* \frac{\partial \psi}{\partial x} \right) \quad (1-35)$$

we arrive at the following results:

$$J_1 = -\frac{e\hbar k_1}{2m} \left[1 - |\rho(E)|^2 \right] |A^{(1)}|^2 = J_{inc} + J_{ref}, \quad x < 0 \quad (1-36)$$

$$J_2 = -\frac{e\hbar k_2}{2m} |\tau(E)|^2 |A^{(1)}|^2 = J_{trans}, \quad x \geq 0 \quad (1-37)$$

If we look at the result at region 1, the probability current density J_1 equals the sum of the incident current and a reflected current. In region 2, we have a transmitted current J_2 . We can, therefore, define a reflection probability $R(E)$ and transmission probability $T(E)$ in the following manner:

$$R(E) = \left| \frac{J_{ref}}{J_{inc}} \right| = \left| \frac{-\frac{q\hbar k_1}{m} |\rho(E)|^2 |A^{(1)}|^2}{-\frac{q\hbar k_1}{m} |A^{(1)}|^2} \right| = |\rho(E)|^2, \text{ and} \quad (1-38)$$

$$T(E) = \left| \frac{J_{trans}}{J_{inc}} \right| = \left| \frac{-\frac{q\hbar k_2}{m} |\tau(E)|^2 |A^{(1)}|^2}{-\frac{q\hbar k_1}{m} |A^{(1)}|^2} \right| = \frac{k_2}{k_1} |\tau(E)|^2. \quad (1-39)$$

Since a particle is either reflected or transmitted, we must have that

$$T(E) + R(E) = 1. \quad (1-40)$$

For our particular problem, this gives us:

$$R(E) = \left(\frac{k_1 - k_2}{k_1 + k_2} \right)^2 \text{ and } T(E) = \frac{4k_1 k_2}{(k_1 + k_2)^2}. \quad (1-41)$$

Therefore, in contrast to the case $E < V_0$, our result for $E > V_0$ does not conform to the predictions of classical physics.

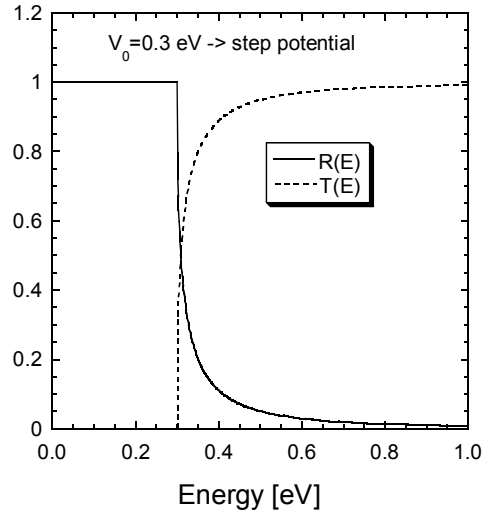


Figure 1-21 A step potential of height $V_0=0.3$ eV. The effective mass of the electrons is assumed to be $0.067 m_0$, where m_0 is the free electron mass. Transmission and reflection coefficient for a step potential as a function of the incident carrier energy.

An interesting thing to plot is the magnitude squared of the wavefunctions in regions (1) and (2) as a function of position. The behavior of $|\psi_i(x)|^2$ ($i=1$ for region 1 and $i=2$ for region 2) is shown on Figure 1-22. We use $E=0.25$ eV and $A^{(1)}=1$. Note that the energy of the particles E is smaller than the barrier height. Therefore, we expect to see standing wave pattern in region 1 [since $R(E)=1$ for this case] and evanescent (exponentially-decaying solution) in region 2.

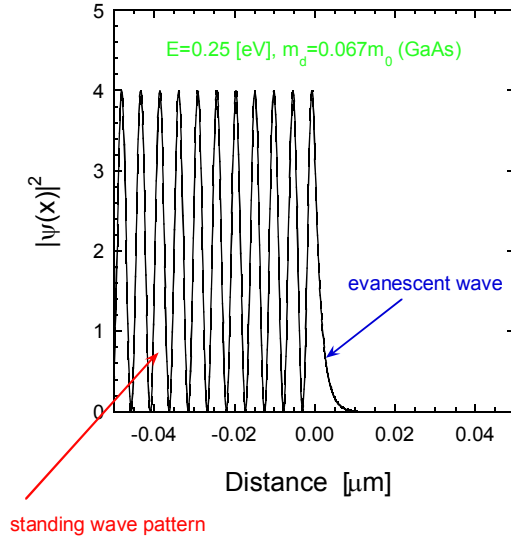


Figure 1-22. The magnitude squared of the wavefunction for carrier energy less than the barrier height.

(D) Tunneling Through a Single Barrier

Consider the potential barrier shown in Figure 1-23, for which the potential energy term appearing in the 1D TISE is of the form:

$$V(x) = \begin{cases} 0, & x < 0 \\ V_0, & 0 \leq x \leq L \\ 0, & x > L \end{cases} \quad (1-42)$$

Following the steps outlined in the previous section, it is easy to show that for energies $E < V_0$, the general solution of the 1D TISE in each of the three regions, is of the form:

$$\begin{aligned} \psi_1(x) &= Ae^{ikx} + Be^{-ikx} \\ \psi_2(x) &= Ce^{-\gamma x} + De^{\gamma x}, \text{ where } k = \sqrt{\frac{2mE}{\hbar^2}} \text{ and } \gamma = \sqrt{\frac{2m(V_0 - E)}{\hbar^2}}. \\ \psi_3(x) &= Ee^{ikx} + Fe^{-ikx} \end{aligned} \quad (1-43)$$

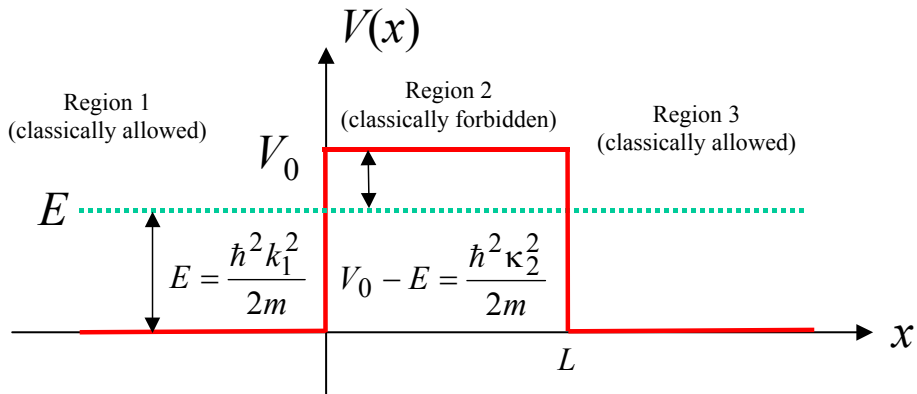


Figure 1-23. Single potential barrier.

The application of the continuity conditions of the wavefunction at the boundaries $x = 0$ and $x = L$, leads to the following relationship between the unknown constants:

$$\begin{aligned}
\psi_1(0) &= \psi_2(0) \rightarrow A + B = C + D \\
\dot{\psi}_1(0) &= \dot{\psi}_2(0) \rightarrow ik(A - B) = -\gamma(C - D) \\
\psi_2(L) &= \psi_3(L) \rightarrow Ce^{-\gamma L} + De^{\gamma L} = Ee^{ikL} + Fe^{-ikL} \\
\dot{\psi}_2(L) &= \dot{\psi}_3(L) \rightarrow -\gamma(Ce^{-\gamma L} - De^{\gamma L}) = ik(Ee^{ikL} - Fe^{-ikL})
\end{aligned} \tag{1-44}$$

Using the above four equations, we can find the relationships between various coefficients, i.e. using matrix representation these relationships can be represented as:

$$\begin{aligned}
\begin{bmatrix} A \\ B \end{bmatrix} &= \begin{bmatrix} \frac{1}{2} \left(1 + i \frac{\gamma}{k} \right) & \frac{1}{2} \left(1 - i \frac{\gamma}{k} \right) \\ \frac{1}{2} \left(1 - i \frac{\gamma}{k} \right) & \frac{1}{2} \left(1 + i \frac{\gamma}{k} \right) \end{bmatrix} \begin{bmatrix} C \\ D \end{bmatrix} = M_1 \begin{bmatrix} C \\ D \end{bmatrix} \\
\begin{bmatrix} C \\ D \end{bmatrix} &= \begin{bmatrix} \frac{1}{2} \left(1 - i \frac{k}{\gamma} \right) e^{(ik+\gamma)L} & \frac{1}{2} \left(1 + i \frac{k}{\gamma} \right) e^{-(ik-\gamma)L} \\ \frac{1}{2} \left(1 + i \frac{k}{\gamma} \right) e^{(ik-\gamma)L} & \frac{1}{2} \left(1 - i \frac{k}{\gamma} \right) e^{-(ik+\gamma)L} \end{bmatrix} \begin{bmatrix} E \\ F \end{bmatrix} = M_2 \begin{bmatrix} E \\ F \end{bmatrix}
\end{aligned} \tag{1-45}$$

In other words, we have the following relationship between the coefficients A and B , and the coefficients E and F :

$$\begin{bmatrix} A \\ B \end{bmatrix} = M_1 \begin{bmatrix} C \\ D \end{bmatrix} = M_1 M_2 \begin{bmatrix} E \\ F \end{bmatrix} = M \begin{bmatrix} E \\ F \end{bmatrix}, \tag{1-46}$$

where the matrix M has elements m_{ij} . Therefore, for coefficients A and E (using the asymptotic condition that $F=0$) we have the following simple relationship: $A=m_{11}E$, i.e. the transmission coefficient is simply given by:

$$T(E) = \left| \frac{E}{A} \right|^2 = \frac{1}{|m_{11}|^2}. \tag{1-47}$$

After a rather straightforward calculation, we arrive at the following expression for the transmission coefficient for particle energies less than the barrier height:

$$T(E) = \left[1 + \left(\frac{\gamma^2 + k^2}{2k\gamma} \right)^2 sh^2(\gamma L) \right]^{-1} \tag{1-48}$$

In the case of a weak barrier ($\gamma L \ll 1$), the expression for the transmission coefficient simplifies to:

$$T(E) \approx \frac{1}{1 + (kL/2)^2}. \tag{1-49}$$

In the opposite limit, i.e. when the barrier is very strong (γL is very large), we have the following approximate expression for the transmission coefficient:

$$T(E) \approx \left(\frac{4k\gamma}{k^2 + \gamma^2} \right)^2 \exp(-2\gamma L). \tag{1-50}$$

For energies larger than the barrier height, i.e. $E > V_0$, using that $\gamma = ik_2$, gives:

$$T(E) = \left[1 + \left(\frac{k^2 - k_2^2}{2kk_2} \right)^2 \sin^2(k_2 L) \right]^{-1} \quad (1-51)$$

The later result is similar to the one obtained in the previous section, i.e. the transmission maxima ($T(E)=1$) occur for $k_2 L = n\pi$.

In Figure 1-24, top panel, we show several results for a potential barrier. The barrier height equals $V_0=0.4$ eV, whereas the barrier width is $L=6$ nm. We also show how the transmission coefficient varies with the width of the barrier, for fixed E and V_0 (Figure 1-24, bottom panel). We consider two cases: particle energies smaller and larger than the barrier height.

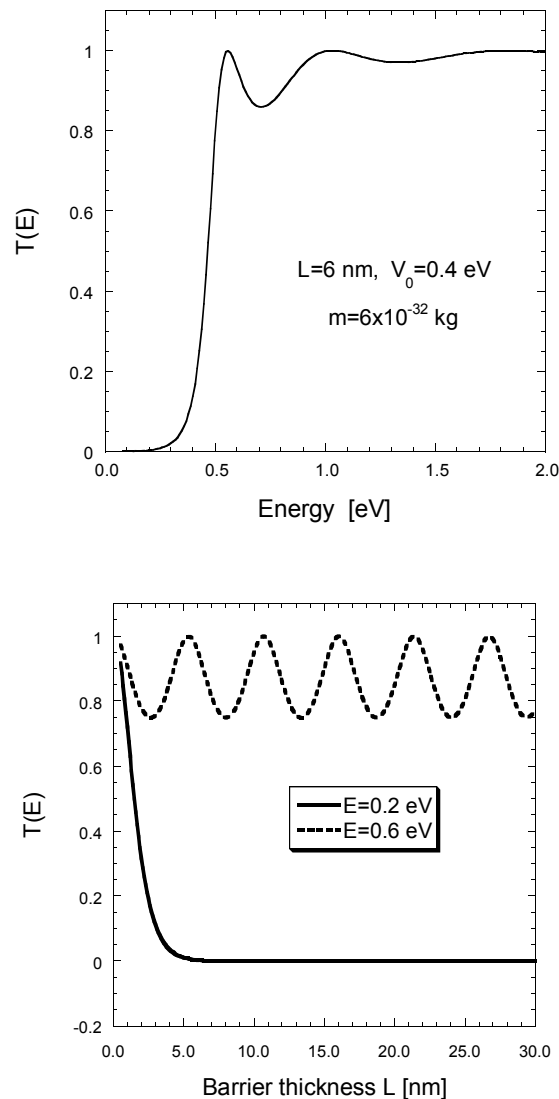


Figure 1-24. Top panel - Variation of the transmission coefficient with energy. Bottom panel - Variation of the transmission coefficient with the barrier thickness.

The following observations can be derived from the results presented in this section:

- Classical physics would predict that no particles with energy $E < V_0$ are transmitted; quantum physics reveals that the probability of transmission of such particles increases hyperbolically with increasing energy.
- Classical physics would predict that all incident particles with $E > V_0$ are transmitted; quantum mechanics shows that this condition - called **total transmission** - occurs only at a few discrete energies. An incident particle with $E > V_0$ that lies between these special values, determined by the condition $k_2 L = n\pi$, may be reflected. The probability of reflection decreases very rapidly with increasing the energy of the particle E .
- For another perspective on transmission and reflection by a barrier, now let us look at the results shown in Figure 1-24, bottom panel. Here, the energy of the particle E and the barrier height are fixed and $T(E)$ is plotted as a function of the barrier width L . This figure shows another bizarre result: for a given energy E , only barriers of certain width will transmit all particles of this energy (transparent barriers). But there is no value of the width such that a barrier of this width reflects all incident particles, because for all values of L , the reflection coefficient $R(E)$ is less than one.
- Because of the hyperbolic decay of the eigenfunction in the classically forbidden region, the amplitude of the eigenfunction in the detector region is reduced from its value in the source region.

1.2.2 Tunneling Through Arbitrary Piecewise-Constant Barrier

We now consider the case of arbitrary number of piece-wise constant potential barriers. As in the case of a simple barrier, the transmission and the reflection coefficients are calculated by solving, either explicitly or implicitly the Schrödinger equation over the domain $x_L < x < x_R$. We assume that outside of the domain of interest (in the asymptotic regions), the wavefunction consists of superposition of traveling waves, and we write the general solution

$$\psi(x) = \begin{cases} a_l e^{ik_l(x-x_l)} + b_l e^{-ik_l(x-x_l)}, & x < x_l \\ a_r e^{-ik_r(x-x_r)} + b_r e^{ik_r(x-x_r)}, & x > x_r \end{cases} \quad (1-52)$$

In the transmission matrix approach outlined in the previous section, the domain is divided into a suitable number of intervals over each of which the potential can be taken to be constant, or perhaps linearly varying. Within each such interval, the wavefunction is expanded in terms of two independent solutions at the chosen energy (oppositely directed traveling waves if the potential is constant). Then the amplitudes of these waves at the two ends of interval i can be related by the propagation matrix P_i :

$$P_i = \begin{bmatrix} e^{ik_i l_i} & 0 \\ 0 & e^{-ik_i l_i} \end{bmatrix} \quad (1-53)$$

The appropriate matching conditions at the boundary between intervals i and $i+1$ must be derived from the form of the Hamiltonian, and are expressed by a matrix B_i which is typically of the form

$$B_i = \frac{1}{2} \begin{bmatrix} 1+r & 1-r \\ 1-r & 1+r \end{bmatrix} \quad (1-54)$$

where $r = v_i / v_{i+1}$, is the velocity ratio. One can then relate the coefficients in the left asymptotic region, incorporate into a vector $\psi_l = [a_l, b_l]^T$, to those in the right asymptotic region, $\psi_r = [b_r, a_r]^T$, by a matrix M formed from the product of the appropriate propagation and boundary matrices:

$$\Psi_r = M\Psi_l = P_m B_{m-1} \dots B_2 P_2 B_1 P_1 \Psi_l \quad (1-55)$$

In practical calculations, the transmission matrix approach has proven to be less than satisfactory, because it is prone to arithmetic overflow. In regions where the wavefunction is evanescent, the P matrices contain real elements equal to the attenuation of the region and its inverse. The inverse is likely to be a very large positive number, and if several evanescent regions are cascaded, the numbers in the matrix will rapidly exceed the dynamic range of floating point variables. This problem is particularly severe when the transmission matrix scheme is applied to multi-band models, because at any given energy many of the bands will be evanescent, but it has also been observed in simple single-band calculations. The transfer matrix method [48,49] has been generalized to multi-dimensional systems by Frenslley [37,50], Lent *et al.*

[51], and to multi-band Hamiltonians by Ting *et al.* [52,53]. In the latter methods the scattering boundary conditions are applied via the quantum transmitting boundary method (QTBM) [54], and the transmission is obtained via the solution of a linear system with dimensions proportional to the size of the device that has to be solved repeatedly. Very recently, a modified version of the QTBM has been developed that expands the scattering solutions in terms of two different closed system wave functions [55]. The calculations are charge self-consistent, but have only been implemented for single-band situations so far.

The second class of simulators is based on the Green's function method to calculate the quantum transport properties, with the coupling to the leads being introduced via the self-energy. The advantage of this approach is the well developed theory of the Green's functions that also allows one to consider inelastic scattering within the nonequilibrium Green's function formalism. A very efficient and widely used algorithm is the recursive Green's function method [56] that has been successfully implemented for two-dimensional devices [57,58] and for small three-dimensional structures such as nano-wires [59]. It's main advantage is that it does not only yield the retarded Green's function that is connected to the S- or T-matrix, but also the less-than Green's function [57], which is needed if inelastic scattering to be considered. The main drawback of the method is its restriction to devices that can be discretized into cross-sectional slices with nearest neighbor interactions only, a condition that cannot be maintained for structures with more than two contacts, since additional contacts inevitably couple more distant slices with one another. For a detailed review of this method please see the article by Lake *et al.* [56].

Thus, although a large variety of methods has been developed in the past decades, the quantum-mechanical ballistic multi-band transport calculation of large two- and three-dimensional structures or devices with more than two Ohmic contacts still presents a significant challenge. To bridge this gap, we have developed a novel and efficient Green's function method to calculate the electronic properties of open quantum systems (such as the transmission, the density of states, the carrier density, and the current density) with arbitrary number of leads in the ballistic limit. To illustrate the idea we first briefly introduce the basics of the Green's function approach used in the ballistic Landauer-Buttiker picture described in Section 3.

1.2.3 Evaluation of the Current Density

To evaluate any physical observables, such as the current density, we must specify how the scattering solutions are statistically weighted in the final result. For the case of a continuous spectrum of states, with δ -function normalization, the derivation of the correct expressions are rather tricky, because we seek expressions for densities of charge, current, energy, etc., rather than total quantities (which are of course infinite in an unbounded system). To illustrate the procedure, let us follow through the derivation of the electron density in a spatially uniform three-dimensional semiconductor in equilibrium. We approximate the conduction band structure by a simple parabolic dispersion relation:

$$E(\mathbf{k}) = E_C + \frac{\hbar^2 \mathbf{k}^2}{2m^*} \quad (1-56)$$

where \mathbf{k} is the wavevector. The probability that each state $|\mathbf{k}\rangle$ will be occupied by an electron is given by the Fermi-Dirac distribution function:

$$f_{FD}[E(\mathbf{k}) - E_F] = \left\{ 1 + e^{\beta[E(\mathbf{k}) - E_F]} \right\}^{-1} \quad (1-57)$$

where E_F is the Fermi level or chemical potential and $\beta = 1/k_B T$, T being the absolute temperature. (To avoid confusion with the transmission probability which is also denoted by T , the absolute temperature will always be shown multiplied by Boltzmann's constant k_B .) Let us now make an *ad hoc* assumption that the semiconductor crystal is a cube with each side of length L , and apply periodic boundary conditions. Then the stationary quantum states are plane waves (normalized to unit amplitude) of the form

$$\psi_{\mathbf{k}}(\mathbf{r}) \sim e^{i\mathbf{k}\cdot\mathbf{r}}. \quad (1-58)$$

Due to the periodic boundary conditions, \mathbf{k} must assume discrete values:

$$\mathbf{k} = \frac{2\pi}{L} (n_x \mathbf{e}_x + n_y \mathbf{e}_y + n_z \mathbf{e}_z), \quad (1-59)$$

where n_x , n_y , and n_z are integers. The total number of electrons in the crystal N is just the sum over all of the states $|\mathbf{k}\rangle$ of the probability that each state is occupied

$$N = 2 \sum_{\mathbf{k}} f_{FD}[E(\mathbf{k})] \quad (1-60)$$

where the factor of 2 comes from the two spin states. Now, because L is large, the allowed values of \mathbf{k} are very closely spaced, and the sum over \mathbf{k} can be well approximated by an integral:

$$\sum_{\mathbf{k}} \rightarrow \left(\frac{L}{2\pi} \right)^3 \int d^3\mathbf{k} \quad (1-61)$$

We can now write an expression for the density of electrons n :

$$n = \frac{N}{L^3} = 2 \int \frac{d^3\mathbf{k}}{(2\pi)^3} f_{FD}[E(\mathbf{k})] \quad (1-62)$$

Notice that the arbitrary crystal dimension L has dropped out of the final expression. In order to evaluate densities using expressions such as Eq. (1-62) it is usually more convenient to transform the integration variable to E . By expressing $d^3\mathbf{k}$ in spherical coordinates and manipulating the dispersion relation one finds:

$$\frac{d^3k}{(2\pi)^3} = \frac{m^* \sqrt{2m^*(E - E_C)}}{2\pi^2 \hbar^3} dE \quad (1-63)$$

Inserting Eq. (1-63) into Eq. (1-62) leads to the usual expression for the electron density in a semiconductor $n = N_C F_{1/2}[\beta(E_F - E_C)]$.

The procedure for evaluating a physical observable in an equilibrium system of infinite extent may thus be generalized from the above discussion. The expectation value of the observable quantity is calculated for each state, taking the scattering states to be normalized to unit amplitude. The density of the observable is then determined by inserting this expectation value into the sum in Eq. (1-62) and evaluating the resulting integral, usually using the relations (1-63). The two most important observables are the electron density $n(x)$ and the current density j (which is independent of position in one dimension and steady-state). The expectation value of the density for a state ψ is simply

$$n_{\psi}(x) = \psi^*(x)\psi(x) \quad (1-64)$$

The expectation value of j is simple, though the operator itself often is not (see Eq. (1-35)). If the dispersion relation $E(k)$ is not parabolic and independent of position, the form of the operator j is *not* given by the simple textbook expression given in Eq. (1-35). The current density operator is instead whatever remains of the kinetic energy term of the Hamiltonian after the application of Green's identity, and this obviously depends upon the form of the Hamiltonian itself. For unit-incident-amplitude scattering states, however, the result is invariably

$$\begin{aligned} \langle \psi_l | J | \psi_l \rangle &= qv_r |t_r|^2 = qv_l T \\ \langle \psi_r | J | \psi_r \rangle &= -qv_l |t_l|^2 = -qv_r T \end{aligned} \quad (1-65)$$

Of course, in equilibrium, these two currents cancel each other (by the principle of detailed balance) and there is no net current flow.

To investigate the transport properties of a quantum system one must generally evaluate the current flow through the system, and this requires that one examine systems that are out of thermal equilibrium. A common situation, in both experimental apparatus and technological systems, is that one has two (or more) physically large regions densely populated with electrons in which the current density is low, coupled by a smaller region through which the current density is much larger. It is convenient to regard the large regions as "electron reservoirs" within which the electrons are all in equilibrium with a constant temperature and Fermi level, and which are so large that the current flow into or out of the smaller "device" represents a negligible perturbation. The reservoirs represent the metallic contacting leads to discrete devices or experimental samples, or the power-supply busses at the system level. Consequently the electrons flowing from a reservoir into the device occupy that equilibrium distribution which characterizes the reservoir. In a simple one-dimensional system with two reservoirs, the electrons flowing in from the left-hand reservoir have $k > 0$ and those flowing from the right-hand reservoir have $k < 0$. Within this picture, the current that is injected from the left-hand reservoir is

$$J_l = 2q \sum_{k_\perp} \int_0^\infty \frac{dk_\parallel}{2\pi} f_{FD} [E(k_\parallel, k_\perp) - E_{Fl}] v_l(k_\parallel, k_\perp) T(k_\parallel, k_\perp) \quad (1-66)$$

and the current injected from the right-hand reservoir is

$$J_r = 2q \sum_{k_\perp} \int_0^\infty \frac{dk_\parallel}{2\pi} f_{FD} [E(k_\parallel, k_\perp) - E_{Fr}] v_r(k_\parallel, k_\perp) T(k_\parallel, k_\perp) . \quad (1-67)$$

In order to simplify the calculation of J further, we must invoke some special properties of the system. The most useful such property is that symmetry which permits the separation of the spatial variables. The separation of variables is possible if the Hamiltonian can be separated into two parts:

$$H = H_\parallel(x, \partial/\partial x) + H_\perp(y, z, \partial/\partial y, \partial/\partial z) . \quad (1-68)$$

(Here the notation H_\parallel and H_\perp is defined with respect to the direction of current transport.) Then the wavefunction separates into a product of two factors:

$$\psi(\mathbf{r}) = \psi_\parallel(x) \psi_\perp(y, z) , \quad (1-69)$$

and the energy can be separated into a product of two terms:

$$E(k) = E_\parallel(k_\parallel) + E_\perp(k_\perp) \quad (1-70)$$

The expression for the total current density J can now be simplified to

$$J = q \int_{V_0}^\infty \frac{dE_\parallel}{2\pi\hbar} T(E_\parallel) [F(E_\parallel - E_{Fl}) - F(E_\parallel - E_{Fr})] , \quad (1-71)$$

where V_0 is the larger of the two asymptotic potentials (minimum energy for a propagating state) and F is the Fermi-Dirac distribution function summed over the transverse states:

$$F(E) = 2 \sum_{k_\perp} \frac{1}{1 + e^{\beta(E + E_\perp)}} \quad (1-72)$$

The form of the sum over k_\perp depends upon the spatial configuration of the tunneling system. Note that the velocity factor does not appear in Eq. (1-71) because it was canceled by the density of states. If the system in question is macroscopically large in its transverse dimensions, the k_\perp form a two dimensional continuum, and $H_\perp = \hbar^2 k^2 / 2m_\perp^*$. Then F can be analytically evaluated and the current density can now be written in the form usually given for the tunneling current:

$$J = q \int_{V_0}^\infty \frac{dE_\parallel}{2\pi\hbar} T(E_\parallel) \ln \left\{ \frac{1 + \exp[-\beta(E_\parallel - E_{Fl})]}{1 + \exp[-\beta(E_\parallel - E_{Fr})]} \right\} \quad (1-73)$$

Note that this expression is valid in general with respect to the dispersion relation in the x direction, but requires a parabolic dispersion relation in the transverse directions. The separation of variables leading to Eq. (1-73) is never rigorously valid in a semiconductor heterostructure. The reason for this is that the transverse effective mass m_\perp^* will vary with semiconductor composition, which varies in the x direction. In principle, one must do at least a two-dimensional integral (if axial symmetry holds, otherwise a three-dimensional integral) as implied by Eq. (1-72). Nevertheless, Eq. (1-73) is widely used to model the current density in heterostructure devices. The error introduced by assuming separation of variables is probably less severe than that due to the assumption of an infinite coherence length.

If the transverse dimensions are constrained, but separation of variables is still possible, the transverse motion of the electrons consists of a discrete set of standing waves or normal modes. Such systems are referred to as ‘‘one-dimensional’’ systems, quantum wires, or electron waveguides. The symbol k_\perp is now interpreted as an index for the discrete transverse modes, and the expression for the current density now becomes

$$J_r = 2q \sum_{k_\perp} \int_0^{-\infty} \frac{dE_\parallel}{2\pi} T(k_\parallel, k_\perp) [f_{FD}(E_\parallel + E_\perp - E_{Fl}) - f_{FD}(E_\parallel + E_\perp - E_{Fr})] \quad (1-74)$$

2. Near-Equilibrium Steady State Transport

The great majority of published work on the subject of quantum transport deals with conditions very near to thermal equilibrium, particularly with very small voltage drop across the transport system. These conditions are known as “linear-response regime”, because the currents induced are linear in the applied voltage. The reason that such circumstances have received so much attention is not due to the technological importance of the linear-response regime, but it is rather due to difficulty of theoretically describing significant departures from equilibrium, as already noted at the end of the previous section. If these departures are negligible, then one may invoke the well-developed machinery of equilibrium statistical physics and simply treat the departure from equilibrium as a small perturbation on the equilibrium state.

One approach to linear response theory is represented by the Kubo formula for the conductivity. In this section we will review the linear response theory and go through the derivation of the Kubo formula for electrical conductivity [60,61]. In that context, we will show that the linear response of a quantum mechanical system can analogously be expressed in terms of functions called Green’s functions [62,63,64,65] which are causal in time.

Let us consider an isolated system, the Hamiltonian of which is denoted by \hat{H} . The dynamical motion of the system determined by \hat{H} is called the natural motion of the system. We suppose that an external force, or field, is applied to the system, the effect of which is represented by the Hamiltonian \hat{H}' that is linear in the external field. Quantum mechanically, the initial steady-state ensemble, which statistically represents the initial state of the system, is specified by the density matrix $\hat{\rho}_0$ satisfying $[\hat{H}, \hat{\rho}_0] = 0$ [66]. For a grand-canonical ensemble, the density matrix is of the form

$$\hat{\rho}_0 = \frac{e^{-\beta(\hat{H} - \mu\hat{N})}}{\text{tr}[e^{-\beta(\hat{H} - \mu\hat{N})}]}, \quad (2-1)$$

where μ is the chemical potential and the temperature is introduced through the parameter $\beta = 1/k_B T$. The density matrix is normalized to unity, i.e. $\text{tr}(\hat{\rho}_0) = 1$.

The response of a quantum mechanical system to an external field can be obtained by solving the appropriate time-dependent Schrödinger equation or, equivalently, for macroscopic systems described statistically, by solving the Liouville equation

$$i\hbar \frac{\partial \hat{\rho}}{\partial t} = [\hat{H} + \hat{H}', \hat{\rho}] \quad (2-2)$$

for the density matrix. In order to find the linear response of the system, one needs to solve Eq. (2-2) to first order in \hat{H}' . This can be achieved with the canonical transformation [67]

$$\hat{\sigma}(t) = \hat{S}(t) \hat{\rho}(t) \hat{S}^\dagger(t), \quad (2-3)$$

where \hat{S} and \hat{S}^\dagger (the adjoint of \hat{S}) are unitary operators satisfying

$$\begin{aligned} -i\hbar \frac{\partial \hat{S}}{\partial t} &= \hat{H} \hat{S} \\ i\hbar \frac{\partial \hat{S}^\dagger}{\partial t} &= \hat{S}^\dagger \hat{H}, \end{aligned} \quad (2-4)$$

respectively. For Hamiltonians that do not depend explicitly on time, $\hat{S}(t) = \exp(i\hat{H}t/\hbar)$. Then the equation of motion for the non-equilibrium density matrix $\hat{\sigma}(t)$ is

$$i\hbar \frac{\partial \hat{\sigma}}{\partial t} = [\hat{H}'(t), \hat{\sigma}(t)], \quad (2-5)$$

where $\hat{H}'(t) = \hat{S}(t)\hat{H}'\hat{S}^\dagger(t)$ is the perturbing Hamiltonian. The solution of Eq. (2-5), to first order in \hat{H}' , is given by

$$\hat{\sigma}(t) = \hat{\rho}_0 - \frac{i}{\hbar} \int_0^t [\hat{H}'(t'), \hat{\rho}_0] dt'. \quad (2-6)$$

In the last equation, it is assumed that the perturbation \hat{H}' is turned on at $t = 0$. For time-independent Hamiltonians, \hat{H} and $\hat{\rho}_0$ commute, and therefore

$$\hat{\rho}(t) = \hat{S}^\dagger(t)\hat{\sigma}(t)\hat{S}(t) = \hat{\rho}_0 - \frac{i}{\hbar} \hat{S}^\dagger(t) \int_0^t dt' [\hat{H}'(t'), \hat{\rho}_0] \hat{S}(t). \quad (2-7)$$

Usually, one is interested in comparing the value of some macroscopic variable represented by the operator \hat{M} with the experiment. The quantity that needs to be compared with the experiment is the quantum and thermal average of \hat{M} at time t given by

$$\langle \hat{M} \rangle = \text{tr} \{ \hat{\rho}(t) \hat{M} \} = \text{tr} \{ \hat{\rho}_0 \hat{M} \} - \frac{i}{\hbar} \text{tr} \left\{ \hat{S}^\dagger(t) \int_0^t dt' [\hat{H}'(t'), \hat{\rho}_0] \hat{S}(t) \hat{M} \right\}. \quad (2-8)$$

The first term on the RHS of Eq. (2-8) is the equilibrium expectation value, and the second term gives the change due to the perturbation. Using the cyclic property of the trace, one can also write

$$\begin{aligned} \langle \hat{M} \rangle &= \text{tr} \{ \hat{\rho}_0 \hat{M} \} - \frac{i}{\hbar} \text{tr} \left\{ \hat{\rho}_0 \int_0^t dt' [\hat{M}(t), \hat{H}'(t')] \right\} \\ &= \text{tr} \{ \hat{\rho}_0 \hat{M} \} - \frac{i}{\hbar} \int_0^t dt' \langle [\hat{M}(t), \hat{H}'(t')] \rangle, \end{aligned} \quad (2-9)$$

where $\hat{M}(t) = \hat{S}(t)\hat{M}\hat{S}^\dagger(t)$.

In semiconductor transport analysis, of major importance is finding the current response of the system due to the time-dependent external field. In other words, one needs to calculate the electrical conductivity that relates the current density J to the electric field $E(\mathbf{r}, t)$ which induces it, through Ohm's law $J_\alpha = \sigma_{\alpha\beta} E_\beta$. Since this is a linear relationship, the conductivity can be found by a proper application of the result given in Eq. (2-9). In the following calculation the electric field is introduced through the vector-potential, so that

$$\mathbf{E}(\mathbf{r}, t) = -\frac{\partial \mathbf{A}(\mathbf{r}, t)}{\partial t}. \quad (2-10)$$

To first order in A (and therefore in E), the perturbing Hamiltonian is

$$\hat{H}' = -\int d^3r j_p(\mathbf{r}) \cdot \mathbf{A}(\mathbf{r}, t) \quad (2-11)$$

where $\hat{j}_p(\mathbf{r})$ is the paramagnetic part of the total symmetrized current-density operator

$$\begin{aligned}
\hat{j}(\mathbf{r}) &= \frac{e}{2m^*} \sum_i \{ [\mathbf{p}_i - e\mathbf{A}(\mathbf{r}, t)] \delta(\mathbf{r} - \mathbf{r}_i) + \delta(\mathbf{r} - \mathbf{r}_i) [\mathbf{p}_i - e\mathbf{A}(\mathbf{r}, t)] \} \\
&= \hat{j}_p(\mathbf{r}) - \frac{ne^2}{m^*} \mathbf{A}(\mathbf{r}, t)
\end{aligned} \tag{2-12}$$

In Eq. (2-12), n is the electron density. The result given in Eq. (2-11) is valid for the Coulomb gauge, $\nabla \cdot \mathbf{A} = 0$. Therefore, the expectation value of the total current-density operator is

$$\langle \hat{j}_\alpha(\mathbf{r}, t) \rangle = -\frac{ne^2}{m^*} A_\alpha(\mathbf{r}, t) + \langle \hat{j}_{p\alpha}(\mathbf{r}, t) \rangle - \frac{i}{\hbar} \int_0^t dt' \langle [\hat{j}_{p\alpha}(\mathbf{r}, t), \hat{H}^I(t')] \rangle \tag{2-13}$$

The first term is a direct-current response to the applied vector potential, which is essentially the displacement current. The second term is assumed to vanish since there is usually no current in a solid in the absence of the electric field. Using the form of \hat{H}^I given in Eq. (2-11), after a straightforward calculation one finds that the current-density components can be obtained from

$$\begin{aligned}
\langle \hat{j}_\alpha(\mathbf{r}, t) \rangle &= -\frac{ne^2}{m^*} A_\alpha(\mathbf{r}, t) + \frac{i}{\hbar} \sum_\beta \int_0^t dt' \int d^3r' \langle [\hat{j}_{p\alpha}(\mathbf{r}, t), \hat{j}_{p\beta}(\mathbf{r}', t')] \rangle A_\beta(\mathbf{r}', t') \\
&= -\frac{ne^2}{m^*} A_\alpha(\mathbf{r}, t) - \sum_\beta \int_0^\infty dt' \int d^3r' G_{\alpha\beta}^r(\mathbf{r}, \mathbf{r}', t - t') A_\beta(\mathbf{r}', t')
\end{aligned} \tag{2-14}$$

where

$$G_{\alpha\beta}^r(\mathbf{r}, \mathbf{r}', t - t') = -\frac{i}{\hbar} \langle [\hat{j}_{p\alpha}(\mathbf{r}, t), \hat{j}_{p\beta}(\mathbf{r}', t')] \rangle \theta(t - t') \tag{2-15}$$

is the so-called retarded two-particle Green's function [68]. The θ -function ensures that the effect at time t depends on the cause only at preceding times. In steady state, the imaginary part of the second term in Eq. (2-15) must cancel the displacement current since the zero-frequency conductivity must be entirely real function. Eq. (2-15) is known as the Kubo formula, in which the current response to an applied electric field is described in terms of the current-current correlation function.

It is also possible to define the advanced Green's function

$$G_{\alpha\beta}^a(\mathbf{r}, \mathbf{r}', t - t') = \frac{i}{\hbar} \langle [\hat{j}_{p\alpha}(\mathbf{r}, t), \hat{j}_{p\beta}(\mathbf{r}', t')] \rangle \theta(t' - t) \tag{2-16}$$

the use of which is more formal. The part of the Green's function that multiplies the θ -function is often called a current-current response function or an after-effect function.

The Eq. (2-14) includes the transient response due to switching on the field at $t = 0$ as well as the forced response. In general, it is the forced response that is usually measured and provided that the Green's function includes dissipative effects, it can be obtained by extending the lower limit of the time integral to $-\infty$. Then, at any finite time, the transients will have died away and one is left with the forced response only. In this case, one frequency component of \mathbf{A} will give rise to corresponding frequency component of \mathbf{J} . Since for simple harmonic perturbation we have

$$\mathbf{A}(\mathbf{r}, \Omega) = -\frac{i}{\Omega} \mathbf{E}(\mathbf{r}, \Omega), \tag{2-17}$$

where Ω is the frequency, the Eq. (2-14) becomes

$$\langle \hat{j}_\alpha(\mathbf{r}, \Omega) \rangle = i \frac{ne^2}{m^* \Omega} E_\alpha(\mathbf{r}, \Omega) + \frac{1}{\hbar \Omega} \sum_\beta \int_0^\infty dt e^{i\Omega t} \int d^3 r' \langle [\hat{j}_{p\alpha}(\mathbf{r}, t), \hat{j}_{p\beta}(\mathbf{r}', 0)] \rangle E_\beta(\mathbf{r}', \Omega). \quad (2-18)$$

The last result was obtained under the assumption that the current-current correlation function is only a function of the time difference $t - t'$. Then, for homogeneous electric fields, the components of the conductivity tensor can be simply calculated from

$$\sigma_{\alpha\beta}(\mathbf{r}, \Omega) = i \frac{ne^2}{m^* \Omega} \delta_{\alpha\beta} + \frac{1}{\hbar \Omega} \int_0^\infty dt e^{i\Omega t} \int d^3 r' \langle [\hat{j}_{p\alpha}(\mathbf{r}, t), \hat{j}_{p\beta}(\mathbf{r}', 0)] \rangle. \quad (2-19)$$

For the inhomogeneous case, the conductivity tensor can be obtained by taking a spatial Fourier transform of Eq. (2-18) and then averaging the result over the spatial volume in order to eliminate any atomic fluctuations. In this case, the final result for the conductivity tensor is [69]

$$\sigma_{\alpha\beta}(\mathbf{q}, \Omega) = i \frac{ne^2}{m^* \Omega} \delta_{\alpha\beta} + \frac{1}{\hbar \Omega} \int_0^\infty dt e^{i\Omega t} \langle [\hat{j}_{p\alpha}(\mathbf{q}, t), \hat{j}_{p\beta}(-\mathbf{q}, 0)] \rangle. \quad (2-20)$$

The results given in Eqs. (2-19) and (2-20) suggest that the complete conductivity tensor for a given frequency of the applied electric field can be rigorously expressed in terms of the electric current components fluctuating spontaneously in the equilibrium state. The basis of this relationship is the fluctuation-dissipation theorem itself [70]. This theorem states that the current is directly proportional to the dissipation arising from the presence of the applied forces. Since the current-current correlation function describes these fluctuations, the Kubo formula is no more than a direct statement of this important theorem for close-to-equilibrium conditions.

Another important result that follows from the discussion in this chapter is the fact that the linear response of the quantum-mechanical system can be expressed in terms of functions that are called Green's functions. Since direct calculation of the retarded two-particle Green's function for a system with strongly interacting particles is somewhat difficult, the path commonly followed is the following:

- One first calculates the one-particle retarded

$$G_r(x_1, x_2) = -\frac{i}{\hbar} \theta(t_1 - t_2) \langle [\hat{\Psi}(x_1), \hat{\Psi}^+(x_2)] \rangle \quad (2-21)$$

and advanced

$$G_a(x_1, x_2) = \frac{i}{\hbar} \theta(t_2 - t_1) \langle [\hat{\Psi}(x_1), \hat{\Psi}^+(x_2)] \rangle \quad (2-22)$$

Green's functions.

- After these functions are calculated, one looks for the relations between these formal Green's functions and the observable ones.

In summary, Eq. (2-19) is a form of the fluctuation-dissipation theorem, relating a transport coefficient, which necessarily characterizes a dissipative process, to the fluctuations about the equilibrium state. Another well-known form of the fluctuation-dissipation theorem is the Einstein relation connecting the mobility and the diffusivity in classical transport theory: $\mu = qD / k_B T$. The Kubo formula expresses the conductivity in terms of the autocorrelation of the current density; if one can calculate this autocorrelation function from the equations of motion, for example, one can evaluate the frequency-dependent conductivity.

2.1 Conductance – The Landauer-Buttiker formula

Let us consider the following system, composed by a one-dimensional channel (a quantum wire) with length L , and two metallic reservoirs with electrochemical potential μ_L and μ_R ($\mu_L > \mu_R$).

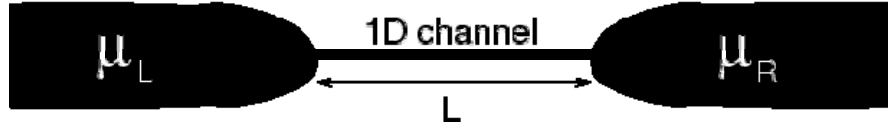


Figure 2-1. Sketch of the considered one-dimensional system.

We suppose that only the first one-dimensional subband is occupied, that the electrons in the channel do not suffer any scattering mechanism, i.e. transport in the channel is ballistic, and that the electrons entering the reservoirs contacts are instantaneously in equilibrium with them. For the moment, let the temperature be equal to zero ($T=0$ K) and the contacts *reflectionless*, that means that the transmission probability from the contact to contact is unitary. Being a two-dimensional confined system, the dispersion relation is equal to $E_i(k) = \varepsilon_i + \hbar^2 k^2 / 2m$, where ε_i is also referred as the *cut-off energy* of i -th 1D subband or *transversal mode*. The positive current, carried by $k > 0$ states in the i th-subband, reads [71]

$$I_i^> = \frac{q}{L} \sum_{k>0} v_i(k) f(E - \mu_L) \quad (2-23)$$

where v_i is the velocity and f is the Fermi-Dirac distribution function. Converting the summation over states into an integral while conserving the number of states in the system, we obtain

$$I_i^> = \frac{q}{\pi} \int_0^\infty v_i(k) f(E - \mu_L) dk \quad (2-24)$$

and from the dispersion relation it follows that

$$I_i^> = \frac{2q}{h} \int_{\varepsilon_i}^\infty f(E - \mu_L) dE \quad (2-25)$$

Equivalently, the negative current reads

$$I_i^< = \frac{2q}{h} \int_{\varepsilon_i}^\infty f(E - \mu_R) dE \quad (2-26)$$

If we now define the function $M(E)$ as

$$M(E) = \sum_i u(E - \varepsilon_i) \quad (2-27)$$

where $u(E)$ is the Heavyside function, we can express the total positive current as

$$I^> = \sum_i I_i^> = \sum_i \frac{2q}{h} \int_{\varepsilon_i}^\infty f(E - \mu_L) dE = \frac{2q}{h} \int_{-\infty}^{+\infty} f(E - \mu_L) M(E) dE \quad (2-28)$$

Working at $T = 0K$, the Fermi-Dirac function is a step function, and considering $M(E)$ constant and equal to an integer M between the energy range $[\mu_L, \mu_R]$, the total current reads

$$I = I^> - I^< = \frac{2q^2}{h} M \frac{\mu_L - \mu_R}{q} \quad (2-29)$$

We note that Eq. (2-29) can be reduced to the form $I = GV$, where V is the applied potential $\frac{\mu_L - \mu_R}{q}$ and G is the conductance, that in this case is

$$G = \frac{2q^2}{h} M \quad (2-30)$$

Since the channel is purely ballistic, the resistance of the channel is equal to zero. This means that all the dissipation of the energy happens in the reservoirs, since a charge propagating from the left transverse the channel conserving the energy, and when it reaches the right reservoir, it loses energy thermalizing in the contact. That is why the inverse of G is often referred as the *contact resistance*.

If we now relax the hypothesis of reflectionless contacts, we have the situation depicted in Figure 2-2

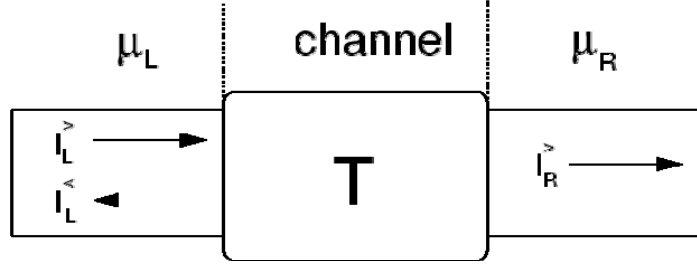


Figure 2-2. Electrons propagating from the left contact undergo elastic scattering and T is the probability of reaching the other reservoir.

If $I_L^>$ is the influx of electrons from the left reservoir, $I_L^<$ is the flux of back-scattered electrons in the left reservoir, and $I_R^>$ is the flux of electrons that has reached the right reservoir (Figure 2-2), if we consider $T(E)$ constant and equal to τ between the energy range $\mu_L - \mu_R$, we deal with the following relations

$$\begin{aligned} I_L^> &= \frac{2q}{h} M (\mu_L - \mu_R) \\ I_L^< &= \frac{2q}{h} M (1 - \tau) (\mu_L - \mu_R) \\ I_R^> &= \frac{2q}{h} M \tau (\mu_L - \mu_R) \end{aligned} \quad (2-31)$$

The total current I is then equal to

$$I = I_L^> - I_L^< = I_R^> = \frac{2q}{h} M \tau (\mu_L - \mu_R) \quad (2-32)$$

and the conductance in case of reflecting contacts can be expressed as

$$G = \frac{2q^2}{h} M \tau \quad (2-33)$$

that is the *Landauer Formula* [72].

We can now study the case in which the temperature is not equal to zero. The total current now reads

$$I = \frac{2q}{h} \int_{-\infty}^{+\infty} [F(E - \mu_L) - f(E - \mu_R)] M(E) T(E) dE \quad (2-34)$$

If $\mu_R = \mu$ and $\mu_L = \mu + \delta\mu$, with $\delta\mu \ll \mu$, Eq. (2-34) becomes

$$I = \frac{2q}{h} \int_{-\infty}^{+\infty} [F(E - (\mu + \delta\mu)) - f(E - \mu)] M(E) T(E) dE \quad (2-35)$$

and by means of the Taylor expansion

$$f(E - (\mu + \delta\mu)) - f(E - \mu) = -\delta\mu \frac{\partial f(E, \mu)}{\partial E} \quad (2-36)$$

we can express Eq. (2-35) as

$$I = \frac{2q}{h} \left[\int_{-\infty}^{+\infty} -\frac{\partial f(E, \mu)}{\partial E} M(E) T(E) dE \right] \frac{\delta\mu}{q} \quad (2-37)$$

that can be reduced to

$$G = \frac{2q}{h} \int_{-\infty}^{+\infty} -\frac{\partial f(E, \mu)}{\partial E} M(E) T(E) dE \quad (2-38)$$

where $-\partial f(E, \mu) / \partial E$ is often referred as the *broadening function*.

As a numerical example, we can consider the case of a QPC with reflectionless contacts in which the transversal potential is a harmonic potential. The transversal modes are equal to

$$E_m = \left(m - \frac{1}{2}\right) \hbar \omega_0. \quad (2-39)$$

In Figure 2-3 we show the conductance at different temperatures, as a function of the electrochemical potential of the reservoir (μ), supposing a small voltage $\delta\mu$ is applied between the two reservoir, and $\hbar\omega_0 = 0.1eV$. As can be noted, as the temperature is increased, the step-like function of the conductance at $T=0\text{ K}$ is smoothed. Such an effect is due to the fact that the broadening function “broadens” over the energy as the temperatures increases: Eq. (2-39) can be then seen as an average over a window that increases as the temperature is increased.

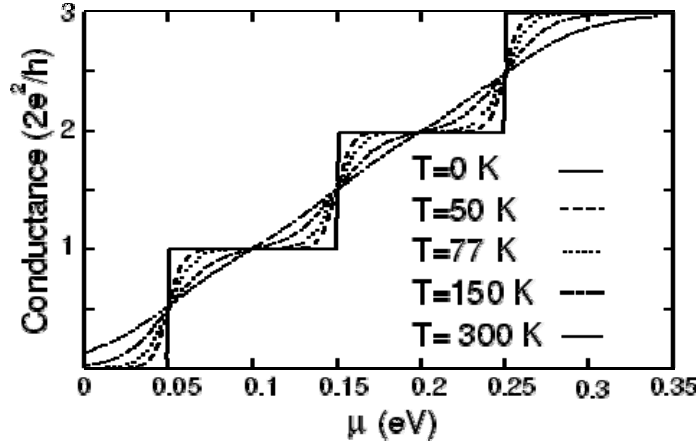


Figure 2-3. Conductance of a QPC with reflectionless contact at different temperatures, in which the transversal potential has been approximated with a harmonic potential ($\hbar\omega_0 = 0.1eV$).

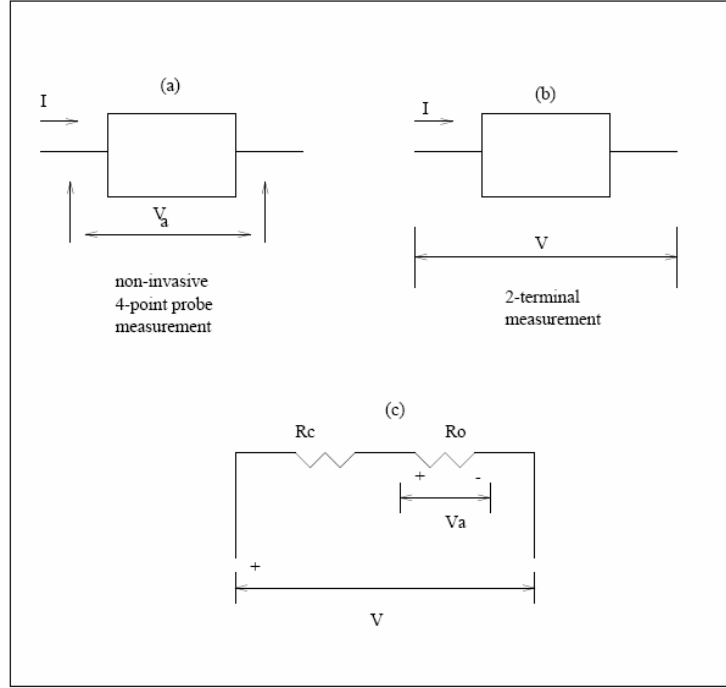


Figure 2-4 (a) $G=(2e^2/h)T/(1-T)=I/V_a$. (b) $G=I/V=(2e^2/h)T$. Here, in addition to measuring the voltage drop across the scattering structure, one also measures the contact resistance. (c) Circuit equivalent of (b) where $V = RcI + V_a$.

3. Far-From-Equilibrium Transport

3.1 Mixed States and Distribution Function

When a system such as an electron device is driven far from equilibrium by the application of an external voltage, both coherent and incoherent processes will generally occur within the device. Coherent processes include tunneling and ballistic transport, and incoherent processes include dissipative scattering via phonons, for example. Coherent effects are described by adding complex-valued amplitudes (that is, values of the wavefunction), which is done implicitly in the solution of Schrödinger's equation above. Incoherent effects are described by superposition of real-valued probabilities. An example of such incoherent superposition is the summation of the current density over energies and transverse modes to obtain the total current density. We can formalize the statistical summation procedure described there into a mathematical object known as the single particle density matrix [73,74] In terms of the continuum position variable x , the density matrix is actually a complex-valued function of two arguments, and has the general form:

$$\rho(x, x') = \sum_i P_i \psi_i(x) \psi_i^*(x') \quad (3-1)$$

where the ψ_i form a complete set of states (*not* necessarily the eigenstates of the Hamiltonian), and the P_i are real-valued probabilities for finding an electron in each state ψ_i . With this definition, the expectation value of any physical observable represented by an operator A is given by:

$$\langle A \rangle = Tr(A\rho) = \lim_{x \rightarrow 0} \int (A\rho)(x, x') dx, \quad (3-2)$$

where A is taken to operate with respect to the first argument of ρ . Inserting Eq. (3-1) into Eq. (3-2) and rearranging the expression, we get the more familiar form for the expectation value:

$$\langle A \rangle = \sum_i P_i \int \psi_i^*(x) A \psi_i(x) dx \quad (3-3)$$

In particular, the particle density is given by

$$n(x) = \rho(x, x), \quad (3-4)$$

and the current density is

$$J(x) = \lim_{x \rightarrow 0} \frac{q\hbar}{im} * \left(\frac{\partial}{\partial x} - \frac{\partial}{\partial x'} \right) \rho. \quad (3-5)$$

If $E(\mathbf{k})$ is non-parabolic, a more complicated expression is required for the current density.

If the motion of the particles described by the density matrix is purely ballistic (no energy loss) and defined by a Hamiltonian H , the equation describing the evolution of the density matrix may be derived by substituting Schrödinger's equation into Eq. (3.1). The result is the Liouville-von Neumann equation for the single-particle density matrix:

$$\frac{\partial \rho}{\partial t} = \frac{1}{i\hbar} [H\rho - \rho H] \equiv L_\rho \rho \quad (3-6)$$

where L_ρ is a linear operator which operates upon the density matrix and is called the Liouville operator. (Since it operates upon ρ , which is itself a quantum-mechanical operator, L_ρ is technically a superoperator.) The Liouville equation acts upon the density matrix by evolving the wavefunctions, but does not change the probabilities P_i . This is a characteristic of ballistic, or conservative, motion. Irreversible, or dissipative, processes involve transitions between quantum states, and are described by operators which modify the probabilities P_i . Such operators are discussed below.

In classical systems, the quantity which describes the state of the system corresponding to ρ is the phase-space distribution function $f_c(r, p)$ where r is now the position and p is the momentum. The classical Liouville equation is

$$\frac{\partial f_c(r, p, t)}{\partial t} = -v \frac{\partial f_c(r, p, t)}{\partial r} + \frac{\partial V}{\partial r} \frac{\partial f_c(r, p, t)}{\partial p} \equiv L_c \quad (3-7)$$

where v is the velocity and V is the potential in which the particles are moving. The particle and current densities are obtainable from the classical distribution function by using

$$\begin{aligned} n(r, t) &= \int dp f_c(r, p, t) \\ J(r, t) &= \int dp v(p) f_c(r, p, t) \end{aligned} \quad (3-8)$$

The Liouville equation, in either the classical or quantum context, describes only ideal, conservative motion. Within the scope of these equations, particles can only oscillate within the system, unless one applies boundary conditions which permit particles to escape from it. The form of the equations (for closed systems) does not describe an approach to a steady-state, neither equilibrium nor non-equilibrium. The reason for this involves the eigenvalue spectrum of L_ρ and L_c . The solutions of Eq. (3-6) will consist of a linear combination of terms with time dependence $\exp(-i\omega_i t / \hbar)$, where $-i\omega_i$ are the eigenvalues of L_ρ .

The Liouville operator [as defined in Eq. (3-6), including the imaginary factor] is anti-Hermitian, so the frequencies ω_i are purely real. Thus the transformation which maps the state of the system at some initial time into some later time is a unitary linear transformation, and we will call the behavior described by such equations "unitary time evolution." Devices of course usually approach a steady state after a sufficiently long time. To describe this behavior, we must incorporate irreversible processes into the equations.

3.1.1 Irreversible Processes and MASTER Equations

Irreversible or energy-dissipating processes always involve transitions between quantum states. Such processes are described, at the simplest level, by master or rate equations [75]. The operators which generate the time-evolution in such equations are of a very different form from that of the Liouville operator. If the state of a system is described by an array of probabilities or occupation factors P_i for a particle to occupy a (stationary) quantum level i , the time evolution of that system is determined by the rates of transition between the levels i . These rates are usually estimated using the "Fermi Golden Rule"

$$W_{ij} = \frac{2\pi}{\hbar} \left| \langle i | H_{\text{int}} | j \rangle \right|^2 \delta(E_i - E_j) \quad (3-9)$$

where H_{int} is the Hamiltonian describing the interaction that causes the transitions, and W_{ij} is the transition rate from state j to state i . The δ -function ensures energy conservation, but it must be remembered that E_i and E_j are the total energy of each state, including, for example, the energy in an emitted phonon. Thus Eq.

(3-9) can describe energy-dissipating processes despite its appearance. If one assumes that these transitions occur independently within any small time interval (the Markov assumption), the transition from state j to state i will produce changes in the corresponding occupation factors:

$$dP_i = -dP_j = W_{ij}P_j dt \quad (3-10)$$

The occupation of state i increases and that of state j decreases as a result of this particular process, and the amount of change depends only upon the occupation of the initial state. (We neglect here the Pauli exclusion principle, which leads to nonlinear master equations.) If we sum over all possible transition processes, we obtain the master equation:

$$\frac{dP_i}{dt} = \sum_j [W_{ij}P_j(t) - W_{ji}P_i(t)] = MP, \quad (3-11)$$

where M is the master operator, whose matrix elements are given by

$$M_{ij} = \begin{cases} W_{ij} & i \neq j \\ -\sum_{j \neq i} W_{ij} & i = j \end{cases} \quad (3-12)$$

Notice the form of this operator and remember that $W_{ii} = 0$. The off-diagonal elements are all positive and the diagonal elements are all negative, with a magnitude equal to the sum of the off-diagonal elements in the same column. (If one considers an open system, the coupling to external reservoirs can lead to master operators in which the magnitude of the diagonal elements exceeds the sum of the off-diagonal elements.) The eigenvalues of an operator of this form will all have real parts less than or equal to zero. Thus the solutions of Eq. (3-11) will consist of a linear combination of terms with a decaying exponential time-dependence, and so will always show a stable approach to some steady state. The Pauli master equation [76] is the most commonly used model of irreversible processes in simple quantum systems. It can be derived from elementary quantum mechanics plus a Markov assumption. There are a number of conceptual problems with the Pauli equation, not the least of which is that it produces violations of the continuity equation. It is nevertheless employed, either explicitly or implicitly, in almost all semi-classical treatments of electron transport in semiconductors.

Master operators most often occur in the description of stochastic (random) processes, where they describe the average behavior of the system. In such cases there will always be fluctuations (noise) about the solution of the master equation. Diffusion phenomena are the classic example of this. The master operator in the classical diffusion equation $\partial n/\partial t = D\nabla^2 n$ is just the laplacian ∇^2 . By examining the form of the finite-difference approximation to the second derivative, it is easy to see that this has the form of a master operator.

3.1.2 The Boltzmann Equation

In this section we will start from semiclassical principles and derive the semiclassical master equation, also known as the Boltzmann transport equation. We first note that in device simulations, to completely specify the operation of a device, one must know the state of each carrier within the device. If carriers are treated as classical particles, one way of specifying the state of the carriers is to solve Newton's equations

$$\frac{d\mathbf{p}}{dt} = -e\mathbf{E} + R(\mathbf{r}, \mathbf{p}, t) \quad \text{and} \quad \mathbf{v}(t) = \frac{d\mathbf{r}}{dt}, \quad (3-13)$$

where $R(\mathbf{r}, \mathbf{p}, t)$ is a random force function due to impurities or lattice vibrations or other imperfections in the system. Alternative approach would be to calculate the probability of finding a carrier with crystal momentum \mathbf{k} at position \mathbf{r} at time t , given by the distribution function $f(\mathbf{r}, \mathbf{k}, t)$, obtained by solving the Boltzmann transport equation (BTE) [77,78,79]. It is important to note that this theory is based on the following assumptions:

- Electrons and holes are independent particles.
- The system is described by a set of Bloch functions [80,81].
- Particles do not interact with each other, but may be scattered by impurities, phonons, etc.
- The number of electrons in an elementary volume ΔV centered around \mathbf{r} , that have

$$\text{wavevectors in the range of } d^3k \text{ centered around } \mathbf{k} \text{ is given by } 2 \times \frac{\Delta V}{8\pi^3} f(\mathbf{r}, \mathbf{k}, t) d^3k$$

Therefore, once the distribution function is specified, various moments of the distribution function can give us particle density, current density, energy density, etc. More precisely

$$n(\mathbf{r}, t) = \frac{1}{V} \sum_{\mathbf{k}} f(\mathbf{r}, \mathbf{k}, t), \text{ particle density} \quad (3-14)$$

$$\mathbf{J}(\mathbf{r}, t) = -\frac{e}{V} \sum_{\mathbf{k}} \mathbf{v}(\mathbf{k}) f(\mathbf{r}, \mathbf{k}, t), \text{ current density} \quad (3-15)$$

$$W(\mathbf{r}, t) = \frac{1}{V} \sum_{\mathbf{k}} E(\mathbf{k}) f(\mathbf{r}, \mathbf{k}, t), \text{ energy density} \quad (3-16)$$

A full quantum-mechanical view to this problem is rather difficult [82,45]. The uncertainty principle states, for example, that we can not specify simultaneously the position and the momentum of the particle. Hence, one needs to adopt a coarse-grained average point of view, in which positions are specified within a macroscopic volume, and momenta are also specified within some interval. If one tries to go straightforwardly and construct $f(\mathbf{r}, \mathbf{k}, t)$ from the quantum-mechanical wavefunctions, difficulties arise since f is not necessarily positive definite.

(A) *Approximations made for the distribution function*

The most difficult problem in device analysis is to calculate the distribution function $f(\mathbf{r}, \mathbf{k}, t)$. To overcome these difficulties, reasonable guess for the distribution function is often made. Two most commonly used approaches are:

- Quasi-Fermi level concept.
- Displaced Maxwellian approximation for the distribution function.\

Under equilibrium conditions $np = n_i^2$, where n is the electron concentration, p is the hole concentration and n_i is the intrinsic carrier concentration which follows from the use of the equilibrium distribution functions for electrons and holes, i.e.

$$f_n(E) = \frac{1}{1 + \exp\left(\frac{E - E_F}{k_B T}\right)}, \quad f_p(E) = 1 - f_n(E) = \frac{1}{1 + \exp\left(\frac{E_F - E}{k_B T}\right)}. \quad (3-17)$$

Under non-equilibrium conditions, it may still be useful to represent the distribution functions for electrons and holes as

$$f_n(E) = \frac{1}{1 + \exp\left(\frac{E - E_{Fn}}{k_B T}\right)} \quad \text{and} \quad f_p(E) = 1 - f_n(E) = \frac{1}{1 + \exp\left(\frac{E_{Fp} - E}{k_B T}\right)}. \quad (3-18)$$

Therefore, under non-equilibrium conditions and assuming non-degenerate statistics, we will have

$$n = N_C \exp\left(\frac{E_{Fn} - E_C}{k_B T}\right), \quad \text{and} \quad p = N_V \exp\left(\frac{E_V - E_{Fp}}{k_B T}\right), \quad (3-19)$$

where N_C and N_V are the effective density of states of the conduction and valence band, respectively, and E_{Fn} and E_{Fp} are the electron and hole quasi-Fermi levels [83,84]. The product

$$np = n_i^2 \exp\left(\frac{E_{Fn} - E_{Fp}}{k_B T}\right) \quad (3-20)$$

suggests that the difference $E_{Fn} - E_{Fp}$ is a measure for the deviation from the equilibrium. However, this can not be correct distribution function since it is even in k , which means that it suggests that current can never flow in a device. The fact that makes it not so unreasonable is that average carrier velocities are usually much smaller than the spread in velocity, given by $\sqrt{2k_B T/m^*} \approx 10^7 \text{ cm/s}$ for $m^* = m_0$ (free electron mass).

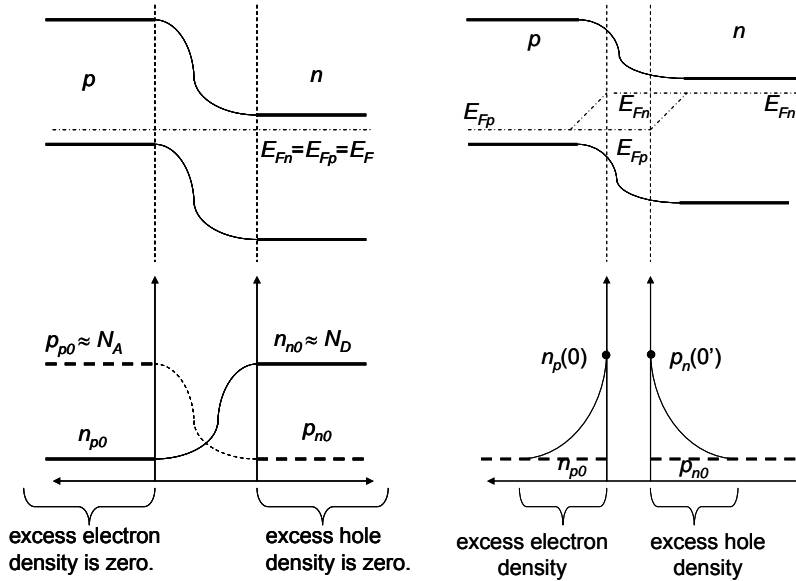


Figure 3-1. Energy band profile of a pn-diode under equilibrium and non-equilibrium conditions. Note that to get the excess electron density (bottom right panel) the electron quasi-Fermi level must move up (top right panel), thus increasing the probability of state occupancy. The same is true for the excess hole concentration, where the hole quasi-Fermi level moves downward.

A better guess for the distribution function $f(\mathbf{r}, \mathbf{k}, t)$ is to assume that the distribution function retains its shape, but that its average momentum is displaced from the origin. For example, particularly suitable form to use is [85]

$$f(\mathbf{r}, \mathbf{k}, t) = \exp\left(\frac{E_{Fn} - E_{C0}}{k_B T}\right) \exp\left(-\frac{\hbar^2}{2m^* k_B T} |\mathbf{k} - \mathbf{k}_d|^2\right) \quad (3-21)$$

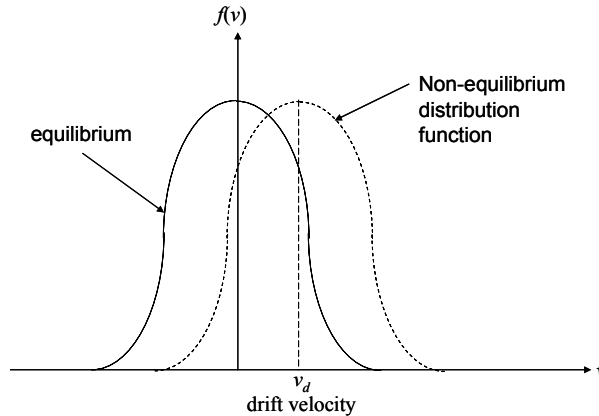


Figure 3-2. Displaced Maxwellian distribution function.

Using this form of the distribution function gives

$$n(\mathbf{r}, t) = \frac{1}{V} \sum_{\mathbf{k}} f(\mathbf{r}, \mathbf{k}, t) = N_C \exp\left(\frac{E_{Fn} - E_{C0}}{k_B T}\right) \quad (3-22)$$

In the same manner, one finds that the kinetic energy density per carrier $u(\mathbf{r}, t)$ is given by

$$u(\mathbf{r}, t) = \frac{1}{2} m^* v_d^2 + \frac{3}{2} k_B T \quad (3-23)$$

The first term on the RHS of Eq. (3.23) represents the drift energy due to average drift velocity, and the second term is the well known thermal energy term due to collisions of carriers with phonons [86].

Since in both cases, the guess for the non-equilibrium distribution has been guided by the form of the equilibrium, they are only valid in near-equilibrium conditions. For far-from-equilibrium conditions, the shape of the distribution function can be rather different [87]. This necessitates the solution of the Boltzmann transport equation that is introduced in the following section.

(B) Boltzmann transport equation

To derive the BTE consider a region of phase space about the point (x, y, z, p_x, p_y, p_z) . The number of particles entering this region in time dt is equal to the number which were in the region of phase space $(x-v_x dt, y-v_y dt, z-v_z dt, p_x-F_x dt, p_y-F_y dt, p_z-F_z dt)$ at a time dt earlier. If $f(x, y, z, p_x, p_y, p_z)$ is the distribution function which expresses the number of particles per quantum state in the region, then the change df which occurs during time dt due to the motion of the particles in coordinate space and due to the fact that force fields acting on the particles tend to move them from one region to another in momentum space is [88]:

$$df = f(x - v_x dt, y - v_y dt, z - v_z dt, p_x - F_x dt, p_y - F_y dt, p_z - F_z dt) - f(x, y, z, p_x, p_y, p_z) \quad (3-24)$$

Using Taylor series expansion, we get

$$\frac{df}{dt} = -\mathbf{v} \cdot \nabla_{\mathbf{r}} f - \mathbf{F} \cdot \nabla_{\mathbf{p}} f \quad (3-25)$$

So far, only the change in the distribution function due to the motion of particles in coordinate space and due to the momentum changes arising from the force fields acting on the particles have been accounted for. Particles may also be transferred into or out of a given region in phase space by collisions or scattering interactions involving other particles of the distribution or scattering centers external to the assembly of particles under consideration. If the rate of change of the distribution function due to collisions, or scattering, is denoted by $(\partial f / \partial t)_{coll}$, the total rate of change of f becomes

$$\frac{df}{dt} = -\mathbf{v} \cdot \nabla_{\mathbf{r}} f - \mathbf{F} \cdot \nabla_{\mathbf{p}} f + \left. \frac{\partial f}{\partial t} \right|_{coll} + s(r, p, t) \quad (3-26)$$

i.e.

$$\frac{df}{dt} + \mathbf{v} \cdot \nabla_{\mathbf{r}} f + \mathbf{F} \cdot \nabla_{\mathbf{p}} f = \left. \frac{\partial f}{\partial t} \right|_{coll} + s(r, p, t) \quad (3-27)$$

The last term on the RHS of Eqs. (3-26) and (3-27) occurs when generation-recombination processes play significant role. Eq. (3-27) represents the Boltzmann transport equation, which is nothing more but a book-keeping equation for the particle flow in the phase space.

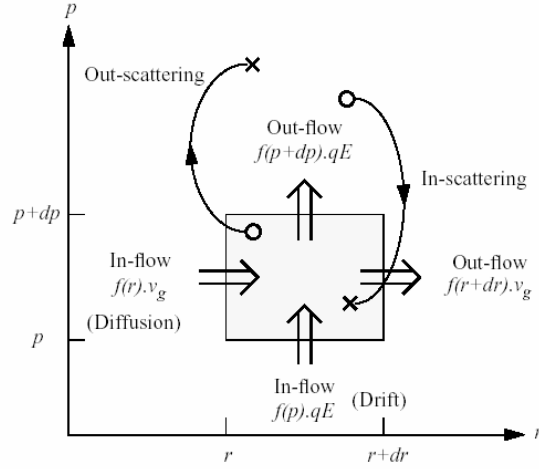


Figure 3-3. A cell in two-dimensional phase space. The three processes, namely drift, diffusion, and scattering, that affect the evolution of $f(\mathbf{r}, \mathbf{p}, t)$ with time in phase space are shown.

The various terms that appear in Eq. (3-27) represent

- $\left(\frac{\partial f}{\partial t} \right)_{forces} = -\mathbf{F} \cdot \nabla_{\mathbf{p}} f$, where $\mathbf{F} = \frac{d\mathbf{p}}{dt} = \hbar \frac{d\mathbf{k}}{dt} = q(\mathbf{E} + \mathbf{v} \times \mathbf{B})$, the total force equals the sum of the force due to the electric field and the Lorentz force due to the magnetic flux density, \mathbf{B} .
- $\left(\frac{\partial f}{\partial t} \right)_{diff} = -\mathbf{v} \cdot \nabla_{\mathbf{r}} f$. This term arises if there is a spatial variation in the distribution function due to concentration or temperature gradients, both of which will result in a diffusion of carriers in coordinate space.
- $\left(\frac{\partial f}{\partial t} \right)_{coll}$ is the collision term which equals the difference between the in-scattering and the out-scattering processes, i.e.

$$\left(\frac{\partial f}{\partial t} \right)_{coll} = \sum_{\mathbf{k}'} \{ S(\mathbf{k}', \mathbf{k}) f(\mathbf{k}') [1 - f(\mathbf{k})] - S(\mathbf{k}, \mathbf{k}') f(\mathbf{k}) [1 - f(\mathbf{k}')] \} = \widehat{C}f \quad (3-28)$$

The presence of $f(\mathbf{k})$ and $f(\mathbf{k}')$ in the collision integral makes the BTE rather complicated integro-differential equation for $f(\mathbf{r}, \mathbf{k}, t)$, whose solution requires a number of simplifying assumptions. In the absence of perturbing fields and temperature gradients, the distribution function must be the Fermi-Dirac function. In this case, the collision term must vanish and the principle of detailed balance gives for all \mathbf{k} and \mathbf{k}' and all scattering mechanisms

$$\frac{S(\mathbf{k}, \mathbf{k}')}{S(\mathbf{k}', \mathbf{k})} = \frac{f_0(\mathbf{k}') [1 - f_0(\mathbf{k})]}{f_0(\mathbf{k}) [1 - f_0(\mathbf{k}')]} \quad (3-29)$$

Therefore, if the phonons interacting with the electrons are in thermal equilibrium, we get

$$\frac{S(\mathbf{k}, \mathbf{k}')}{S(\mathbf{k}', \mathbf{k})} = \exp\left(\frac{E_{\mathbf{k}} - E_{\mathbf{k}'}}{k_B T}\right) \quad (3-30)$$

This relation must be satisfied regardless of the origin of the scattering forces. If, for example, we assume $E_{\mathbf{k}} > E_{\mathbf{k}'}$, then $S(\mathbf{k}, \mathbf{k}')$ which involves emission must exceed $S(\mathbf{k}', \mathbf{k})$ which involves absorption. Note that the BTE is valid under assumptions of semi-classical transport: effective mass approximation (which incorporates the quantum effects due to the periodicity of the crystal); Born approximation for the collisions, in the limit of small perturbation for the electron-phonon interaction and instantaneous collisions; no memory effects, i.e. no dependence on initial condition terms. The phonons are usually treated as in equilibrium, although the condition of non-equilibrium phonons may be included through an additional equation [89].

(C) Scattering Processes

Free carriers (electrons and holes) interact with the crystal and with each other through a variety of scattering processes which relax the energy and momentum of the particle. Based on first order, time-dependent perturbation theory, the transition rate from an initial state \mathbf{k} in band n to a final state \mathbf{k}' in band m for the j th scattering mechanism is given by Fermi's Golden Rule [90]

$$\Gamma_j[n, \mathbf{k}; m, \mathbf{k}'] = \frac{2\pi}{\hbar} \left| \langle m, \mathbf{k}' | V_j(\mathbf{r}) | n, \mathbf{k} \rangle \right|^2 \delta(E_{\mathbf{k}'} - E_{\mathbf{k}} \mp \hbar\omega) \quad (3-31)$$

where $V_j(\mathbf{r})$ is the scattering potential of this process, $E_{\mathbf{k}}$ and $E_{\mathbf{k}'}$ are the initial and final state energies of the particle. The delta function describes conservation of energy, valid for long times after the collision is over, with $\hbar\omega$ the energy absorbed (upper sign) or emitted (lower sign) during the process. The total rate used to generate the free flight is then given by

$$\Gamma_j[n, \mathbf{k}] = \frac{2\pi}{\hbar} \sum_{m, \mathbf{k}'} \left| \langle m, \mathbf{k}' | V_j(\mathbf{r}) | n, \mathbf{k} \rangle \right|^2 \delta(E_{\mathbf{k}'} - E_{\mathbf{k}} \mp \hbar\omega) \quad (3-32)$$

There are major limitations to the use of the Golden rule due to effects such as *collision broadening* and *finite collision duration time* [91]. The energy conserving delta function is only valid asymptotically for times long after the collision is complete. The broadening in the final state energy is given roughly by $\Delta E \approx \hbar/\tau$, where τ is the time after the collision, which implies that the normal $E(\mathbf{k})$ relation is only recovered at long times. Attempts to account for such *collision broadening* in Monte Carlo simulation have been reported in the literature [92,93], although this is still an open subject of debate. Inclusion of the effects of *finite collision duration* in Monte Carlo simulation have also been proposed [94,95]. Beyond this, there is still the problem of dealing with the quantum mechanical phase coherence of carriers, which is neglected in the scatter free-flight algorithm of the Monte Carlo algorithm, and goes beyond the semi-classical BTE description.

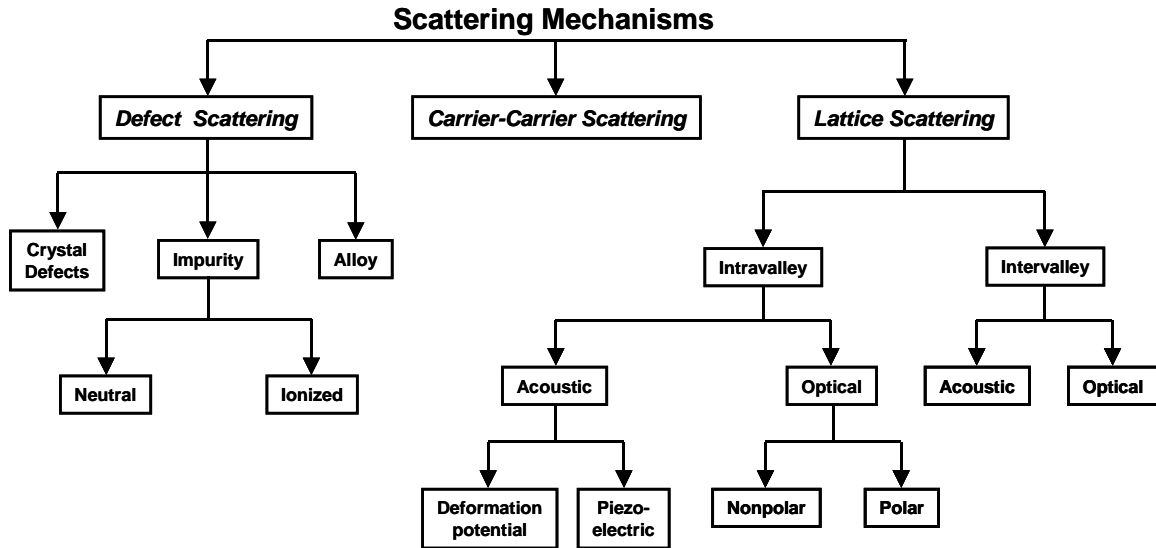


Figure 3-4. Scattering mechanisms in a typical semiconductor.

Figure 3-4 lists the scattering mechanisms one should in principle consider in a typical Monte Carlo simulation. They are roughly divided into scattering due to crystal defects, which is primarily elastic in nature, lattice scattering between electrons (holes) and lattice vibrations or phonons, which is inelastic, and finally scattering between the particles themselves, including both single particle and collective type excitations. Phonon scattering involves different modes of vibration, either acoustic or optical, as well as both transverse and longitudinal modes. Carriers may either emit or absorb quanta of energy from the lattice, in the form of phonons, in individual scattering events. The designation of inter- versus intra-valley scattering comes from the multi-valley band-structure model of semiconductors, and refers to whether the initial and final states are in the same valley or in different valleys.

The basic Monte Carlo algorithm may be used to track a single particle over many scattering events in order to simulate the steady-state behavior of a system. However, for improved statistics over shorter simulation times, and for transient simulation, the preferred technique is the use of a *synchronous ensemble* of particles, in which the basic Monte Carlo algorithm is repeated for each particle in an ensemble representing the (usually larger) system of interest until the simulation is completed. Since there is rarely an identical correspondence between the number of simulated charges, and the number of actual particles in a system, each particle is really a *super-particle*, representing a finite number of real particles. The corresponding charge of the particle is weighted by this super-particle number. Figure 3-5 illustrates an ensemble Monte Carlo simulation in which a fixed time step, Δt , is introduced over which the motion of all the carriers in the system is synchronized. The squares illustrate random, instantaneous, scattering events, which may or may not occur during a given time-step. Basically, each carrier is simulated only up to the end of the time-step, and then the next particle in the ensemble is treated. Over each time step, the motion of each particle in the ensemble is simulated independent of the other particles. Nonlinear effects such as carrier-carrier interactions or the Pauli exclusion principle are then updated at each time step.

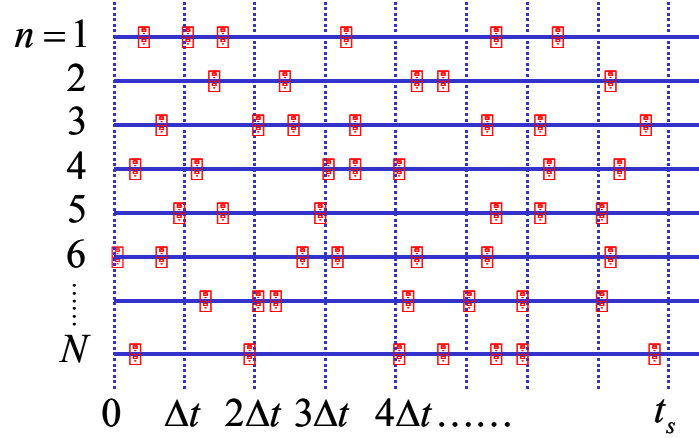


Figure 3-5. Ensemble Monte Carlo simulation in which a time step, Δt , is introduced over which the motion of particles is synchronized. The squares represent random scattering events.

(D) Statistical Averaging

The non-stationary one-particle distribution function and related quantities such as drift velocity, valley or subband population, etc., are then taken as averages over the ensemble at fixed time steps throughout the simulation. For example, the drift velocity in the presence of the field is given by the ensemble average of the component of the velocity at the n th time step as

$$\bar{v}_z(n\Delta t) \cong \frac{1}{N} \sum_{j=1}^N v_z^j(n\Delta t), \quad (3-33)$$

where N is the number of simulated particles and j labels the particles in the ensemble. This equation represents an estimator of the true velocity, which has a standard error given by

$$s = \frac{\sigma}{\sqrt{N}}, \quad (3-34)$$

where σ^2 is the variance which may be estimated from [96]

$$\sigma^2 \cong \frac{N}{N-1} \left\{ \frac{1}{N} \sum_{j=1}^N (v_z^j)^2 - \bar{v}_z^2 \right\}. \quad (3-35)$$

Similarly, the distribution functions for electrons and holes may be tabulated by counting the number of electrons in cells of k -space. From Eq. (3-34), we see that the error in estimated average quantities decreases as the square root of the number of particles in the ensemble, which necessitates the simulation of many particles. Typical ensemble sizes for good statistics are in the range of $10^4 - 10^5$ particles. Variance reduction techniques to decrease the standard error given by Eq. (3-34) may be applied to enhance statistically rare events such as impact ionization or electron-hole recombination.

(E) Ensemble Monte Carlo

An overall flowchart of a typical Ensemble Monte Carlo (EMC) simulation is illustrated in Figure 3-6. After initialization of run parameters, there are two main loops, and outer one which advances the time step by increments of ΔT until the maximum time of the simulation is reached, and an inner loop over all the particles in the ensemble (N), where the Monte Carlo algorithm is applied to each particle individually over a given time step.

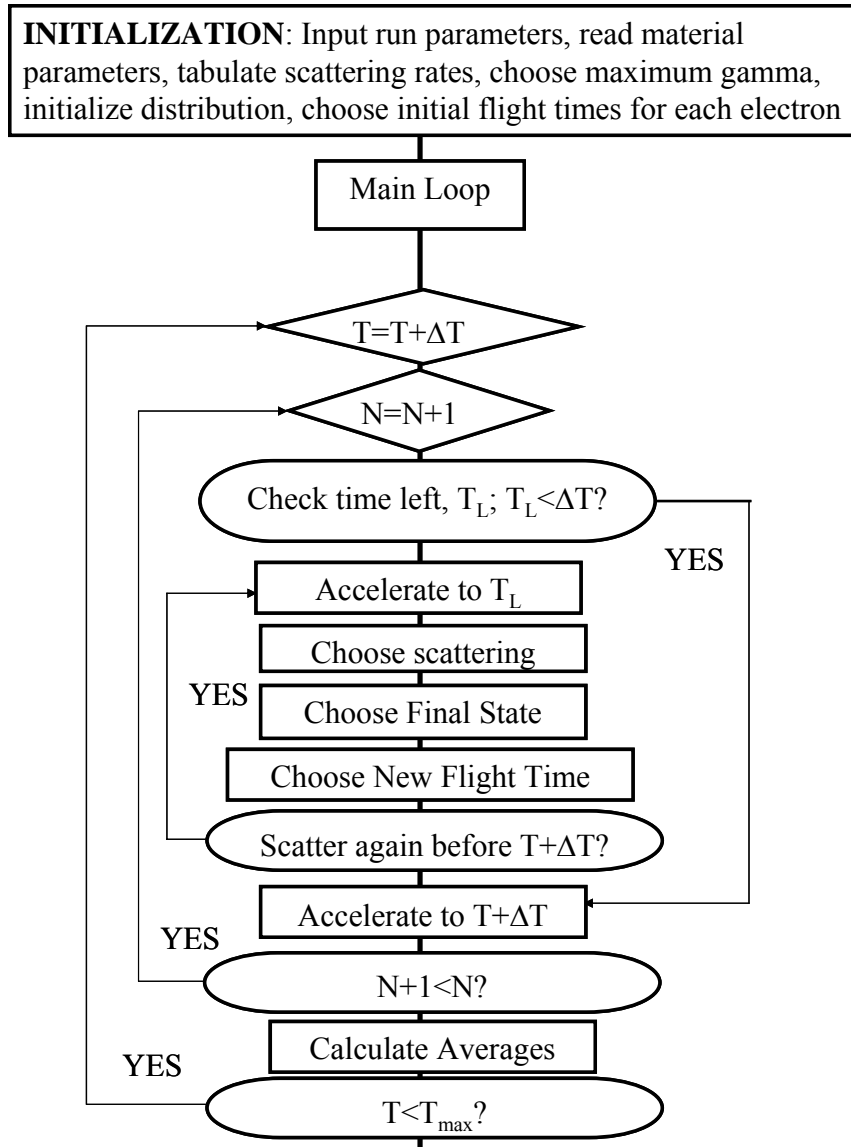


Figure 3-6 Flow chart of an Ensemble Monte Carlo (EMC) simulation.

As an example of the calculated results for the EMC algorithm illustrated in Figure 3-6, Figure 3-7 shows the calculated velocity in the direction of the electric field (drift velocity) versus time for GaAs at 300K for various electric fields, in which the ensemble of carrier is initially in equilibrium, and then a constant electric field is abruptly turned on at zero. The model used here is a non-parabolic three valley model, consisting of a central valley surrounded by satellite valleys in the X and L directions. Scattering mechanisms included are polar optical phonon scattering, acoustic deformation potential scattering, intervalley non-polar optical scattering, and ionized impurity scattering (impurity concentration = $1.0 \times 10^{14}/\text{cm}^3$). The first thing to note is that there is a transient period which may last over several picoseconds, before the carriers reach a steady state situation. For very short times, the motion of particles is almost ballistic (free of scattering) as they accelerate freely in time. As scattering begins to occur, the carrier acceleration slows, and the velocity reaches a peak (overshoot) before settling to a steady state. The overshoot velocity becomes more pronounced at higher fields, and is related to differences in the momentum and energy relaxation times in the system associated with scattering, as well as inter-valley transfer which occurs when carriers are accelerated high enough in energy to overcome the energy difference of the valleys (approximately 0.28 eV in GaAs).

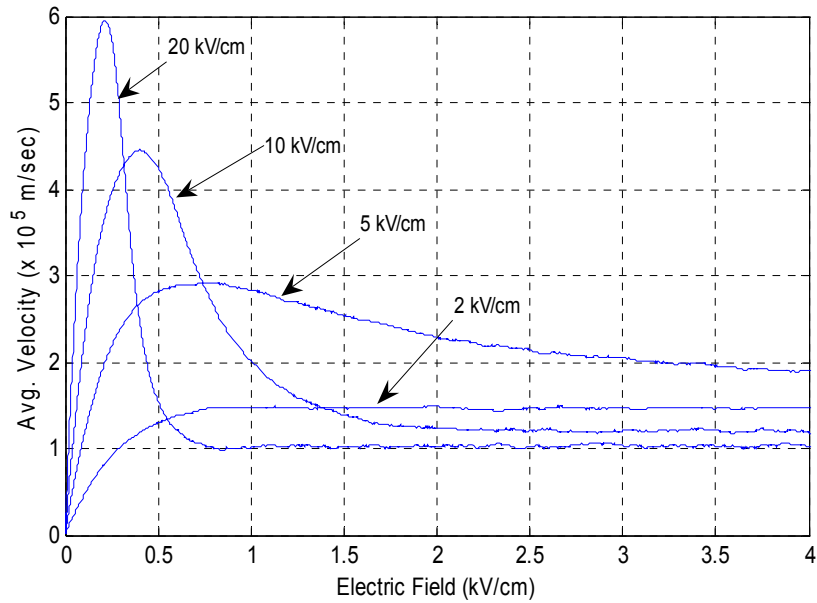


Figure 3-7 Drift velocity versus time in an EMC simulation for electrons in GaAs at 300K for various electric fields.

Figure 3-8 shows the steady state drift velocity versus electric field, calculated by waiting until the electron velocity in Figure 3-7 reaches steady state, and then performing averages in time and over the ensemble to calculate the stationary velocity for a given field. As can be observed, the velocity versus field is initially linear in the field, with the slope given by the low field mobility of GaAs. At the peak of the velocity-field curve, the velocity saturates and then decreases, due to the transfer of carriers from the higher mobility central valley, to the lower mobility satellite valleys. This mechanism is responsible for a region of negative resistance and corresponding Gunn oscillations due to the ensuing instability associate with negative resistance [97].

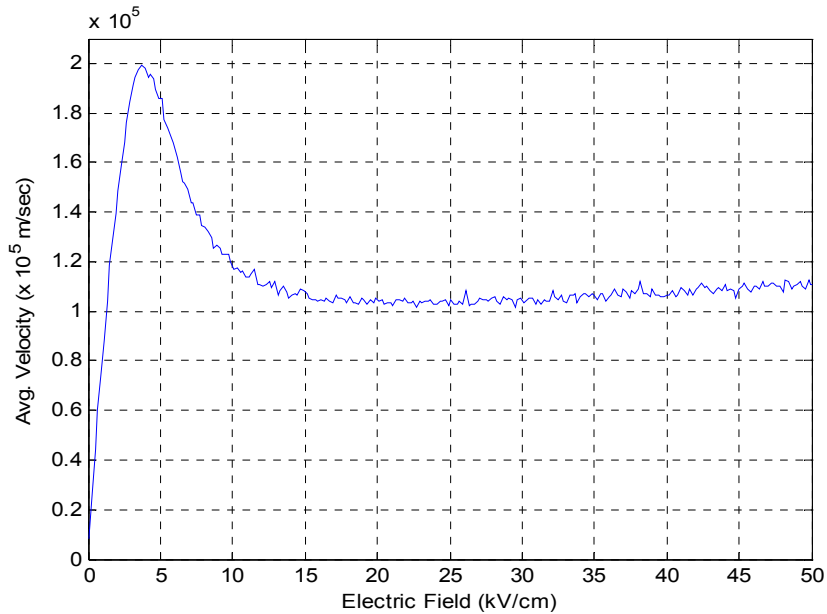


Figure 3-8 Average drift velocity versus electric field for Bulk GaAs at 300K.

3.2 The Wigner Distribution Function

The Wigner distribution function is a mathematical transform of the density matrix which approaches the classical distribution function f_c as the system becomes classical (with large dimensions, slowly varying potentials, and/or high temperatures) [98,99]. This representation of the statistical state has proven to be useful in modeling quantum-effect devices such as the resonant-tunneling diode [100,101].

To derive the Wigner function from the density matrix $\rho(x, x')$ defined in Eq. (3-6) one rewrites the arguments (x, x') as $r = \frac{1}{2}(x + x')$ and $r' = x - x'$, and then Fourier transforms r' into a momentum variable p . Thus

$$f_W(r, p) = \int_{-\infty}^{+\infty} dr' \rho(r + \frac{1}{2}r', r - \frac{1}{2}r') e^{-ipr'/\hbar} \quad (3-36)$$

Applying the same procedure (which is known as the Wigner-Weyl transformation) to the Liouville-von Neumann equation gives:

$$\frac{\partial f_W}{\partial t} = -\frac{p}{m^*} \frac{\partial f_W}{\partial r} - \frac{1}{\hbar} \int_{-\infty}^{+\infty} \frac{dp'}{2\pi\hbar} V_W(r, p - p') f_W(r, p') \quad (3-37)$$

where the kernel of the potential operator is given by:

$$V_W(r, p) = 2 \int_0^{\infty} dr' \sin(pr'/\hbar) [V(r + \frac{1}{2}r') - V(r - \frac{1}{2}r')] \quad (3-38)$$

Let us examine the form of these equations. Because Eq. (3-37) is derived from Eq. (3-6) by a mathematical transformation, we would expect that it should also describe unitary time evolution. The condition for unitary evolution is that L_w be an anti-Hermitian operator. The potential operator is anti-Hermitian [because $V_w(r, -p) = -V_w(r, p)$], and the drift term is anti-Hermitian if periodic boundary conditions are imposed. On the other hand, we have seen that if initial conditions are imposed, the drift term is a master operator, and the equation then describes irreversible time evolution. This is the origin of the usefulness of the Wigner representation for describing electron devices. One applies boundary conditions to f_w so as to fix the distribution of electrons entering the domain:

$$\begin{aligned} f_W(x_l, p) |_{p>0} &= f_l(p) \\ f_W(x_r, p) |_{p<0} &= f_r(p) \end{aligned} \quad (3-39)$$

where $f_l(p)$ and $f_r(p)$ are the distribution functions in the left- and right-hand contacts (reservoirs), respectively. Because these boundary conditions introduce irreversibility into the Liouville equation, one can now evaluate the time-evolution of a device, and observe an approach to steady-state. Inelastic processes such as phonon scattering may be included in a semi-classical way by adding the Boltzmann collision term [102], or by even simpler schemes such as the relaxation time approximation [101]. The open-system Wigner function approach has proved to be of use in understanding the behavior of resonant-tunneling diodes. This technique permits evaluation of steady-state behavior in the form of the $I(V)$ curve, and calculations of the large-signal transient response and small-signal ac response. The $I(V)$ curves derived from this model show the expected negative differential-conductance region, but the ratio of the peak to valley currents is always smaller than that obtained from the tunneling theory, and is often less than that observed experimentally. Recently, Tsuchiya and co-workers have developed an improved formulation of the Liouville equation which takes the spatial variation of the effective mass into account, and which leads to larger peak-to-valley ratios than the simpler theory [103].

3.3 Green's Functions

In single-time approaches, such as the density matrix (first proposed and used by von Neumann [104]) and Wigner function approaches [98], the spectral density function is integrated out of the problem, but the full spatially non-local nature of the potential interactions is retained. The problem with the density matrix in many semiconductor applications is that it is a real space function, and only considers quantum interference effects occurring between two separated points in space. When the physical problem is one that is better understood in terms of phase-space distributions, the Wigner function approach has an advantage over the density matrix method.

Both of the previously mentioned approaches include correlations in space, but they do not consider that there may be correlations in the time domain (they are Markovian in time). Transport equations that overcome the limitations of the previously mentioned methods and include the quantum effects properly, may be derived by the non-equilibrium Green's function technique. The introduction of quantum-field theoretical methods in non-equilibrium statistical mechanics evolved from the work of Martin and Schwinger [105] and Schwinger [106]. Significant further developments in this area were due to Kadanoff and Baym [107]. Essentially the same theory was developed by Konstantinov and Perel [108], Dzyaloshinski [109], Keldysh [110], Abrikosov *et al.* [111] and others. The equivalence between the Kadanoff-Baym and Keldysh formalism was elegantly demonstrated by Langreth [112]. Several review articles focusing on different aspects of this subject exist in the literature [113,114].

Green's functions or correlation functions [115], which are thermodynamic averages of the products of field operators $\hat{\Psi}(x_1)$ and $\hat{\Psi}^+(x_2)$, play a fundamental role in statistical physics and they constitute the connection between experimentally relevant quantities and conveniently calculable ones.

In the discussion of the linear response of the system, we already introduced the retarded and advanced Green's functions through the definitions

$$G_r(x_1, x_2) = -\frac{i}{\hbar} \theta(t_1 - t_2) \left\langle \left\{ \hat{\Psi}(x_1), \hat{\Psi}^+(x_2) \right\} \right\rangle \quad (3-40)$$

and

$$G_a(x_1, x_2) = \frac{i}{\hbar} \theta(t_2 - t_1) \left\langle \left\{ \hat{\Psi}(x_1), \hat{\Psi}^+(x_2) \right\} \right\rangle, \quad (3-41)$$

respectively. As previously mentioned, the brackets symbolize an ensemble average, or a summation over the proper basis. In general, at zero temperature and in equilibrium, the brackets denote the ground state of the interacting system. At finite temperature and in equilibrium, the brackets imply thermodynamic average over all possible states of the system. For systems out of equilibrium, the brackets signify the need to average over the available states of the system, since the concept of thermodynamic averaging is 'ill defined'.

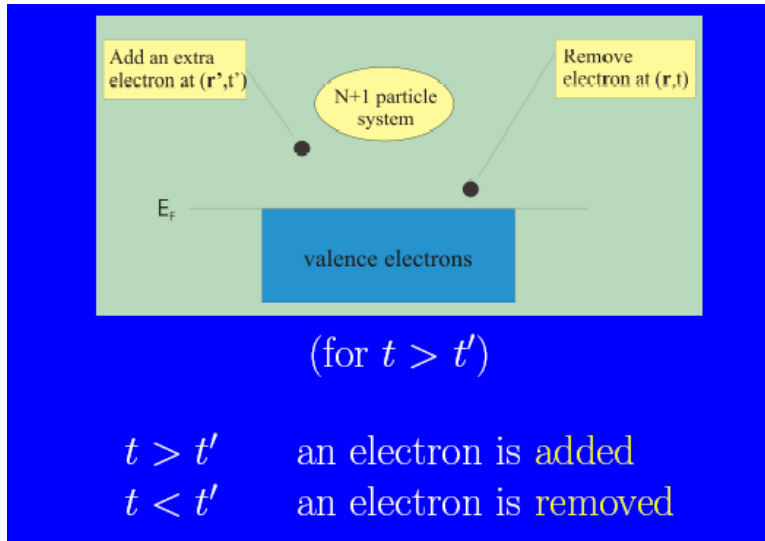


Figure 3-9. Cartoon of single-electron Green's function.

For simple Hamiltonians, the Green's functions (see Figure 3-9) can be obtained from the Schrödinger or Liouville equation. For complicated Hamiltonians, one usually applies the perturbation approach (see Figure 3-10 for simple explanation) in which the generation of the perturbation series relies upon the S-matrix expansion of the unitary operator

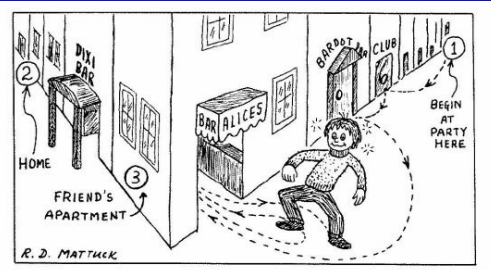
$$T \exp \left[-\frac{i}{\hbar} \int_{t_1}^{t_2} dt \hat{V}(t) \right]$$

where $\hat{V}(t)$ is the perturbing potential interaction written in the interaction representation and T is the time-ordering operator. The time-ordering operator placed in front of several functions of time is actually an instruction to order the functions according to their time-ordering with the earliest time to the right.

A drunken man is going home (2) from a party (1).
On his way home are many bars ...

Possibility that the man arrives at (2):

$$\begin{aligned}
 P(2, 1) &= P_0(2, 1) && \text{(zero order)} \\
 &+ P_0(2, A)P(A)P_0(A, 1) + P_0(2, B)P(B)P_0(B, 1) + \dots && \text{(first order)} \\
 &+ P_0(2, B)P(B)P_0(B, A)P(A)P_0(A, 1) + \dots && \text{(second order)} \\
 &+ \dots && \text{(higher orders)}
 \end{aligned}
 \tag{8}$$



[from R. D. Mattuck, *A Guide to Feynman Diagrams in the Many-Body Problem* (Dover, New York, 1992)]

Figure 3-10. Classical analog of the perturbation series expansion – the drunken man problem.

In equilibrium situations, the states of the system are well defined for times $t_1 \rightarrow -\infty$ and $t_2 \rightarrow +\infty$. However, the above scheme can not be applied to non-equilibrium situations, which is the normal case in nearly all active semiconductor devices. The basic reason is the fact that, in general, within the evolution, no state of the system in the future may be identified with any of the states in the past. The procedure for avoiding this dilemma was proposed by Blandin *et al.* [116]. They suggested that the integration path in the S-matrix be a time-loop shown in Figure 3-11.

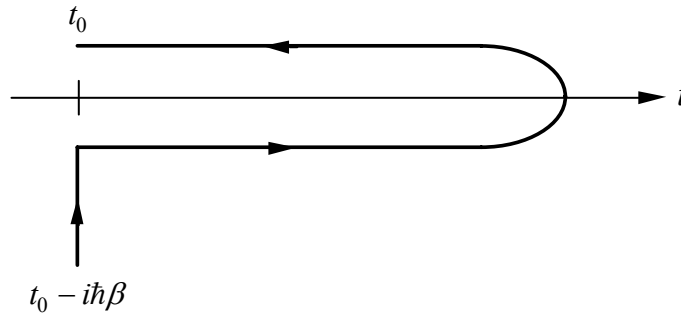


Figure 3-11. Integration path for real-time Green's functions proposed by Blandin *et al.*

With this choice of the contour, the evolution of the system is assumed to begin with a thermal Green's function at $t' = t_0 - i\hbar\beta$, where β is the inverse temperature. It then evolves into the non-equilibrium, but non-interacting Green's function at t_0 . The contour then extends forward in time, up to the maximum of (t_1, t_2) , at which point it is returned backward in time at t_0 , where presumably, one knows the states of the

system. For situations, where initial correlations can be neglected, such as far-from-equilibrium stationary transport, we can discard the contribution of the contour from $t_0 - i\beta$ to t_0 and invoke the limit $t_0 \rightarrow -\infty$.

When dealing with nonequilibrium situations, one also needs to define less-than

$$G^<(x_1, x_2) = \frac{i}{\hbar} \langle \hat{\Psi}^+(x_2) \hat{\Psi}(x_1) \rangle \quad (3-42)$$

and greater-than

$$G^>(x_1, x_2) = -\frac{i}{\hbar} \langle \hat{\Psi}(x_1) \hat{\Psi}^+(x_2) \rangle \quad (3-43)$$

correlation functions, which have time arguments that are on different branches of the contour, as shown on Figure 3-12a. One can also define Green's functions that have time arguments on the same branch of the contour (Figure 3-12b). These are time-ordered

$$G_t(x_1, x_2) = \theta(t_1, t_2) G^>(x_1, x_2) + \theta(t_2, t_1) G^<(x_1, x_2) = G_r + G^< = G_a + G^> \quad (3-44)$$

and anti-time-ordered

$$G_{\bar{t}}(x_1, x_2) = \theta(t_2, t_1) G^>(x_1, x_2) + \theta(t_1, t_2) G^<(x_1, x_2) = G^> - G_r = G^< - G_a \quad (3-45)$$

Green's functions. The function $\theta(t_1, t_2)$ is defined on the contour with the property: $\theta(t_1, t_2) = 1$ if t_1 is later on the contour than t_2 and $\theta(t_1, t_2) = 0$ if earlier.

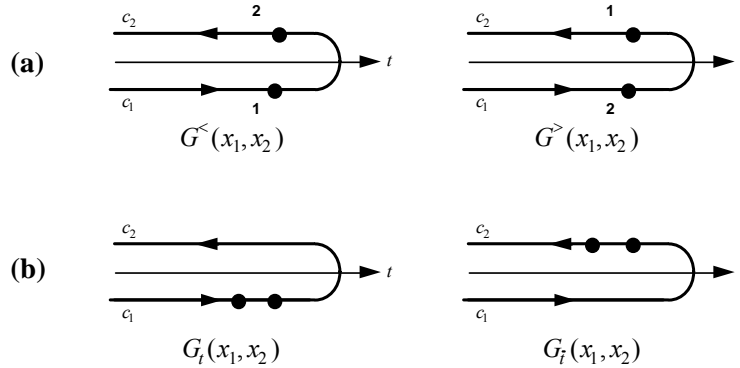


Figure 3-12. Definition of the four Green's functions that have time arguments on different legs (a) and same legs (b) of the time-loop path.

The natural question that arises after this rather formal introduction of the various Green's functions is whether we can associate some physical meaning to them. If we look at the RHS of the equation for $G^<$, for $x_1 = x_2$ it is proportional to the particle density, and for equal times it corresponds to the reduced single-particle density matrix. Hence, this quantity describes the kinetics of the system. By similar arguments, one finds that $G^>$ may be seen as corresponding to the density of the missing particles (holes). The spectrally decomposed time-ordered and anti-time-ordered Green's functions possess propagator forms, whereas the retarded and advanced Green's functions contain the spectral properties of the system, i.e., the information of the single-particle energy renormalizations.

Similar types of Green's functions may also be defined for phonon fields:

$$D^>(x_1, x_2) = -i \langle \hat{H}_{e-ph}(x_1) \hat{H}_{e-ph}(x_2) \rangle \quad (3-46)$$

$$D^<(x_1, x_2) = -i \langle \hat{H}_{e-ph}(x_2) \hat{H}_{e-ph}(x_1) \rangle \quad (3-47)$$

$$D_t(x_1, x_2) = \theta(t_1, t_2) D^>(x_1, x_2) + \theta(t_2, t_1) D^<(x_1, x_2) \quad (3-48)$$

$$D_{\bar{t}}(x_1, x_2) = \theta(t_2, t_1) D^>(x_1, x_2) + \theta(t_1, t_2) D^<(x_1, x_2) \quad (3-49)$$

$$D_r(x_1, x_2) = D_t - D^< = D^> - D_{\bar{t}} - \theta(t_1, t_2) (D^> - D^<) \quad (3-50)$$

$$D_a(x_1, x_2) = D_t - D^> = D^< - D_{\bar{t}} - \theta(t_2, t_1) (D^> - D^<) \quad (3-51)$$

where $\hat{H}_{e-ph}(x)$ is the second quantized form of the perturbation due to the electron-phonon interaction

$$\hat{H}_{e-ph}(x) = \sum_{\mathbf{q}} M_{\mathbf{q}\lambda} e^{i\mathbf{q}\cdot\mathbf{r}} \left(\hat{a}_{\mathbf{q}\lambda} e^{-i\omega_{\mathbf{q}\lambda} t} + \hat{a}_{-\mathbf{q}\lambda}^+ e^{i\omega_{\mathbf{q}\lambda} t} \right). \quad (3-52)$$

The main difference between the expressions for fermion and boson fields is that $D^>$ and $D^<$ have the same sign which follows from commutation relations obeyed by boson fields. In equilibrium, greater-than and less-than Green's functions can be expressed in terms of the phonon occupation number

$$N_{\mathbf{q}\lambda} = \langle \hat{a}_{\mathbf{q}\lambda}^+ \hat{a}_{\mathbf{q}\lambda} \rangle:$$

$$D^>(x_1, x_2) = -i \sum_{\mathbf{q}} |M_{\mathbf{q}\lambda}|^2 \left[(N_{\mathbf{q}\lambda} + 1) e^{-i\omega_{\mathbf{q}\lambda}(t_1-t_2)} + N_{\mathbf{q}\lambda} e^{i\omega_{\mathbf{q}\lambda}(t_1-t_2)} \right] e^{i\mathbf{q}\cdot(\mathbf{r}_1-\mathbf{r}_2)} \quad (3-53)$$

$$D^<(x_1, x_2) = -i \sum_{\mathbf{q}} |M_{\mathbf{q}\lambda}|^2 \left[(N_{\mathbf{q}\lambda} + 1) e^{i\omega_{\mathbf{q}\lambda}(t_1-t_2)} + N_{\mathbf{q}\lambda} e^{-i\omega_{\mathbf{q}\lambda}(t_1-t_2)} \right] e^{i\mathbf{q}\cdot(\mathbf{r}_1-\mathbf{r}_2)}. \quad (3-54)$$

The equilibrium form of the retarded phonon Green's function is

$$D_r(x_1, x_2) = -2\theta(t_1 - t_2) \sum_{\mathbf{q}} |M_{\mathbf{q}\lambda}|^2 e^{i\mathbf{q}\cdot(\mathbf{r}_1-\mathbf{r}_2)} \sin[\omega_{\mathbf{q}\lambda}(t_1 - t_2)]. \quad (3-55)$$

The other three Green's functions are found easily from these three. The meaning of the various boson Green's functions is equivalent to the one for fermion fields.

A quantity that now possesses a simple perturbation expansion on the contour, is the contour-ordered Greens function

$$G_{c_k}(x_1, x_2) = -i \langle T_{c_k} \left(\hat{\Psi}(x_1) \hat{\Psi}^+(x_2) \right) \rangle \quad (3-56)$$

where c_k is the contour from Figure 3.12 and T_{c_k} is the contour-ordering operator. To obtain a perturbation expansion for the contour-ordered Green's function, one could employ the standard functional derivative method due to Schwinger, as explained in detail by Kadanoff and Baym, or use the equivalent procedure based on the statistical Wick's theorem. The Feynman rules, which result from the application of the Wick decomposition to a perturbation expansion, are similar to the conventional ground-state Feynman rules. The only difference is that to each line, we now associate the contour-ordered Green's function, which can also be written in the matrix notation [117,118]

$$\mathbf{G}_{\mathbf{C}}(x_1, x_2) = \begin{bmatrix} G_t & G^< \\ G^> & G_{\bar{t}} \end{bmatrix} \quad (3-57)$$

by the prescription that the ij-component of the $\mathbf{G}_{\mathbf{C}}$ -matrix (or Keldysh matrix) be defined as $G_{c_k}(x_1, x_2)$

for t_1 and t_2 residing on the segments c_i and c_j , respectively. For example, we can write the 11 element

by assuming contour ordering as

$$G_{11}(x_1, x_2) = \theta(t_1, t_2)G^>(x_1, x_2) + \theta(t_2, t_1)G^<(x_1, x_2) \equiv G_t(x_1, x_2) . \quad (3-58)$$

In a similar manner the 22-term yields the anti-time ordered Green's function G_t^- . The 12-term is the one that has time t_1 on leg c_1 and time t_2 on leg c_2 . Since the upper branch of the contour (leg c_2) corresponds to later times compared to the lower branch (leg c_1), t_2 is later on the contour than t_1 . Thus, the contour ordering operator in Eq. (3-56) will flip the order of the creation and annihilation operators, which results in a sign change into the definition of the contour-ordered Green's function. Then

$$G_{12}(x_1, x_2) = \frac{i}{\hbar} \langle \hat{\Psi}^+(x_2) \hat{\Psi}(x_1) \rangle = G^<(x_1, x_2) . \quad (3-59)$$

By similar arguments, the 21-term becomes $G^>$.

The components of \mathbf{G}_C are not linearly independent, and by performing a nonunitary transformation it is possible to remove part of the redundancy. Such a coordinate transformation, often called *rotation in Keldysh space*, is given by $\mathbf{G}_K \rightarrow \mathbf{L} \tau_3 \mathbf{G}_C \mathbf{L}^+$, where

$$\mathbf{L} = \frac{1}{\sqrt{2}} \begin{bmatrix} 1 & -1 \\ 1 & 1 \end{bmatrix}, \quad \mathbf{L}^+ = \frac{1}{\sqrt{2}} \begin{bmatrix} 1 & 1 \\ -1 & 1 \end{bmatrix} \quad \text{and} \quad \tau_3 = \begin{bmatrix} 1 & 0 \\ 0 & -1 \end{bmatrix} . \quad (3-60)$$

The new matrix Keldysh Green's function is of the form

$$\mathbf{G}_K(x_1, x_2) = \begin{bmatrix} G_r & G_K \\ 0 & G_a \end{bmatrix}, \quad (3-61)$$

where $G_K = G^> + G^<$ is the so-called Keldysh Green's function.

Under the assumption that the individual field operators are based upon wave-functions which satisfy the Schrödinger equation, one can calculate the equations of motion for the various Green's functions. For single-point potentials, such as those arising from the solution of the Poisson equation, the equations of motion for \mathbf{G}_K are of the form:

$$\left(i\hbar \frac{\partial}{\partial t_1} - H_o(\mathbf{r}_1) - V(\mathbf{r}_1) \right) \mathbf{G}_{K0}(x_1, x_2) = \delta(x_1, x_2) \mathbf{I} , \quad (3-62)$$

$$\left(-i\hbar \frac{\partial}{\partial t_2} - H_o(\mathbf{r}_2) - V(\mathbf{r}_2) \right) \mathbf{G}_{K0}(x_1, x_2) = \delta(x_1, x_2) \mathbf{I} , \quad (3-63)$$

where \mathbf{I} is the identity matrix, $V(\mathbf{r})$ is the single-point potential and the subscript "0" is added to indicate the non-interacting form of the Green's functions. These, so-called bare Green's functions, in the *quasi-particle picture* often used in quantum-field theory represent the *bare particle* just added to the system.

We have stated several times earlier that transport arises as a balance between the driving forces and the dissipative forces. In the Green's functions formalism, these dissipative mechanisms are introduced through the self-energy terms. Using the perturbation expansion, one can define the self-energy functions Σ as an irreducible part of the Green's function. The self-energy may also in principle be introduced variationally. For electrons in a solid, important self-energy contributions are provided by the driving forces and electron interaction with other electrons, ions in the lattice, impurities and phonons. Expressing the self-energy functions in the matrix form, it is possible to write the equations of motion for the full Green's function as

$$\left(i\hbar \frac{\partial}{\partial t_1} - H_o(\mathbf{r}_1) - V(\mathbf{r}_1) \right) \mathbf{G}_K(x_1, x_2) = \delta(x_1, x_2) \mathbf{I} + \int dx_3 \Sigma_K(x_1, x_3) \mathbf{G}_K(x_3, x_2) , \quad (3-64)$$

$$\left(-i\hbar \frac{\partial}{\partial t_2} - H_o(\mathbf{r}_2) - V(\mathbf{r}_2) \right) \mathbf{G}_K(x_1, x_2) = \delta(x_1, x_2) \mathbf{I} + \int dx_3 \mathbf{G}_K(x_1, x_3) \Sigma_K(x_3, x_2) , \quad (3-65)$$

where the self-energy matrix is the one that has the Keldysh form

$$\Sigma_K = \begin{bmatrix} \Sigma_r & \Sigma_K \\ 0 & \Sigma_a \end{bmatrix} . \quad (3-66)$$

The self-energy functions (see Figure 3-13) introduced in Eq. (3-66) can be either *one-point* or *two-*

point functions [119]. One-point functions are those self-energy functions which only depend upon a single time variable. Those which depend upon two time variables are two-point functions. One-point functions seem mostly to arise from the perturbations due to external fields. The two-point self-energy functions arise from the particle interactions which cause the particle to scatter. One example is particle-phonon interactions. Particle-particle interactions are usually taken to be instantaneous: the important terms of Hartree and exchange are often approximated as one-point functions. However, further contributions such as correlation or screened exchange are two-point functions. Impurities introduce a potential which is a one-point function, but the multiple scattering from this impurity makes the self-energy a two-point function. The self-energy functions can also be *local* (one-point potentials) and *nonlocal* (two-point functions) in space. In the quasi-particle picture, the full Green's function represents the *dressed particle*.

The corresponding Dyson's equations for the Keldysh matrix Green's function, given in Eq. (3-61), are

$$\mathbf{G}_K = \mathbf{G}_{K0} + \mathbf{G}_{K0} \Sigma_K \mathbf{G}_K \quad (3-67)$$

$$\mathbf{G}_K = \mathbf{G}_{K0} + \mathbf{G}_K \Sigma_K \mathbf{G}_{K0} \quad (3-68)$$

where the matrix \mathbf{G}_{K0} explicitly depends on the initial distribution. The product notation on the RHS of Eqs. (3-67) and (3-68) represents integration over the internal variables. From the matrix equations given in Eq. (3-67), one can find the equations obeyed by the various Green's functions. The equations of motion for the less-than and greater-than Green's functions are

$$G^{>,<} = \left(1 + G_r \Sigma_r\right) G_o^{>,<} \left(1 + \Sigma_a G_a\right) + G_r \Sigma^{>,<} G_a . \quad (3-69)$$

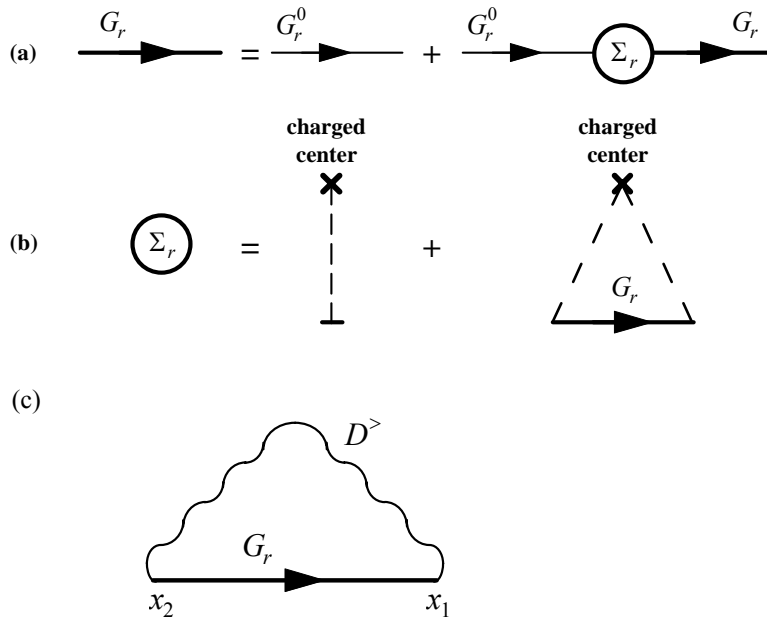
Because the self-energy functions $\Sigma^{>,<}$ depend upon $G^{>,<}$, the equations of motion given in Eq. (3-69) are integral equations. The equations of motion for time-ordered and anti-time-ordered Green's functions are

$$G_{t,\bar{t}} = \left(1 + G_r \Sigma_r\right) G_{ot,o\bar{t}} \left(1 + \Sigma_a G_a\right) + G_r \Sigma_{\bar{t},t} G_a . \quad (3-70)$$

The retarded and advanced Green's functions satisfy the Dyson equation

$$G_{r,a} = G_{r,a}^o + G_{r,a}^o \Sigma_{r,a} G_{r,a} . \quad (3-71)$$

While six different Green's functions were introduced in this section, only four of these are considered to be independent in the general non-equilibrium situation, and only two are independent in the equilibrium system.



- (a) Dyson's equation for the retarded Green's function. Thin line corresponds to the bare (non-interacting) Green's function. Thick lines represent the interacting (full, or renormalized) Green's function.
- (b) Self-consistent Born approximation for scattering from impurities.
- (c) Self-consistent Born approximation for electron-phonon interaction.

Figure 3-13. Feynman diagrams representation of the Dyson equation (a), single point (b) and two-point (b) self-energy functions.

In general, the equation for $G^<$ and the Dyson equation are coupled equations: for example, the retarded self-energy may depend on the less-than correlation function. This leads to enormous complications, and one usually attempts to split the calculation into two steps: first the retarded/advanced Green's functions are determined, and then used as an input to the equation for $G^<$. Since many one-electron properties of the system (density of states, scattering rates and quasi-particle lifetimes) require the knowledge of the spectral density function $A = i(G_r - G_a)$, the solution of the Dyson equation for G_r (see Figure 3-14), apart from serving as an input to the equation for $G^<$, is of particular interest. For example, after integrating the spectral density function over momentum states, one gets the density-of-states function. Once retarded, advanced and less-than Green's functions are known, all other Green's functions are calculated as a linear combination of these three. The less-than Green's function is actually the quantity we usually want, because various moments of this function will give us the quantities of interest, such as particle occupation of particular state, current density, etc.

The Green's function is conveniently obtained from the Dyson equation.
 Exemplary derivation of the Dyson equation using Hartree and exchange terms

$$\begin{aligned}
 \uparrow &= \uparrow + \uparrow \text{---} \bigcirc + \uparrow \text{---} \bigcirc \text{---} \uparrow + \uparrow \text{---} \bigcirc \text{---} \bigcirc \text{---} \uparrow + \uparrow \text{---} \bigcirc \text{---} \bigcirc \text{---} \bigcirc \text{---} \uparrow + \dots \\
 &= \left[1 + \uparrow \text{---} \bigcirc + \uparrow \text{---} \bigcirc \text{---} \uparrow + \uparrow \text{---} \bigcirc \text{---} \bigcirc \text{---} \uparrow + \uparrow \text{---} \bigcirc \text{---} \bigcirc \text{---} \bigcirc \text{---} \uparrow + \dots \right] \uparrow \\
 &= \sum_n \left[\uparrow \text{---} \bigcirc + \uparrow \text{---} \bigcirc \text{---} \uparrow \right]^n \uparrow \\
 &= \frac{1}{1 - \left[\uparrow \text{---} \bigcirc + \uparrow \text{---} \bigcirc \text{---} \uparrow \right]} \uparrow \\
 &= \frac{1}{1 - \uparrow \text{---} \left[\text{---} \bigcirc + \text{---} \bigcirc \text{---} \uparrow \right]} \uparrow = \frac{1}{1 - \uparrow \Sigma} \uparrow
 \end{aligned}$$

Σ : (irreducible or proper) self-energy

Figure 3-14. Diagrammatic representation of the solution of the Dyson equation for the advanced and the retarded Green's functions.

Starting from the equations of motion for less-than and retarded Green's functions, Mahan and Hansch [120] obtained the correct gauge-invariant equation for the quantum distribution function $f(\mathbf{k}, \omega, \mathbf{r}, t)$ for bulk semiconductors in linear response regime. Very recently, an alternative approach has been proposed in which one transforms to a new basis defined by the eigenfunctions of the external potential. For the uniform field case, this means that the kinetic equation and the Dyson equation are Airy-transformed [121,122,123]. The basic idea for this approach was to treat collisional broadening and intra-collisional field-effect on equal footing. The equivalence between the Kubo formula and Generalized Kadanoff-Baym equation (in linear response) was also demonstrated [124]. Real-time Green's functions were also used in the formalism developed by Jauho and Wilkins [125] that allows a nonperturbative calculation of the effect of the electric field on electron-impurity scattering. The extension of the approach presented in this section, to include the effect of the initial correlations, has been carried out by Wagner [126]. We have applied the Green's functions formalism to study transport properties of silicon inversion layer using all relevant scattering mechanisms that dominate the transport properties of the structure at different temperatures. We are currently investigating ballistic transport in FinFET devices using the CBR technique for solving the Green's function problem introduced in Chapter 4 of this review article.

4. CBR Method for the Solution of the 3D Green's Function Method as Applied to Modeling 2D/3D FinFET Devices

An efficient method based on Green's function approach, termed as Contact Block Reduction (CBR) method^{127,128} that is presented next has been developed at Walter Schottky Institute and ASU and used by the group from ASU to calculate self-consistently transport properties in nanoscale 10 nm gate length FinFET device operating in the ballistic regime. The method rigorously separates the open system problem into the solution of a suitably defined closed system (energy-independent) eigen-problem and the energy-dependent solution of a small linear system of equations of size determined by the contact regions that couple the closed system to the leads. The calculation of the charge density of the open system throughout the device can be performed with an effort comparable to a single calculation of a small percentage of the eigenstates of a closed system.

The CBR method allows one to calculate 2D or 3D ballistic transport properties of a device that may have any shape, potential profile, and most importantly any number of external leads. In this method, quantities like the transmission function and the charge density of an open system can be obtained from the eigenstates of the corresponding closed system defined as $H^0 |a\rangle = e_a |a\rangle$, and the solution of a very small linear algebraic system for every energy step E . The retarded Green's function $\mathbf{G}^R(E)$ can be calculated via the Dyson equation through a Hermitian Hamiltonian \mathbf{H}^0 of a closed system represented by¹²⁹,

$$\mathbf{G}^R(E) = \mathbf{A}^{-1}(E) \mathbf{G}^0(E), \quad \mathbf{A}(E) = \mathbf{I} - \mathbf{G}^0(E) \mathbf{S}(E) \quad (4-1)$$

$$\mathbf{G}^0(E) = \sum_a \frac{|a\rangle\langle a|}{E - e_a}$$

The inversion of the matrix \mathbf{A} can be easily performed using the property of the self-energy \mathbf{S} in real space representation: it is non-zero only at boundary regions of the device, which are in *contact* with the external leads. We denote these boundary regions (=contacts) with index C , and the rest of the device with index D (see Figure 4-1).

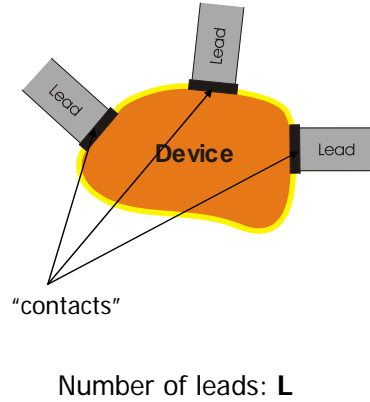


Figure 4-1. Graphical description of the device and the leads.

As a result, the Green's function matrix of the open system can be written in the following form:

$$\mathbf{G}^R = \begin{pmatrix} \mathbf{G}_C^R & \mathbf{G}_{CD}^R \\ \mathbf{G}_{DC}^R & \mathbf{G}_D^R \end{pmatrix} \quad (4-2)$$

$$= \begin{pmatrix} \mathbf{A}_C^{-1} \mathbf{G}_C^0 & \mathbf{A}_C^{-1} \mathbf{G}_{CD}^0 \\ \mathbf{A}_{DC} \mathbf{A}_C^{-1} \mathbf{G}_C^0 + \mathbf{G}_{DC}^0 & -\mathbf{A}_{DC} \mathbf{A}_C^{-1} \mathbf{G}_{CD}^0 + \mathbf{G}_D^0 \end{pmatrix}$$

The left-upper matrix block $\mathbf{G}_C^R = \mathbf{A}_C^{-1} \mathbf{G}_C^0$ fully determines the transmission function whereas the left-lower block \mathbf{G}_{DC}^R determines the density of states, charge density, etc. The particle density $n(\mathbf{r})$ can be obtained using,

$$n(\mathbf{r}) = \sum_{a,b} \langle \mathbf{r} | a \rangle \langle b | \mathbf{r} \rangle x_{ab}, \quad (4-3)$$

where x_{ab} is the density matrix and is given by,

$$x_{ab} = \mathop{\text{a}}\limits_0^L \int_0^L X_{ab}^{(l)}(E) f_l(E) dE,$$

$$X_{ab}^{(l)}(E) = \frac{1}{2p} \frac{\text{Tr}(\langle a | \mathbf{B}_C^{-1} \mathbf{G}_C^{(l)} \mathbf{B}_C^{-1} | \rangle)}{(E - e_a + ih)(E - e_b - ih)} \Big|_{h \rightarrow 0^+}, \quad (4-4)$$

$$\mathbf{G}_C = i \mathbf{S}_C - \mathbf{S}_C^\dagger \mathbf{B}_C = \mathbf{1}_C - \mathbf{S}_C \mathbf{G}_C^0.$$

In Eq. (4-4), L denotes the total number of external leads of the device, index l denotes individual lead number and $f_l(E)$ is the distribution function associated with lead l . The integration in Eq. (4-4) is performed over the energy interval, where both the density matrix distribution $X_{ab}^{(l)}(E)$ and the distribution function $f_l(E)$ are non-negligible. Consequently, the density matrix distribution defines the lower integration limit, and the distribution function $f_l(E)$ the upper integration limit. The advantage of using Eq. (4-2)-(4-4) for determining electron density is in *splitting numerical costs* between calculation of position-independent density matrix and position-dependent, but energy-independent charge density in Eq. (4-3). Then the total numerical cost can be estimated as $N_{nc} = N_{eigen}^2 N_E + N_{eigen}^2 N_{grids}$, where N_E is number of energy steps, N_{eigen} is number of eigenstates to be used, and N_{grids} is the number of grid points in real space. Note the absence of a large terms like $N_E \cdot N_{grids}$.

However, a slightly different approach to calculate particle density can be adopted that is also very efficient. This approach appears to be more suitable for self-consistent calculation. For a self-consistent calculation using a predictor-corrector approach described below, it is important to have an expression for the local density of states (LDOS), $r(\mathbf{r}, E)$. To obtain the expression for the LDOS using CBR algorithm, we note that the lower-left block \mathbf{G}_{DC}^R of the matrix in Eq. (4-2) can be also written in the following form,

$$\mathbf{G}_{DC}^R = \mathbf{G}_{DC}^0 \mathbf{B}_C^{-1} \quad (4-5)$$

Next, using the formula $r(\mathbf{r}, E) = \langle \mathbf{r} | \mathbf{G}^R \mathbf{G} \mathbf{G}^{R\dagger} | \mathbf{r} \rangle / 2p$ and performing simple algebraic manipulations one gets

$$r(\mathbf{r}, E) = \frac{1}{2p} \mathop{\text{a}}\limits_m \left| G_{rm}^R \right|^2 G_{mm},$$

$$G_{rm}^R = \langle \mathbf{r} | \mathbf{G}_{DC}^0 \mathbf{B}_C^{-1} | m \rangle \quad (4-6)$$

$$= \mathop{\text{a}}\limits_{m', a} \frac{\langle \mathbf{r} | a \rangle \langle a | m' \rangle}{E - e_a} \langle m' | \mathbf{B}_C^{-1} | m \rangle.$$

The term G_{rm}^R in Eq. (4-6) is the retarded Green's function in a mixed space and mode representation, and the second line in this equation is the CBR expression for it. One can check now that the total numerical cost of LDOS using Eq. (4-6) can be estimated as

$$N_r = N_E N_{grids} N_{eigen} N_{modes}, \quad (4-7)$$

where N_{modes} is the number of non-zero elements in \mathbf{G}_C , which is diagonal in the mode representation (thus N_{modes} is the number of propagating modes). It is usually much more efficient to express the quantities with index C (contacts) in mode representation, due to the possible mode reduction. The advantage of using Eq. (4-6) is the absence of quadratic and higher order terms with N_{grids} or N_{eigen} .

4.1 Bound States Treatment

It is important to note that the density matrix x_{ab} in Eq. (4-4) and the derived quantities may also account for bound states, if they are present in the system. Indeed, as it has been shown in Ref. [??], the

term $X_{ab}^{(l)}(E)$ does not disappear when the coupling to the leads (represented by S_c and G_c terms) is zero (i.e. when system states are not coupled to the outside world), but instead results in

$$X_{ab}^{(l)}(E) = \frac{3}{4} \frac{3}{4} \frac{3}{4} \frac{3}{4} \frac{3}{4} \frac{3}{4} \frac{3}{4} \frac{3}{4} \frac{3}{4} d_{ab} d(E - E_a) \quad (4-8)$$

that assures the inclusion of bound states into the total charge density. We point out, however, that in the case of numerical evaluation of Eq.(4-8), the delta-functions corresponding to the bound states should be integrated analytically, leading to the expression

$$x_{ab}^{TOTAL} = \sum_{g \in BS} d_{ab} d_{ag} f(e_g) + \sum_{l=1}^L \int dE X_{ab}^{(l)}(E) f_l(E) \quad (4-9)$$

where the sum with index g is performed over all bound states (BS) in the system. While in the *idealized* ballistic case, it is generally unclear how these states are occupied if the bias is applied, however, in a presence of small scattering in the system these quasi-bound states (QBS) can be viewed as states that get occupied as a result of scattering of carriers coming from one of the leads $l = 1..L$. In the later case, if one knew ‘*from what lead has a carrier come from*’, one could assign to the carrier the corresponding distribution function. Exploring this idea, one can make an assumption that the distribution function $f(e_g)$ of the quasi-bound state $|g\rangle$ depends on the “coupling strength” to the outside leads. If a quasi bound state $|g\rangle$ is coupled more strongly to lead l , then it is reasonable to expect that its distribution function is close to the one of lead l . Generally, one can speculate that if the scattering is small, then the quasi-bound states can be occupied according to the following approximate formula

$$x_{ab}^{BS} = \sum_{g \in BS} d_{ab} d_{ag} \sum_{l=1}^L F_{gl} f_l(e_g) / \sum_{l=1}^L F_{gl} \quad (4-10)$$

where the coupling strength, F_{gl} , of state g to lead l is given by

$$F_{gl} = \sum_{m=1}^{M_l} \left| \langle g | c_m^{(l)} \rangle \right|^2 \quad (4-11)$$

The summation in Eq. (4-11) is performed as the squares of the absolute values of projections of states $|g\rangle$ over M_l transverse modes $c_m^{(l)}$ in lead l . F_{gl} can be used to determine what states $|g\rangle$ should be treated as “quasi-bound” ones. We find this approach to be essential, in particular, for a superior convergence of the self-consistent cycle. An example of using the coupling strength for determining the quasi-bound states is given in Figure 4-2. The solid circles represent the coupling strength F_{gl} of an eigenstate $|g\rangle$ to the lead l (for simplicity data for only one (source) lead are shown on Figure 4-2). We see that the vast majority of eigenstates are strongly coupled to the lead, except the lower 6 circles, for which $F_{gl} < 0.2$. It is possible, therefore, to introduce a threshold in coupling strength (for example $F_{th} = 0.19$), so that eigenstates with coupling strength less than the threshold would be identified as QBS. Furthermore, every peak in the DOS corresponds to a certain QBS (there are 6 peaks and 6 QBS shown in Figure 4-2). While the former property is not always the case (some QBS do not result in resonant peaks in the DOS), it is generally possible to find a QBS “responsible” for every resonant peak in the DOS. Therefore, most of the hard-to-integrate resonant peaks in the DOS can be eliminated, by excluding the responsible weakly-coupled eigenstates from the eigenstate set $\{|a\rangle\}$, which we use to calculate the retarded Green’s function G^0 of a closed device. These excluded states then are taken into account with the following resulting expression for the charge density:

$$n(\mathbf{r}) = \sum_{a \in BS} \left| \langle \mathbf{r} | a \rangle \right|^2 x_{aa}^{BS} + \sum_{a, b \in BS} \langle \mathbf{r} | a \rangle \langle b | \mathbf{r} \rangle x_{ab} \quad (4-12)$$

If the explicit relation between the charge density and the LDOS is desired, the following formula can be use instead:

$$n(\mathbf{r}) = \sum_{a \in BS} |\langle \mathbf{r} | a \rangle|^2 x_{aa}^{BS} + \sum_{l=1}^L \int_0^{\infty} r_l(\mathbf{r}, E) f_l(E) dE \quad (4-13)$$

where the LDOS due to lead l is $r_l(\mathbf{r}, E)$.

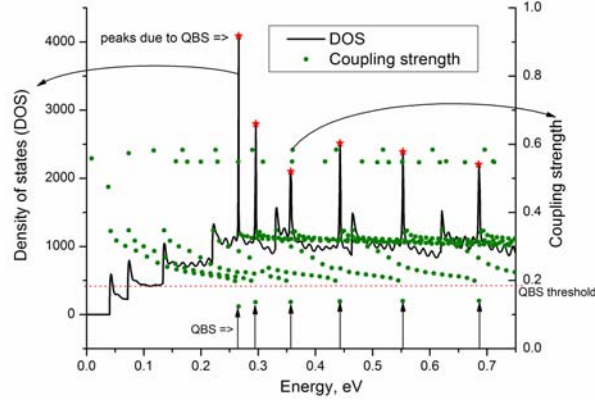


Figure 4-2 Quasi-bound state (QBS) detection using the coupling strength in Eq. (52). The graph shows DOS energy dependence (solid curve, left-hand scale) and the coupling strength for the device eigenstates (solid circles, right-hand scale). Note the resonant peak of open-system DOS at each QBS.

4.2 Energy Discretization

For an efficient numerical implementation of a self-consistent scheme, the choice of the energy grid is of high importance. To integrate the continuous part of the carrier density, the LDOS is discretized in energy space and then a simple numerical integration is done by summing up the values for each energy step weighted by the Fermi distribution and the energy grid spacing DE_k with k being the index of the energy grid. Using a regular grid with constant grid spacing, the integral over the peak deriving from the resonant states is very poor since the relative distance between the nearest energy grid point E_k and the resonant energy E_m is, generally, arbitrary. In addition, the resonant energy is slightly shifted with each iteration step, leading to a varying integration error during the self-consistent cycle, which acts as an obstacle against convergence for any self-consistent algorithm. Thus, a solution to this problem is to use the physical information about the system and employ an *adaptive energy grid* that resolves each known peak with a local energy grid of a few tens of grid points that is fixed to the resonant energy E_m . The location of resonant states is easy to find, since the resonant energies are close to (selected) eigenstates of the closed system. Another advantage of the CBR method is that these eigenstates are already known, since in this method the solution for the open system is being expressed in the basis of the closed system. As a result, the integration error is reduced compared to the case of using regular grid and remains constant within the iteration, since the grid is locally fixed to the shifted mode energies.

4.3 Self-Consistent Solution

The self-consistent solution of the ballistic or quasi-ballistic transport properties of an open device requires repeated solution of the Schrödinger and Poisson equations. In principle, it is possible to simply iterate the solution of the Schrödinger and Poisson equations and with enough damping this will yield a converged result. However, this approach leads to hundreds of iteration steps for each bias point that do not pose a reasonable scheme. To improve the convergence of a highly non-linear set of coupled equations, such as the Schrödinger-Poisson problem, the Newton algorithm is usually the first choice. However, the exact Jacobian for the Schrödinger-Poisson set cannot be derived analytically, and its numerical evaluation is rather costly (while certainly possible, see e.g. Ref. [??]). In the case of a closed system this problem has been solved using the *predictor-corrector* approach^{130,131}. The aim of this method is to find a good

approximation for the quantum density as a function of the electrostatic potential where an expression for the Jacobian is known. In this work we adopted this approach to open systems. At first, the Schrödinger equation is solved for the closed system with the Hartree potential, $j_H(\mathbf{r})$, and the exchange and correlation potential, $j_{XC}(\mathbf{r})$ taken into account. Then the local density of states $r(\mathbf{r}, E)$ of the open system is calculated using the CBR method. The Hartree potential j_H and carrier density n are then used to calculate the residuum, F , of the Poisson equation using,

$$F[j_H] = \mathbf{A}j_H - (n - N_D) \quad (4-14)$$

where \mathbf{A} is the matrix derived from the discretization of the Poisson equation. If the residuum is smaller than a predetermined threshold the solution is taken to be a converged one. If the residuum is still too large, the correction to the Hartree potential $Dj_H(\mathbf{r})$ is calculated in the predictor step, where the predictor carrier density $n_{pr}(\mathbf{r})$ is calculated, assuming it to be the functional of the change $Dj_H(\mathbf{r})$ in the Hartree potential as follows :

$$\begin{cases} n_{pr}(\mathbf{r}) = 2 \sum_{l=1}^L \int_0^{\infty} r_l(\mathbf{r}, E) f\left(\frac{E + Dj_H(\mathbf{r}) - E_F^{(l)}}{k_B T}\right) dE \\ \mathbf{A}(j_H(\mathbf{r}) + Dj_H(\mathbf{r})) = n_{pr}(\mathbf{r}) - N_D(\mathbf{r}) \end{cases} \quad (4-15)$$

where $f(x) = [1 + \exp(x)]^{-1}$ for 3D systems or the corresponding Fermi integral for systems with lower dimensions, the energy $E_F^{(l)}$ is the Fermi energy level in lead l , and a factor of 2 is taken into account for the spin degeneracy of the electrons. Note that the Jacobian for the system Eq. (4-14) can be easily found analytically:

$$\begin{aligned} J_{\mathbf{r}\mathbf{r}'} &= \frac{\mathcal{J}[F(\mathbf{r})]}{\mathcal{J}[Dj_H(\mathbf{r}')] } = \mathbf{A}_{\mathbf{r}\mathbf{r}'} + \frac{\mathcal{J}[n_{pr}(\mathbf{r})]}{\mathcal{J}[Dj_H(\mathbf{r}')] } = \\ &= \mathbf{A}_{\mathbf{r}\mathbf{r}'} + d_{\mathbf{r}\mathbf{r}'} \frac{2}{k_B T} \sum_{l=1}^L \int_0^{\infty} r_l(f-1) f dE. \end{aligned} \quad (4-16)$$

After applying the Newton method, the obtained correction to the Hartree potential Dj_H and the corresponding carrier density are used to update the Hartree j_H , exchange m_X^{LDA} and correlation m_C^{LDA} potentials for the next iteration ($i + 1$) as follows:

$$\begin{aligned} j_H^{(i+1)} &= j_H^{(i)} + Dj_H^{(i)}, \\ j_{XC}^{(i+1)} &= m_X^{LDA}[n_{pr}^{(i)}[Dj_H^{(i)}]] + m_C^{LDA}[n_{pr}^{(i)}[Dj_H^{(i)}]] \end{aligned} \quad (4-17)$$

The loop is repeated until convergence is achieved, that is $|Dj_H^{(i)}| < e$, with e being the absolute error of the potential. We find that typically only very few (5-7) solutions of the Schrödinger equation are necessary to yield a solution with 3 converged digits in the potential and currents.

4.4 Device Hamiltonian, Algorithm and Some Numerical Details

In this work FinFET devices with varying fin width (4 nm ~ 12 nm) have been simulated. With 12 nm fin width the simulation real space domain is fairly large. While the CBR method for quantum transport simulation can be used with any multi-band Hamiltonians, including the tight-binding and $k \times p$, in this work, we choose to adopt the effective mass model and finite difference discretization scheme to be able to simulate relatively 'large' FinFET device within a reasonable time frame. The structure and the size of the corresponding effective mass Hamiltonian are determined by the dimensionality of the transport problem and the number of real space grid points. Due to the presence of non-equivalent valleys in Si, we need to solve the open-system problem for each valley, and then add up the contributions from different valleys (weighting them with the corresponding valley degeneracy).

In ultra-scaled nano-transistors source, drain and gate regions are usually heavily doped, therefore it is important to include quantum-mechanical effects of exchange and correlation. In this work this is done via the local density approximation (LDA). The phenomenological scattering on the phonons using the relaxation time approximation has been taken into account. Since this phonon scattering model relies on

phenomenological parameters, in this work we present results that include into account this phenomenological scattering on phonons as well as purely ballistic ones (that do not depend on such parameters).

After the initial guess for the potential and the *initial* number of device eigenstates, the CBR loop is started. For each CBR-Poisson iteration the following tasks are performed: i) transverse lead modes are calculated; ii) eigen-problem is solved for closed- system with von Neumann boundary conditions at the contacts; iii) open-system solution is constructed. The simulator has been modified to incorporate the automatic determination of the required number of device eigenstates and lead modes for each iteration to yield desired accuracy. Due to this dynamic nature of eigenstate and lead modes determination, CPU time can be saved and also memory requirements have been optimized.

PARAMETER\OPERATION REGIME	SUBTHRES HOLD	ON-STATE
Number of grid points/mesh size	17169/2.5 Å	17169/2.5 Å
Number of device eigenstates used in calculation (averaged over valleys)	470 (2.7%)	270 (1.5%)
Number of total lead transverse modes used in calculation (averaged over valleys)	39 (20 %)	31 (16 %)
Average absolute error of potential (eV)/ Average number of converged digits of the current	$10^{-5} / 3$	$10^{-5} / 3$
Average number of CBR-Poisson iterations	5	6

Table 4-1 Convergence data and average number of generalized von Neumann eigenstates used for construction of open system solution.

The accuracy e also determines the upper error norm for the functional F ; if $\|F\| < e$ then the solution is considered to be converged and the next bias point can be processed, otherwise the predictor-corrector approach is invoked to determine correction Dj to the potential. With updated potential j CBR routine is called again and the loop continues until convergence is achieved. Note that the CBR module is called for each non-equivalent Si valley to obtain the LDOS and transmission function for each valley; then the total charge density, currents, etc. are calculated as the corresponding sums. Table 4-1 shows the average values of required number of device eigenstates and lead modes in off- and on- state of a FinFET device being simulated.

One can see that the CBR method allows us to use a small fraction of device eigenstates and lead modes to get a well-converged solution within 5-6 iterations, on average. It is significant that this excellent rate of convergence has been observed on a wide variety of devices with different doping profiles and geometries. However, in order to achieve this result, a combination of *all* the steps has to be performed. For example, in the absence of QBS detection, the average number of iterations would be about 20-30, and in some cases there could be no converged solution at all. Similarly, it would be significantly harder to achieve any convergence in the absence of adaptive energy discretization, etc. However, we find that the full scheme presented in this section, resolves convergence problems in most cases. As a real-life example, the convergence of the non-linear Poisson equation for a FinFET with different gate voltages changing from +0.2 to -1.0 V and fixed drain-to-source voltage (0.1 V) is shown in Figure 4-3. The corresponding error in the source-drain current is also plotted in the same figure. The maximum error in the source-drain current values is 2% for the potential accuracy fixed at 2×10^{-5} eV. No convergence-tuning parameters of any kind have been used in the simulation: the energy grid, energy cut-off, number of eigenstates, lead modes, etc., are automatically determined by the CBR simulator in every iteration and for each bias point.

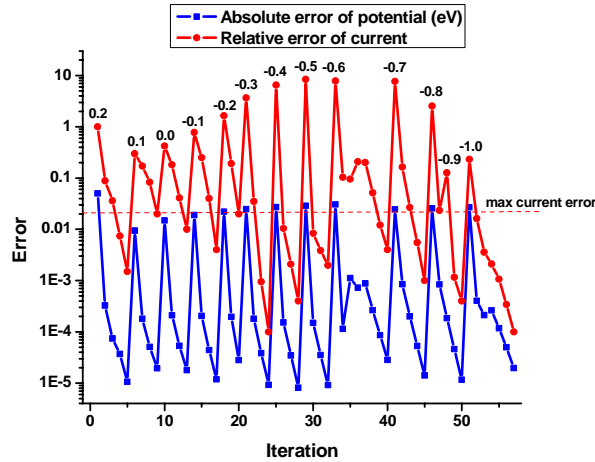


Figure 4-3 Residuum of non-linear Poisson equation and corresponding relative error in drain current (w.r.t. converged solution of a higher order) for different gate voltages. The gate voltage values are shown above of each segment of the curve.

4.5 Simulation Example – 2D Results

Over the decade many novel structures have been proposed for the nanoscale regime of operation, among them fully depleted MOSFETs, in particular Double-Gate (DG) MOSFETs emerged as the leading candidate for the ultimate scaling of silicon MOSFETs down to 10 nm (see Table (??)). In these devices effective control of the gate over the channel has been enhanced by using multiple gates and thinning of body thickness¹³². For a given insulator thickness theoretical study shows that DG devices can be scaled to the lowest channel length keeping the short channel effects within acceptable limits¹³³. Theoretically, cylindrical or surround-gate MOSFET is found to show the best gate control of channel but realization of this structure from fabrication point of view is quite challenging^{134, 135}. Different orientation of double-gate MOSFETs have been proposed¹³⁶ as shown in Figure 4-4. In type I device¹³⁷ the current direction is in plane but gate-to-gate direction is normal to the wafer plane. The fabrication process with this type of devices is complex and contacting the bottom gate is rather difficult. Type II devices¹³⁴ have gate-to-gate direction in plane but the current direction is perpendicular to the plane. This type of devices suffers from inability to easily control the channel and source/drain doping profiles¹³⁸. Type III devices¹³⁹ have the advantage of both in-plane gate-to-gate direction and in-plane current direction but the width of the device is normal to the plane.

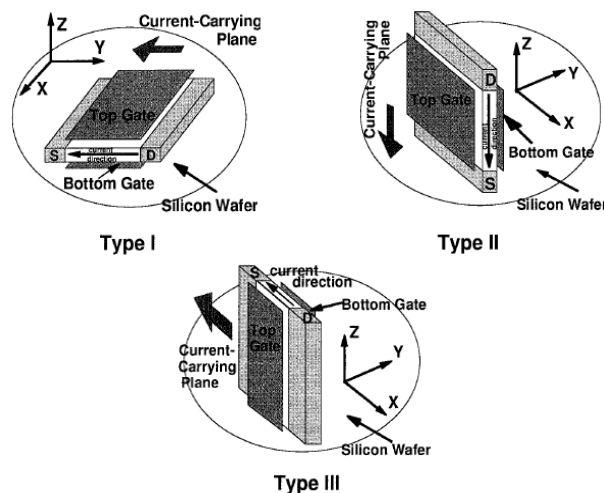


Figure 4-4 Three possible orientations of DG MOSFETs in silicon wafer (adapted from Wong *et. al.* [136]).

The major disadvantages of these double gate MOSFETs are: (i) non-planer structure as opposed to the planar structure of conventional bulk MOSFETs, (ii) self-alignment of the gates with each other and with source/drain and, (iii) formation of ultra thin silicon film. FinFET^{140, 141, 142} is a special category of type III devices in which the height is reduced to maintain quasi-planar topography for the ease of fabrication¹⁴³. In FinFETs gates are automatically self-aligned with each other¹⁴⁰ and also the packing density is large compared to other double-gate structures¹³⁶.

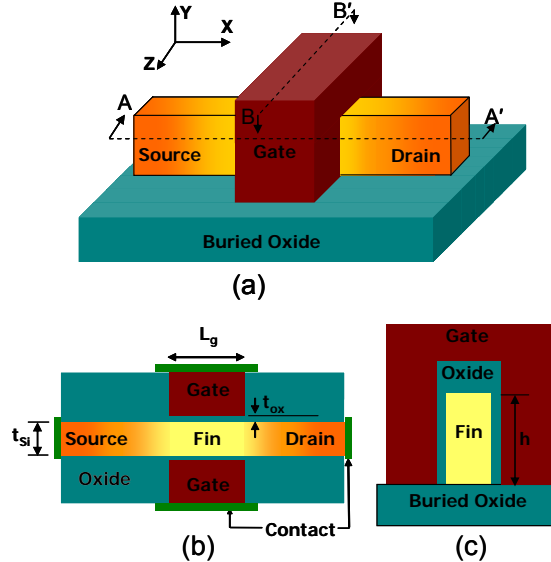


Figure 4-5. (a) - 3D schematic view of a prototype FinFET, (b) - top view along A-A' cross section and, (c) - side view along B-B' cross-section.

The geometry of a typical FinFET device is shown in Figure 4-5. The fin thickness, t_{si} , is considered to be the most important process parameter as it controls the carrier mobility as well as threshold voltage. The fin is made thin enough when viewed from above, as shown in Figure 26 (b), so that both gates simultaneously control the entire fully depleted channel film. Usually the top surface of the fin is covered by a thicker oxide compared to the thickness of the side gates (front and back), t_{ox} ; therefore channels form only along the vertical surfaces of the fin. The fin height, h here is equivalent to the “gate width” of the conventional bulk MOSFET. Therefore, the effective channel width in FinFET devices is equal to $2h$ when only side gates are considered. For higher drive current different channel width is achieved by introducing multiple fins in parallel. In that case, the resultant width of the channel can be represented as $2 \cdot h \cdot N_{fins}$ with N_{fins} being the number of fins.

In this work we have modified our 2D CBR simulator in such a way that semiconductor devices on wafers of *arbitrary crystallographic orientation* can be simulated. This was necessary to match the experimental data¹⁴⁴ for a FinFET device of which the channel is on (110) wafer plane. The conventional approach assumes wafer in (001) plane, and with the real space axes X, Y, Z being along crystallographic directions [100] and [010] and [001] respectively, the effective mass tensor is diagonal and the Schrödinger equation can be discretized and solved accordingly. However, for FinFET devices with channel oriented in (110) wafer plane, the effective mass tensor is non-diagonal (see e.g. [??]). The resulting Schrödinger equation has mixed second derivative and first derivatives terms of which the coefficients are the non-diagonal element of effective mass tensor. Considering 2D simulation, it is possible to eliminate the mixed second derivative term by rotating the device in real space by a suitable angle. The first derivative term can be eliminated with the wave-function change of variable after the elimination of second derivative terms. As a result, to simulate FinFET devices with channel orientation in (110) wafer plane it is sufficient to use modified effective masses along device coordinates. Note that the above procedures are rigorously valid for 2D (and 1D) transport simulations; a full 3D simulation with the wave-function depending on the device depth (e.g. fin height) would require a somewhat different treatment of the coefficients in the discretized Schrödinger equation containing effective masses. Regarding 2D simulation, however, we assume that the

wave-functions depend on device length and width directions, but neglect the explicit height dependence, thus assuming that the *transport* in this 3D FinFET device is two-dimensional (2D).

With the inclusion of the modifications specified above, sets of 2D simulation have been performed in order to match experimental data with fin thickness of 12 nm and physical gate oxide thickness of 1.7 nm. In the experiment the gate electrode consisted of dual doped n^+ / p^+ polysilicon. Also the gate insulator is nitrided oxide for which the dielectric constant might not be exactly the same as that of SiO_2 ¹⁴⁵. However, in our simulations we use the same device geometry (fin width, gate length and gate oxide thickness) but assume n^+ polysilicon gate and SiO_2 as the gate insulator. The effects of top gate on transport are assumed to be negligible considering much thicker gate oxide compared to side gate oxide. As mentioned earlier, the experimental FinFET device has been fabricated with the channel oriented in (110) wafer plane. In our simulations we also adopt the same wafer plane and assume that carrier propagation is along $[1\bar{1}0]$ crystallographic direction.

In order to obtain the closest match to the experimental results, a series of simulations with different combinations of doping profiles (source/drain doping concentration) and gate-source/drain underlap regions (which defines the doping gradient) have been performed. The doping profile which gives the closest fit of simulation results to the transfer characteristics of experimental FinFET at low drain bias can be described as - source/drain doping of $7 \times 10^{18} \text{ cm}^{-3}$ which follow a Gaussian envelope over a gate-source/drain underlap length of 12 nm to reach the body doping of 10^{15} cm^{-3} . The resulting doping gradient is around 3 nm/dec. We use uniform doping of $7 \times 10^{18} \text{ cm}^{-3}$ in the gate electrodes. Since the exact doping profile in the gate electrode is not specified, the simulated transfer characteristics can be shifted in voltage-scale (gate voltage) to match experimental data. However, it is important to mention that in selecting the above mentioned doping profiles as the appropriate one to match experimental data we consider simultaneously that: (i) the value of subthreshold slope being in good correspondence to that obtained in experiment, (ii) over the gate voltage range of interest (-0.8V to 0V) the current values are reasonably close to the experimental data and, (iii) at very low gate voltage transfer characteristics do not show any bending which we do not observe in experimental data.

As one can see from Figure 4-6, the transfer characteristics obtained using the above mentioned doping profile gives current values close to the experimental ones in the subthreshold regime at a drain voltage of 0.1 V. In order to check that this result is not a coincidence, the transfer characteristics with the same geometry but with a high drain voltage of 1.2 V have been calculated, and found to be in good correspondence with the experimental data (see Figure 4-10).

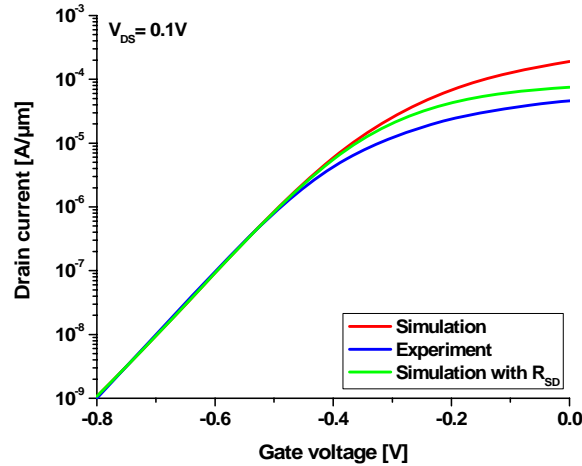


Figure 4-6 Comparison of simulated transfer characteristics to the experimental data at low drain bias of 0.1V.

One can see from Figure 4-6 that above threshold, with the increase in gate voltage, the deviation between simulated and experimental data increases rapidly. We predict that presence of very high parasitic series source/drain resistance, a critical issue in FinFET device, might be a reason for the smaller value of the drain current at high gate bias in the experiment. In order to examine the influence of the series parasitic source/drain resistance, R_{SD} we extract the value of R_{SD} of experimental device from a plot of total

resistance, R_{tot} (sum of device resistance, R_{int} and series parasitic source/drain resistance, R_{SD}) vs. gate voltage as shown in Figure 4-7.

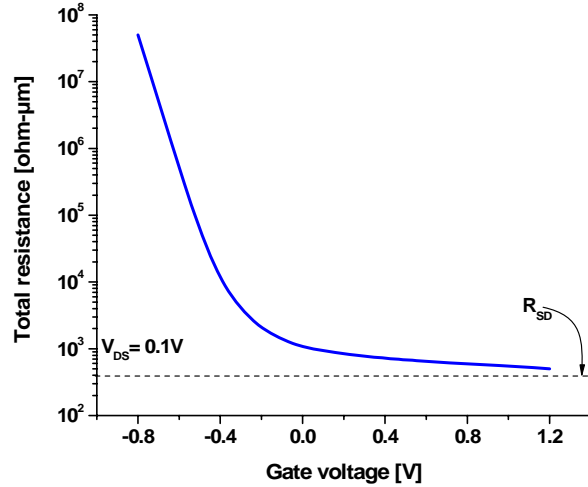


Figure 4-7 Total resistance, R_{tot} as a function of gate voltage at low drain bias.

For sufficiently large value of gate voltage, R_{int} becomes very small and one can reasonably assume that $R_{tot} \gg R_{SD}$. The value of total parasitic series source/drain resistance extracted for the experimental device is found to be around $400 \Omega - \mu\text{m}$. Including the effects of R_{SD} the modified transfer characteristics is also shown in Figure 4-6 and one can see that the simulation result is very close to the experimental findings even at high gate bias. After including the effects of series resistance still we see some deviation of simulation results from the experiment in on-state. It is well known that in nanoscale devices, the presence of an unintentional dopant in the channel is highly probable¹⁴⁶. Even if the fin is lightly doped, the unavoidable background doping might give rise to a one ionized dopant being present at a random location within the channel. Also, if an electron becomes trapped in a defect state at the interface or in the silicon body, it will introduce a fixed charge in the channel region. Depending on its position and applied bias, this unintentional dopant can significantly alter the device behavior, particularly when the channel is very lightly doped. An unintentional dopant sitting at a random location within the channel introduces a localized barrier which impedes the carrier propagation. The impact is significantly larger for an unintentional dopant sitting at the beginning of the fin near the source end compared to other probable positions¹⁴⁷.

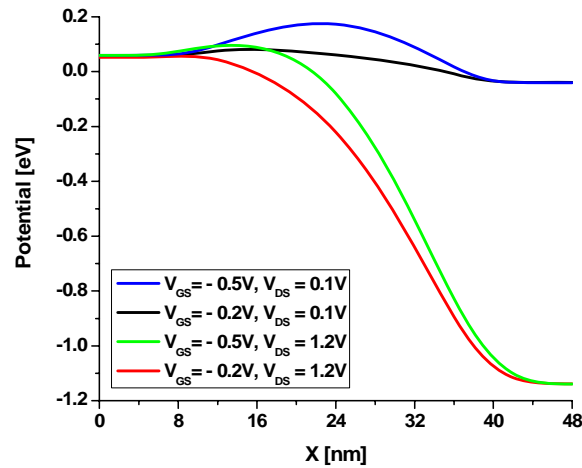


Figure 4-8 1D potential profiles along the length of the device in subthreshold and on-state with low and high drain biases.

Figure 4-8 depicts the effective 1D potential profiles along X direction at the center of the fin in subthreshold regime and on-state at low and high drain bias. Also shown in Figure 4-9 is the corresponding 1D lateral electric field profiles along X direction. At low drain bias ($V_{DS} = 0.1V$) and low gate voltage ($V_{GS} = -0.5V$) the intrinsic barrier is already high enough (as shown in Figure 4-8) so that the effects of the localized barrier introduced by the unintentional dopant can be assumed negligible. Therefore, over the subthreshold regime, we see a very good correspondence between simulation and experiment. For higher gate voltages ($V_{GS} = -0.2V$), the intrinsic barrier is reduced significantly (Figure 4-8).

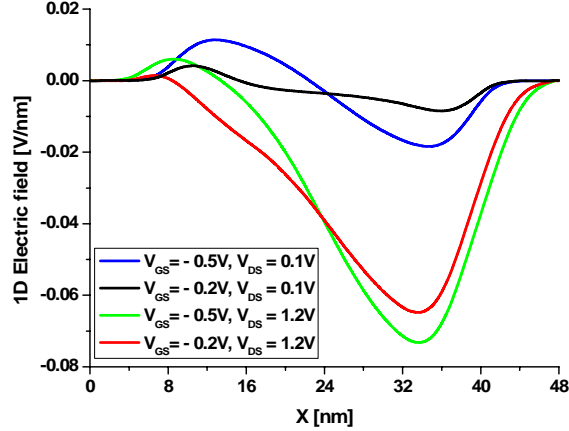


Figure 4-9 1D electric field along the length of the device for the potential profiles in Figure 4-8.

Also the lateral electric field is reduced due to the increased effects of transverse electric field (Figure 4-9). Thus, the localized barrier due to unintentional dopant is expected to influence the value of the drain current around device turn-on point. Therefore, at low drain bias ($V_{DS} = 0.1V$), the deviation between simulated drain current and experimental value increases with increasing gate voltage (above threshold) up to some cut-off beyond which the inversion electrons start to screen the potential of a single dopant ion. Consequently the influence of unintentional dopant on drain current gradually diminishes at much higher value of the gate voltage beyond threshold voltage which is also evident from Figure 4-6.

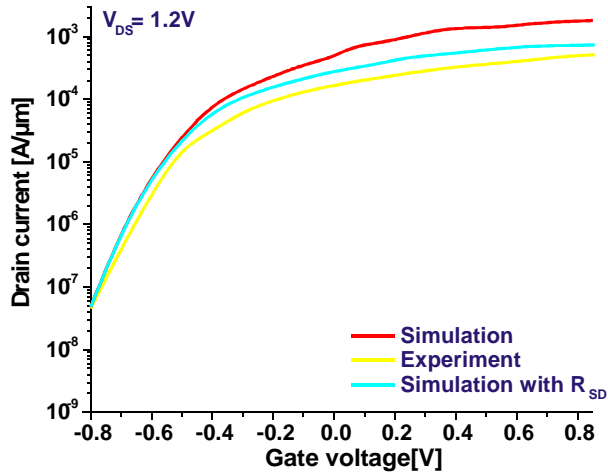


Figure 4-10 Comparison of simulated transfer characteristics to the experimental data at $V_{DS} = 1.2V$.

For higher value of drain voltage ($V_{DS} = 1.2V$), we see some discrepancies between the simulation results and experimental data in *both* subthreshold and at high gate volt-ages as shown in Figure 4-10. Inclusion of series parasitic source/drain resistance reduces the drain current, but still the experimental values of drain current remains much smaller than the simulation results. In this case, due to significant

DIBL effects, intrinsic barrier reduces, compared to the case with low drain bias, for both low and high gate voltages as shown in Figure 4-8. In subthreshold regime, the intrinsic barrier is much lower for $V_{DS} = 1.2V$ than for $V_{DS} = 0.1V$. Consequently, the discrepancy between the experiment and simulation can be explained by more ‘noticeable’ (with respect to intrinsic barrier) effect of localized barrier due to *unintentional dopant*, which was much less significant for low drain bias. As the gate voltage increases, the effects of unintentional dopant become even more pronounced, which may explain the high voltage trend in Figure 4-10. We note that the position of the unintentional dopant is crucial in determining its effects on drain current. At high drain bias, unintentional dopant at the source side, will affect the drain current stronger than impurities at other locations.

Finally, we note that the subthreshold slope of 125 mV/dec have been reported for the *n*-FinFET in the experiment¹⁴⁴. The corresponding value as obtained from our simulation is 120 mV/dec. The value of DIBL(at $I_D = 3 \times 10^{-6} A/\mu m$) as extracted from the transfer characteristics of the experimental device is 145 mV/V and the corresponding value calculated from our simulation considering the effects of series parasitic source/drain resistance is 160 mV/V. These numbers clearly show that the experimentally fabricated¹⁴⁴ 10 nm FinFET device was very far from optimal. Consequently, the 10 nm device characteristics could be significantly improved by a proper tuning of device geometry. In the work¹⁴⁸ we have used our CBR simulator to optimize the device geometry and doping profile of a 10 nm FinFET device to meet most of the performance matrices defined by ITRS [5] for high performance 10 nm double-gate devices, which are expected to be commercially available around 2015.

4.6 Simulation Example – 3D Results

In the previous section simulation results obtained for DG FinFETs using 2D CBR simulator have been presented. In using 2D simulator the implicit assumptions taken into account are, (i) height (‘Y’ coordinate) of the FinFET is much larger compared to the width (‘Z’ coordinate) (Fig. 4-5(a)) so that it is reasonable to assume that carriers are not confined along Y direction and solution of Schrödinger equation can be represented by simple plane waves and, (ii) the device can be viewed as a combination of parallel identical slices along the Y direction. However, the latter assumption becomes invalid once the top gate in a FinFET and the buried oxide at the bottom of the device are taken into account. In addition, the first assumption is not always necessarily true for ultra-scaled devices as the Y directed length can be comparable to Z directed length. For example, considering 10 nm FinFET and ‘area efficiency’ as the factor for the determination of fin height, the minimum height could be as low as 10 nm. In this case even the dimension of ultra thin fin ($\gg 4$ nm) is comparable to the fin height ($\gg 10$ nm) and the first assumption may lead to not very accurate picture of carrier transport. Moreover, effects like discrete doping, unintentional dopant in the channel and surface roughness all are inherently three dimensional in real space and 2-D simulator should not be used to estimate the impacts of these effects in device characteristics.

One of the major goals of this research was to extend the 2-D CBR simulator to a fully self-consistent 3-D version so that device characteristics of ultra-small devices can be predicted with higher accuracy. However, the computational cost in 3D simulator goes up significantly from that of 2D simulator as it solves 3D open-system Schrödinger and 3D Poisson equations. Moreover, in 3D domain contact cross-sections are two dimensional instead of one dimensional for the 2D case. One would then, within the CBR formalism, need to solve 2D Schrödinger equation for the lead modes of all contacts and 3D Schrödinger equation for device eigenstates for each equivalent valley.

Regarding 3D simulation of DG and TG FinFETs, an ultra-scaled geometry has been considered to keep computational time within reasonable limits. The entire device domain is $30 \text{ nm} \times 12 \text{ nm} \times YL$ where YL varies from 4 nm to 8 nm. Fin widths of 4 nm and gate oxide thickness of 1.2 nm have been assumed. Buried oxide is not simulated in this work. Source/drain doping of $2 \times 10^{19} \text{ cm}^{-3}$ is used with a doping gradient of 1.25 nm/decade over the source/drain-body junctions. Uniform doping of $2 \times 10^{19} \text{ cm}^{-3}$ has been assumed for the gate electrodes. For DG FinFETs fin height of 4 nm and 8 nm have been considered while for TG simulation only fin height of 4 nm is used. In addition, a simplified TG structure (as shown in Fig. 7.1) has been used to save computational time.

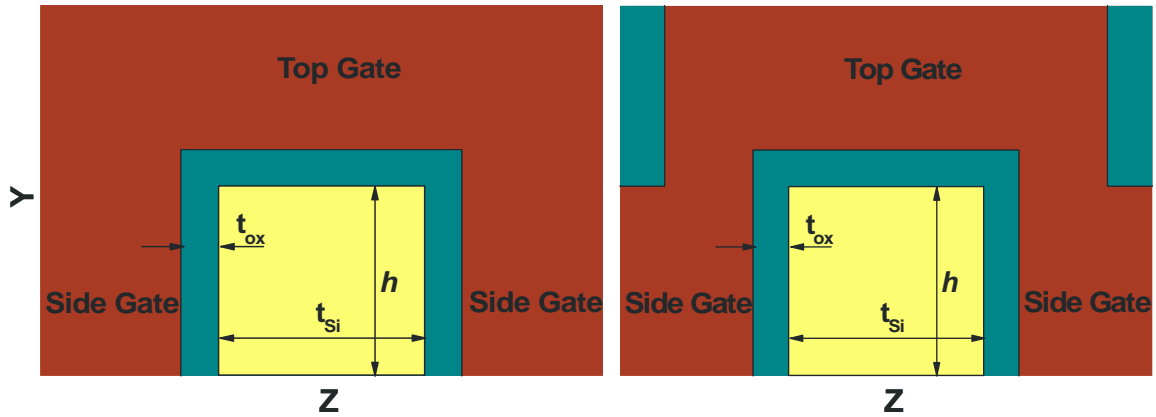


Figure 4-11 Left panel – side view of the FinFET geometry along line B-B' in Fig. 4.5(a), Right panel – side view of simplified structure used for 3-D simulations in this work

4.6.1 DG FinFET: 2D vs. 3D Simulation

At first the computational efficiency of both 2D and 3D simulator has been compared. The 2D computational domain is $30\text{ nm} \times 12\text{ nm}$ while the 3D computational domain is $30\text{ nm} \times 12\text{ nm} \times 4\text{ nm}$. Fig. 4-12 shows the convergence in terms of Poisson residuum as the gate voltage varies from -0.3V to 0.4V at a fixed drain bias of 0.4V . Each segment of the curves corresponds to a particular gate voltage and number of dots/square denotes the number of iteration needed to achieve convergence for that bias condition.

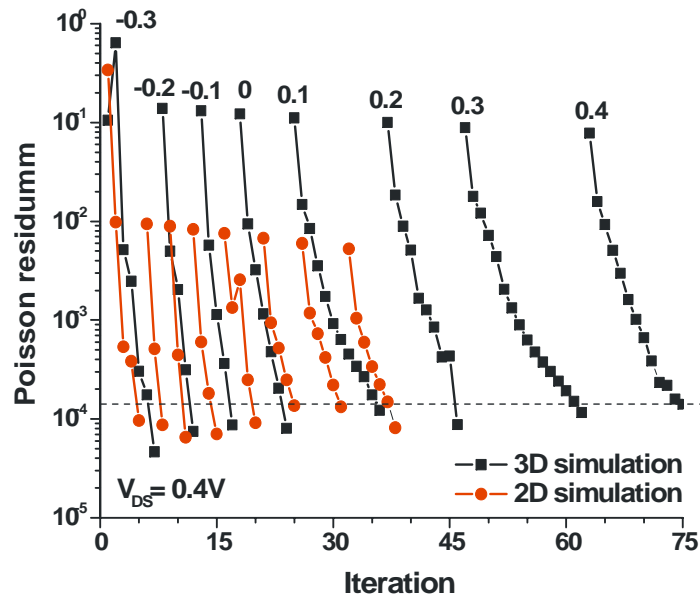


Figure 4-12 Residuum of non-linear Poisson equation for 2-D and 3-D case. The number of dots/square along each segment of the curves corresponds the number of iteration for that bias point. The value of applied voltage is shown close to each segment.

It is evident from the results presented in Fig. 4-12 that both 2D and 3D simulations converge well for each of the bias points with an average Poisson residuum smaller than 10^{-4} eV. Table 4-2 shows different parameter related to computational efficiency of 2D and 3D simulator for each iteration and valley.

Table 4-2 COMPUTATIONAL EFFICIENCY OF 2D AND 3D SIMULATOR FOR EACH BIAS POINT AND AVERAGED OVER VALLEYS

Parameter	2D simulation	3D simulation
Grid points	2356	25916
Number of device eigenstates	90 (3.8%)	260 (1%)
Avg. number of iterations per bias	5	9
Eigen-solver time (per iteration)	15s	800s
Open-system solver time(per iteration)	36s	2200s

The impact of fin height on the device characteristics of DG FinFETs has been investigated using the 3-D simulator and the results have been compared with the corresponding 2-D simulations. For 3-D simulations, fin heights of 4 nm and 8 nm have been examined while the 2-D simulation is performed considering a slice in XZ plane.

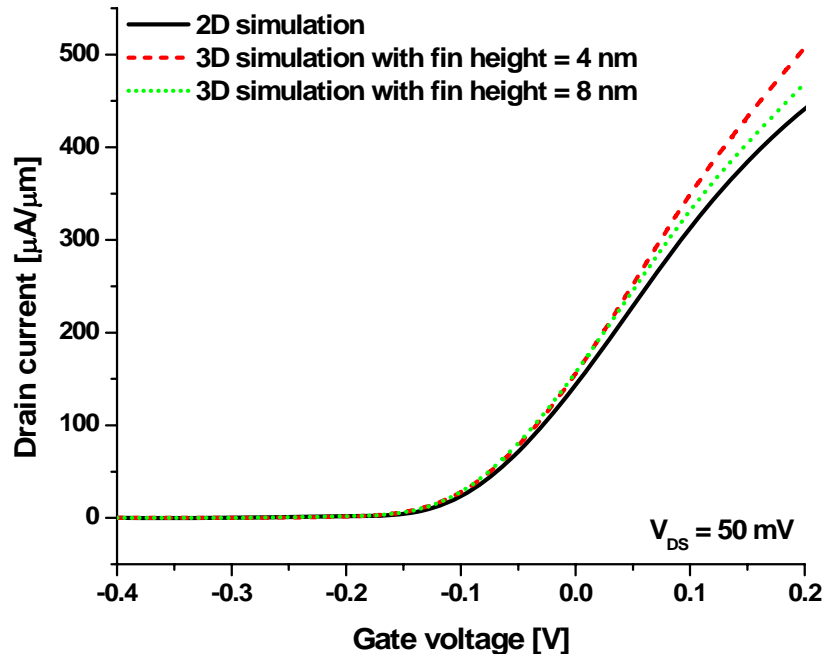


Figure 4-13 Transfer characteristics of the considered FinFET with fin heights of 4 nm and 8 nm along with the 2-D simulation result at a drain bias of 50 mV.

The corresponding transfer characteristics of the FinFET are shown in Figure 4-13 at a drain bias of 50mV. It is important to mention that simulations results presented here are purely ballistic (i.e. no scattering has been included). Note that the values of the drain current are normalized by the height of the fin, h (per fin height value). For a fin height of 4 nm the drain current is significantly higher than the 2D case at high gate voltages. As the fin height increases from 4 nm to 8 nm the drain current decreases and approaches the values obtained with the 2D simulation. This behavior can be explained as an effect of increasingly stronger quantum confinement as fin height is reduced. The electron density obtained from 2D simulation along with average electron density (taken over the fin height of 4 nm), and electron density at the middle of the fin height for a fin height of 4 nm and 8 nm are shown in Figure 4-14. Two-dimensional simulation assumes no confinement effects along the height direction. As the fin height is reduced, confinement along height direction gradually becomes stronger and the density of electrons around the center (along the height

direction) of the fin increases. As can be seen from the results presented in Figure 4-14, for a fin height of 4nm the average (taken over fin height) electron density is smaller than the 2D density but the electron density in a XZ plane at the middle of the fin height is significantly larger than the 2D case due to the strong quantum confinement. Therefore, in this case the current is mainly determined by the confined electron around the center region of the fin even though the average density is smaller than the 2D case. As the fin height is increased from 4 nm to 8 nm, the confinement along the height direction becomes relatively weaker and consequently the electron density at the middle of the fin height also reduces. Therefore, the drain current also reduces and approaches the 2-D value.

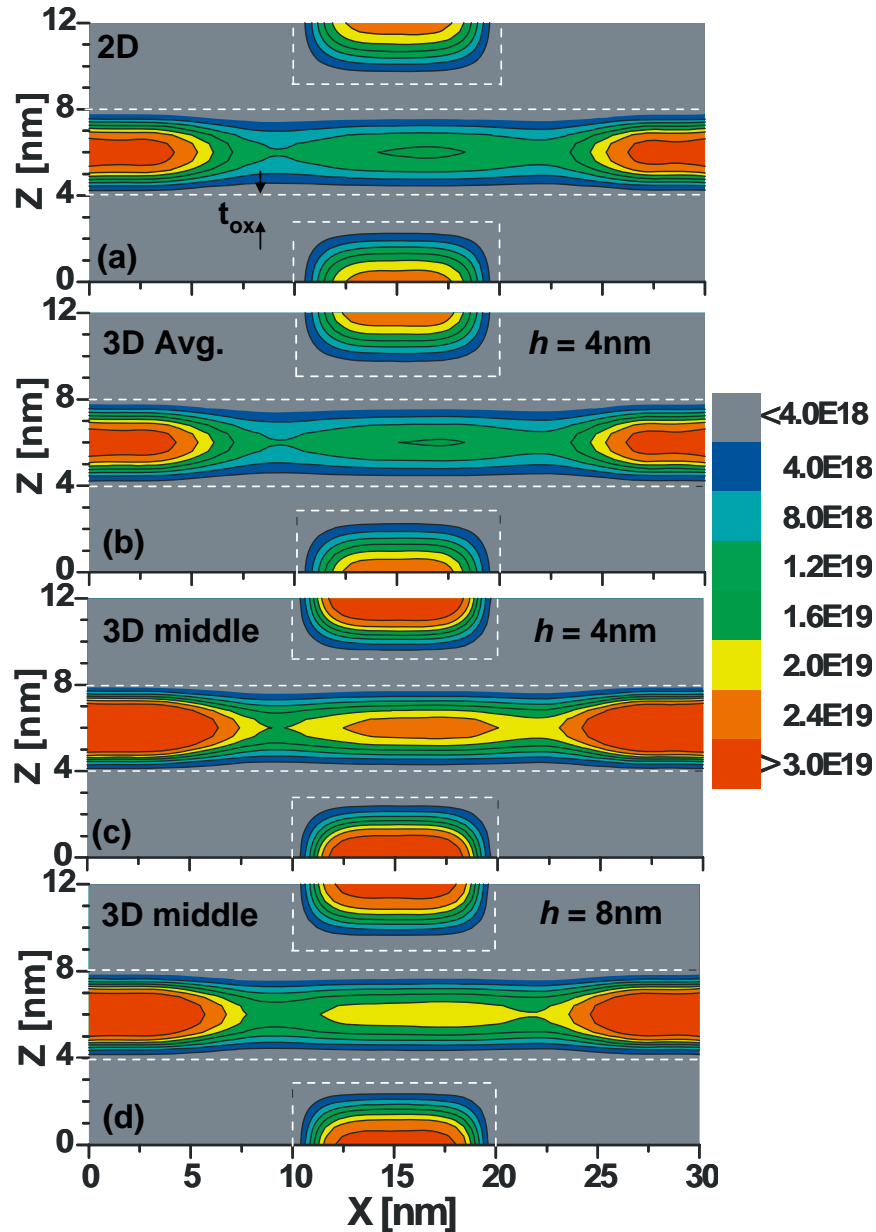


Figure 4-14 2D electron density, (a) 2D simulation, (b) averaged over h with $h = 4\text{nm}$, (c) at $y = h/2$ with $h = 4\text{ nm}$ and, (d) at $y = h/2$ with $h = 8\text{nm}$ The applied biases are $V_{GS} = 0.2\text{V}$ and $V_{DS} = 0.05\text{V}$

The net *gate leakage* normalized by the fin height is shown in Figure 4-15 for both 2D and 3D simulations. Interestingly, these results clearly show that a 2D gate leakage simulation can also serve as a very good approximation for *double-gate* devices: in this case the location of resonant peaks in gate

currents is not affected by 3D effects, only the magnitude of the peaks is affected due to the change of electron density in the channel, induced by the confinement in height direction. Predictably, this property is no longer true for *tri-gate* devices, as it is demonstrated in the following sub-section.

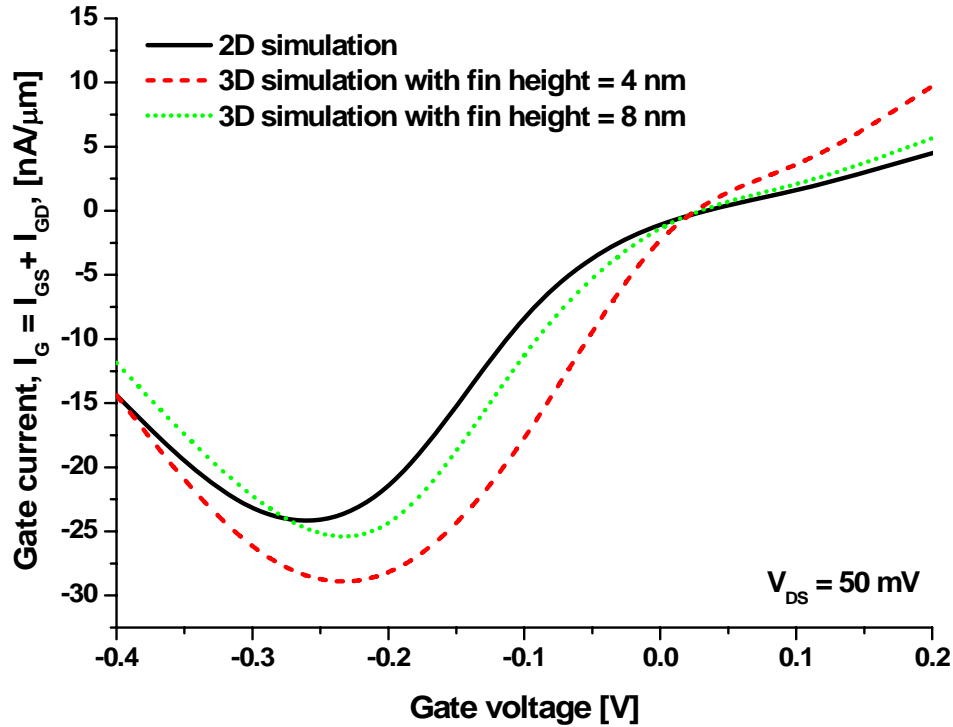


Figure 4-15 Net gate leakage as a function of gate voltage for the considered FinFET with fin height of 4 nm and 8 nm along with the corresponding 2-D simulation result.

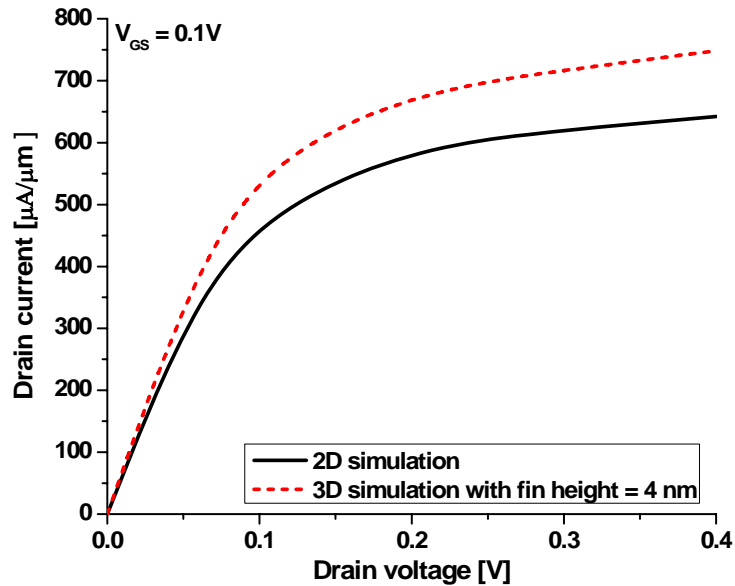


Figure 4-16 Output characteristics of the considered FinFET obtained from 3-D simulation with $h = 4$ nm along with the corresponding 2-D simulation at $V_{GS} = 0.1$ V.

Finally, we compare the 2D and 3D output characteristics of DG FinFET. The results obtained from 3D simulations for gate voltage of 0.1V are shown in Figure 4-16 along with the corresponding 2D simulation. One can see that the output characteristics exhibit behavior similar to the transfer characteristics as fin

height is reduced. The percentage increment in drain current (calculated as $(I_{3D} - I_{2D}) / I_{2D} \cdot 100$) is nearly constant for different drain voltages and is equal to 15%.

4.6.2 Double-Gate(DG) vs. Tri-Gate(TG) FinFET

In order to investigate the influence of the top gate on carrier transport, FinFET with active top gate has also been simulated. In this case the top gate oxide thickness is set to the same value as the side gate oxide thickness, 1.2 nm. Figure 4-17 shows the transfer characteristics of the DG and TG FinFETs at a drain bias of 0.4V.

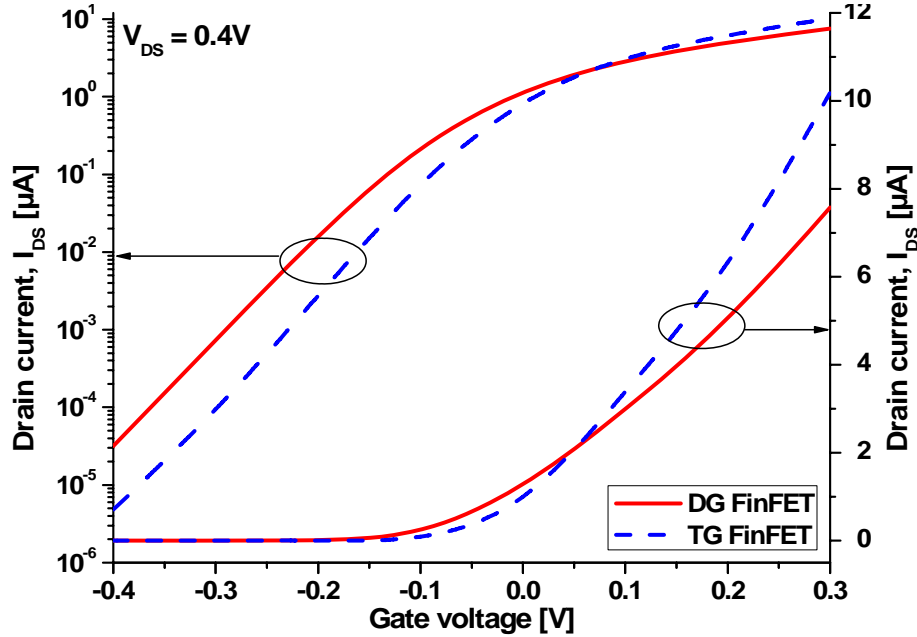


Figure 4-17 Transfer characteristics of double-gate and tri-gate FinFETs at a drain bias of 0.4V (linear and semi-log scale).

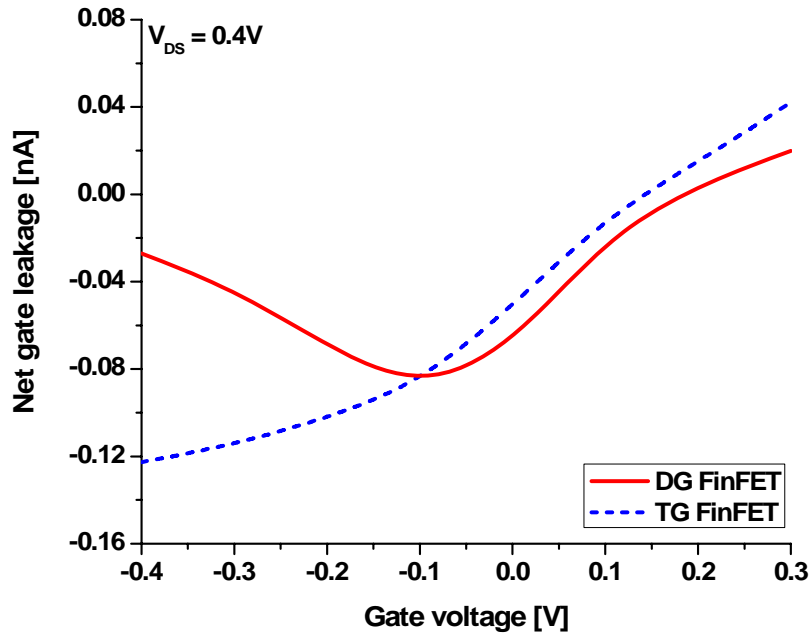


Figure 4-18 Net gate leakage vs. gate voltage for DG and TG FinFETs at $V_{DS} = 0.4V$.

From Figure 4-17 one can see significant improvement in device turn-on and turn-off characteristics which is mainly due to the additional control of the top gate over the channel. Figure 4-18 shows the net gate leakage as a function of gate voltage for a fixed drain bias of 0.4V. Adding the top gate significantly increases the off-state gate leakage as can be seen from Figure 4-18.

The 3-D electron densities for DG and TG FinFETs for an applied gate voltage of 0.2V and drain bias of 0.4V are shown in Figure 4-19.

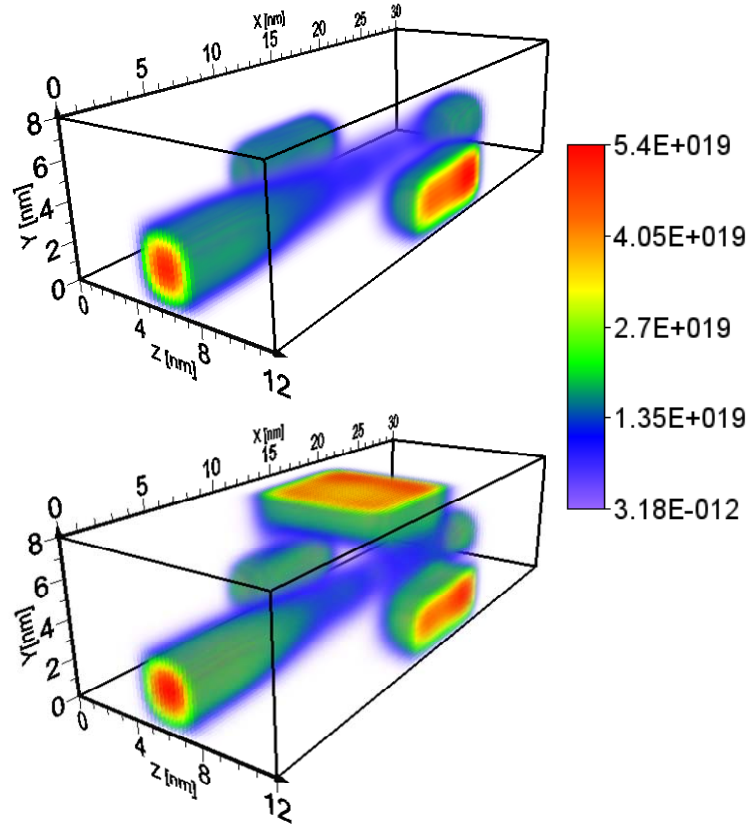


Figure 4-19 3-D electron density for DG and TG FinFETs for $V_{GS} = 0.2V$ and $V_{DS} = 0.4V$.

One can see from these simulation results that electrons are equally and strongly confined around the center of the fin from Y and Z directions since $t_{Si} = h = 4$ nm. Also shown in Figure 4-20 are the electron densities in YZ plane at $X = 15$ nm for DG and TG FinFETs. Inclusion of the top gate increases the channel electron density significantly and consequently, yield higher drain current.

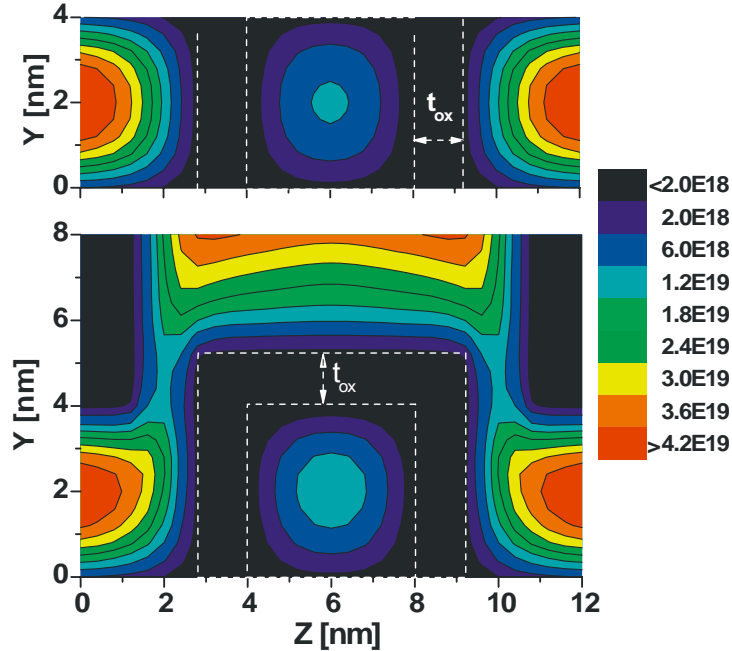


Figure 4-20 Electron densities in YZ plane at $X = 15 \text{ nm}$ at $V_{GS} = 0.2\text{V}$ and $V_{DS} = 0.4\text{V}$ for DG (top panel) and TG (bottom panel) FinFETs.

The output characteristics of DG and TG FinFETs for an applied gate voltage of 0.1V are shown in Figure 4-21. One can see the ‘flattening’ of the output characteristics for TG FinFET at smaller drain voltage compared to DG FinFET. Therefore, TG FinFET would provide less pronounced effects of channel length modulation by drain voltage.

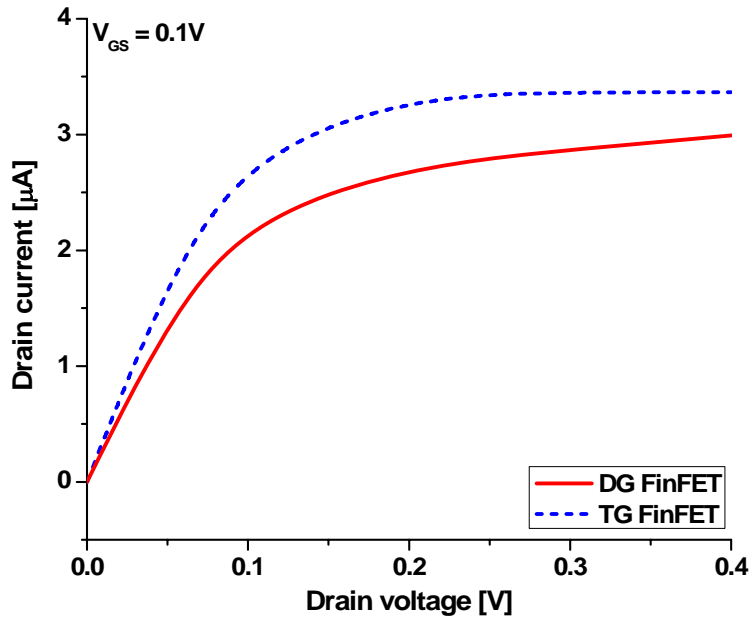


Figure 4-21 Output characteristics of double-gate and and tri-gate FinFETs at a gate voltage of 0.1V .

Table 4-3 summarizes different performance matrices obtained from the simulations for DG and TG FinFETs. For the considered FinFET, adding a top gate increases the on-current increases by 35% and decreases the subthreshold source-to-drain leakage by 85%. **However**, the total off-current

$(I_{SD,LEAK} + |I_G|)$ approximately doubles when top gate is added due to the significant increase in the off-state gate leakage as shown in Table 4-3. It is important to note that for all the simulations presented here, SiO₂ as the gate dielectric and thin gate oxide thickness of 1.2 nm are assumed. It is possible to reduce the amount of leakage by using high-K dielectric in the gate oxide. Overall, device characteristics can be improved by adding a top gate.

Table 4-3 PERFORMANCE MATRICES OBTAINED FROM THE SIMULATIONS OF DG AND TG FINFETS

Parameter	DG	TG
$I_{ON} = I_{DS} @ V_{DS} = 0.4V, V_{GS} = 0.3V$ [μA]	7.57	10.18
$I_{SD,LEAK} = I_{DS} @ V_{DS} = 0.4V, V_{GS} = -0.4V$ [nA]	0.0319	0.0048
Subthreshold swing [mV/dec]	73	70
$ I_G @ V_{DS} = 0.4V, V_{GS} = -0.4V$ [nA]	0.027	0.122

4.6.3 Effects of an Unintentional Dopant : DG vs. TG FinFET

Using 3D simulator it is possible to analyze device behavior in the presence of an unintentional dopant within the channel region. There have been a number of theoretical investigations on the evaluation of the effects due to unintentional dopants in the channel region. Effects of a point defect in nano-wire MOSFETs has been reported where the point defect is characterized by a macroscopic Coulomb tail treated in the mode-space approach, plus a short range on-site perturbation potential energy, treated exactly. It is found in that work that subthreshold current may vary in a factor of 10 according to the position of the impurity. Very recently, a full 3-D real-space RGF simulator has been presented for the study of the impacts of stray charges on the behavior of ultra-small nano-wire transistor. It was concluded that the influence of a single dopant on device behavior is dramatic, *independent* of its spatial location. This rather unusual result is likely due to the very specific device geometry considered in that work: the cross-sectional dimensions of the nano-wire channel (2.2 nm×2.2 nm) were small enough to be comparable to characteristic Coulomb radius of a single impurity. Surely enough, in this case the impurity simply “blocks” the channel almost entirely and the resulting device I-V characteristics are no longer of a ‘normal’ FET. Basically, Martinez *et al.* [??] have demonstrated in their work that for extremely narrow (~2 nm) wires a single impurity leads to a complete degradation of *any* transistor characteristics. However the question remains, whether the effect of a single impurity is *really* so dramatic for nano-transistors with more realistic Si thickness. For a comparison, ITRS estimates for 10 nm gate length devices Si body thickness of 6 nm.

In this work the presence of an unintentional dopant (UD) in a FinFET with 4 nm Si thickness, which corresponds to the optimization rules developed in this work, has been investigated. The UD is treated in real-space and included in the corresponding doping in the Poisson equation. Two extreme positions of an unintentional dopant have been investigated - (a) when the dopant is sitting close to the source end of the channel ($x = 11$ nm, $y = 2$ nm, $z = 6$ nm) and, (b) when sitting close to the drain end of the channel ($x = 19$ nm, $y = 2$ nm, $z = 6$ nm). Figure 4-22 shows the schematic of the two positions of the unintentional dopant being simulated in this work.

An unintentional dopant sitting in a random location in the channel give rise to localized potential barrier (LB) the impact of which heavily depends on the operating condition of the device. Figure 4-23 presents the 2-D conduction band profile at $y = h/2$ with $h = 4$ nm showing the LB for case (a) at $V_{GS} = 0.2$ V, and $V_{DS} = 50$ mV.

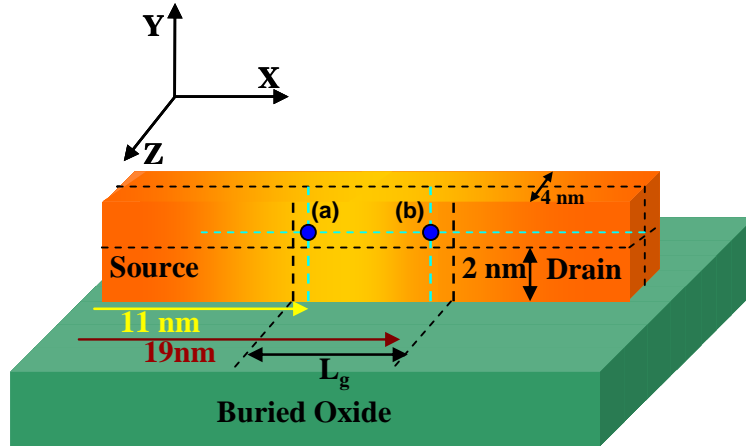


Figure 4-22 Simplified schematic of the considered FinFET showing the positions of the unintentional dopant simulated in this work.

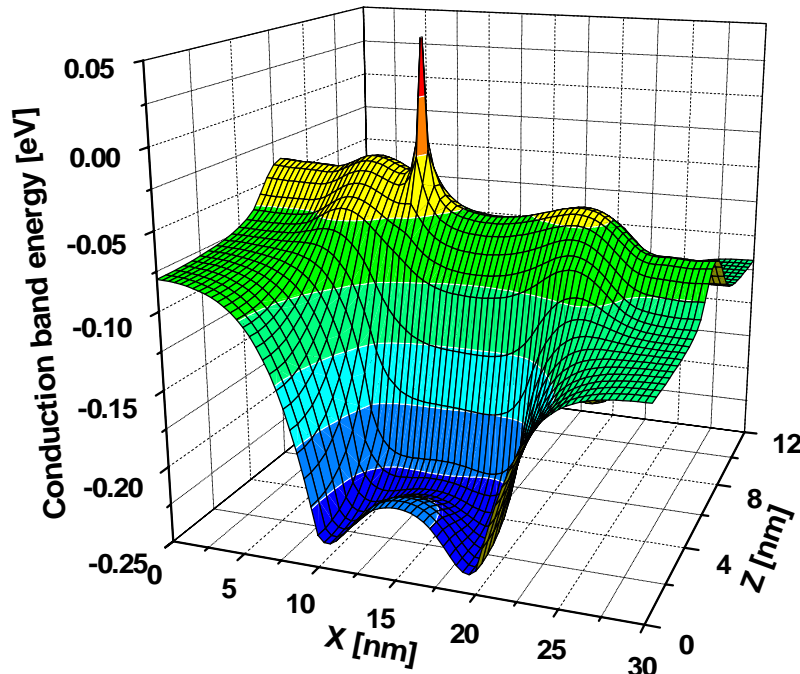


Figure 4-23 Two-dimensional potential energy profile at $y = h/2$ with $h = 4$ nm showing the localized barrier produced by an unintentional dopant sitting near the source end of the channel ($x = 11$ nm, $y = 2$ nm, $z = 6$ nm) at $V_{GS} = 0.2$ V and $V_{DS} = 50$ mV.

Transfer characteristics for these two cases are shown in Figure 4-24 at a drain bias of 0.4V. When UD is sitting near the source end of the channel its impact is more significant than the case when it is sitting near the drain end as can be seen from Figure 4-24.

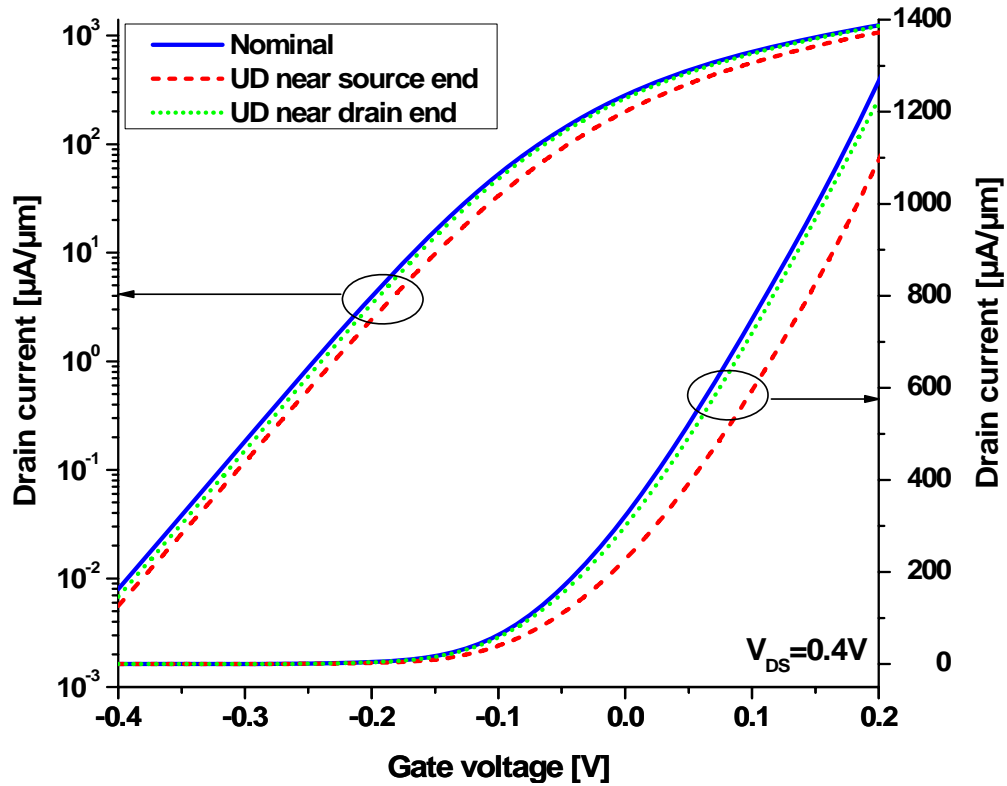


Figure 4-24 Transfer characteristics of the considered DG FinFET in the presence of an unintentional dopant. Solid curve- no UD present, dashed curve- when the UD is sitting near the source end of the channel (case (a)), dotted curve- when the UD is sitting near the drain end of the channel (case (b)).

The percentage reduction in drain current due to the presence of an unintentional dopant for the above mentioned two cases are shown in Figure 4-25. Also shown in the same figure the percentage reduction in drain current for case (a) in the linear regime of operation ($V_{DS} = 50\text{mV}$). One can see from Figure 4-25 that the reduction in drain current heavily depends on the position of the dopant and applied biases. Also for a given drain bias, the reduction in drain current in subthreshold is smaller than the value around threshold and with increasing gate voltage above threshold, the reduction in drain current gradually decreases.

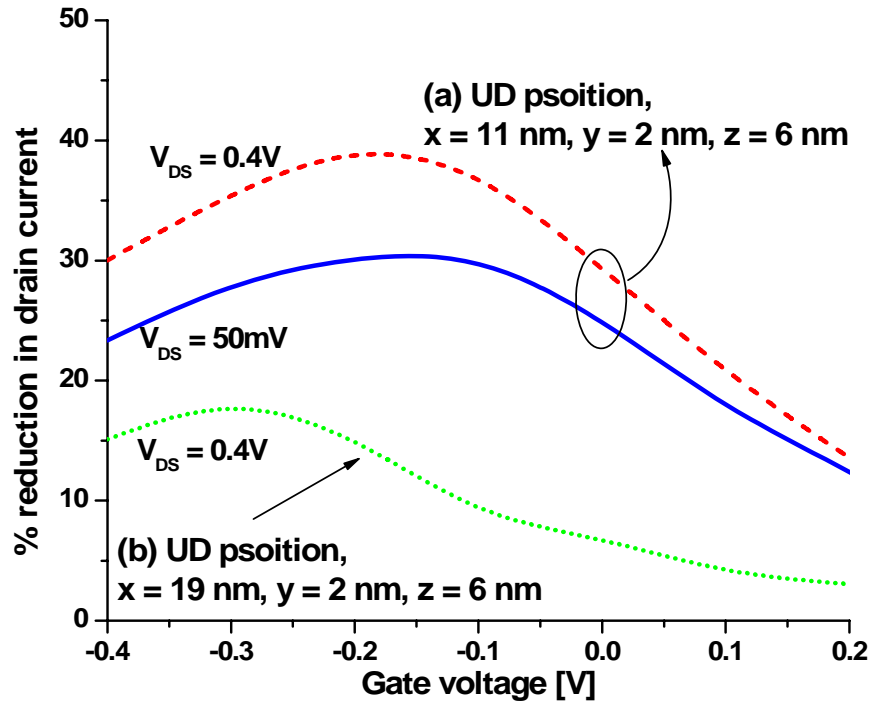


Figure 4-25 Percentage reduction in drain current due to the presence of an unintentional dopant. Solid curve- case (a), at linear regime, dashed curve- case (a), in saturation regime and, dotted curve- case (b), at saturation.

To explain this dependency of the reduction in drain current on applied biases, the 1-D potential energy profiles at the middle of the fin height along X direction for applied gate voltage of -0.4V (with $V_{DS} = 0.4\text{V}$ and 50mV), -0.2V and 0.2V with $V_{DS} = 0.4\text{V}$ are shown in Figure 4-26. One can see from Figure 4-26 that for an applied gate voltage of -0.4V , the height of the intrinsic barrier is higher when the drain bias is low compared to the case with high drain bias. Therefore, for the particular position of the dopant simulated here (case (a)), the relative dominance of the localized barrier at low drain bias is less significant compared to the case with high drain bias. As a result the reduction in drain current at low drain bias is smaller than the reduction at high drain bias. This is true for the entire gate voltage sweep range used in this work.

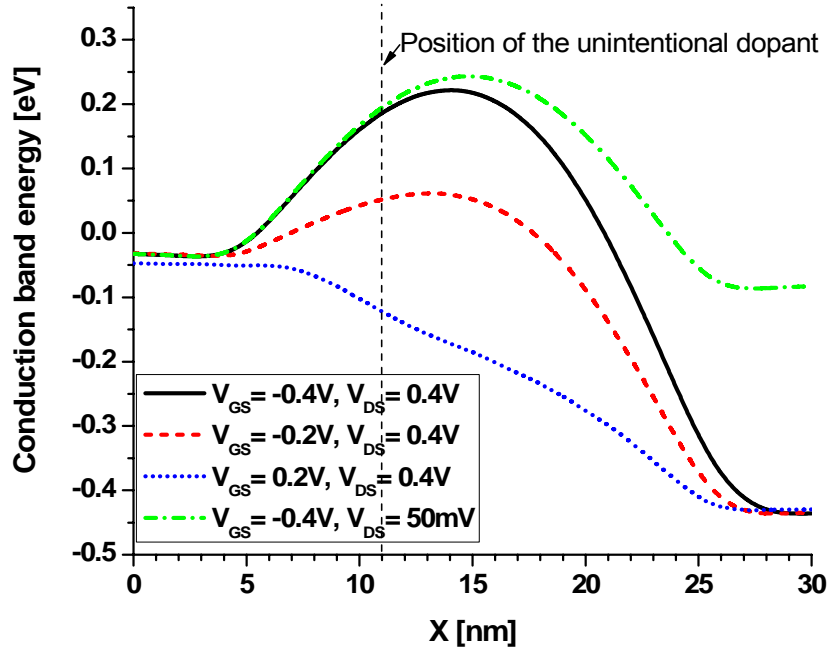


Figure 4-26 One-dimensional conduction band energy profiles along X direction at different bias condition at $y = h/2$ with $h = 4$ nm, $z = 6$ nm.

Considering case (a) at a fixed drain bias of 0.4V, in subthreshold regime ($V_G = -0.4V$) the intrinsic barrier is already high enough and the impact of the LB due to the unintentional dopant is not that prominent but still present as the carriers would see it before reaching the intrinsic barrier. As the gate voltage increases ($V_G = -0.2V$) the intrinsic barrier is lowered significantly and LB becomes the dominant one and consequently drain current reduces significantly. Again for sufficiently high gate voltage ($V_G = 0.2V$) the intrinsic barrier is diminishing and for the particular position of UD (case (a)), carriers would not feel the LB as can be seen from Figure 4-26. Moreover, carriers, after overcoming the intrinsic barrier, are already accelerated by the lateral electric field. Therefore, the reduction in drain current gradually decreases with increasing gate voltage.

Interestingly, the impact of an unintentional dopant on carrier transport is less significant when a top gate is added. The percentage reduction in drain current for TG and DG FinFETs for the case (a) is shown in Figure 4-27.

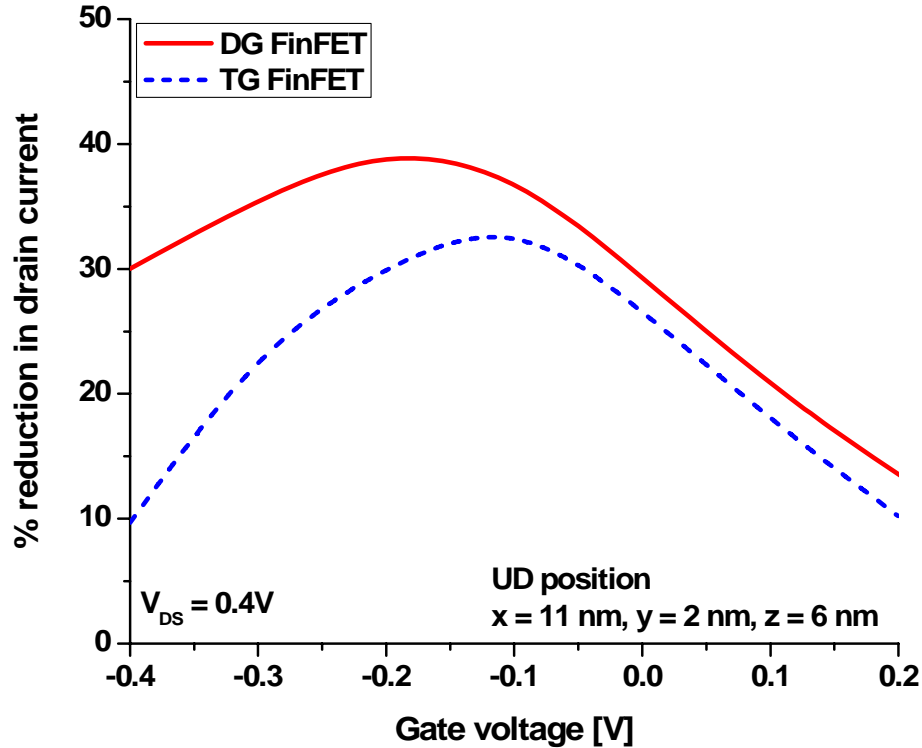


Figure 4-27 Percentage reduction in drain current due to the presence of an UD ((case (a)) for double-gate and tri-gate FinFETs at $V_{DS} = 0.4V$.

Overall, it is clear that TG devices are more efficient in suppressing impurity scattering effects. In on-state the current is reduced by 10% and 14% for TG and DG devices respectively from the corresponding nominal values (Figure 4-27), which gives about 30% better impurity suppression rate for TG over DG devices. Thus, in addition to the 25% improvement in on-current, 6 time reduction of the subthreshold source-drain leakage (Table 4-3), TG FinFETs also have the advantage of being less sensitive to impurity scattering effects when compared to DG FinFETs.

The better impurity suppression in TG devices is illustrated in Figure 4-28, where the 1-D conduction band profiles along the Z direction crossing the position of the unintentional dopant (case (a)) are shown for TG and DG FinFETs.

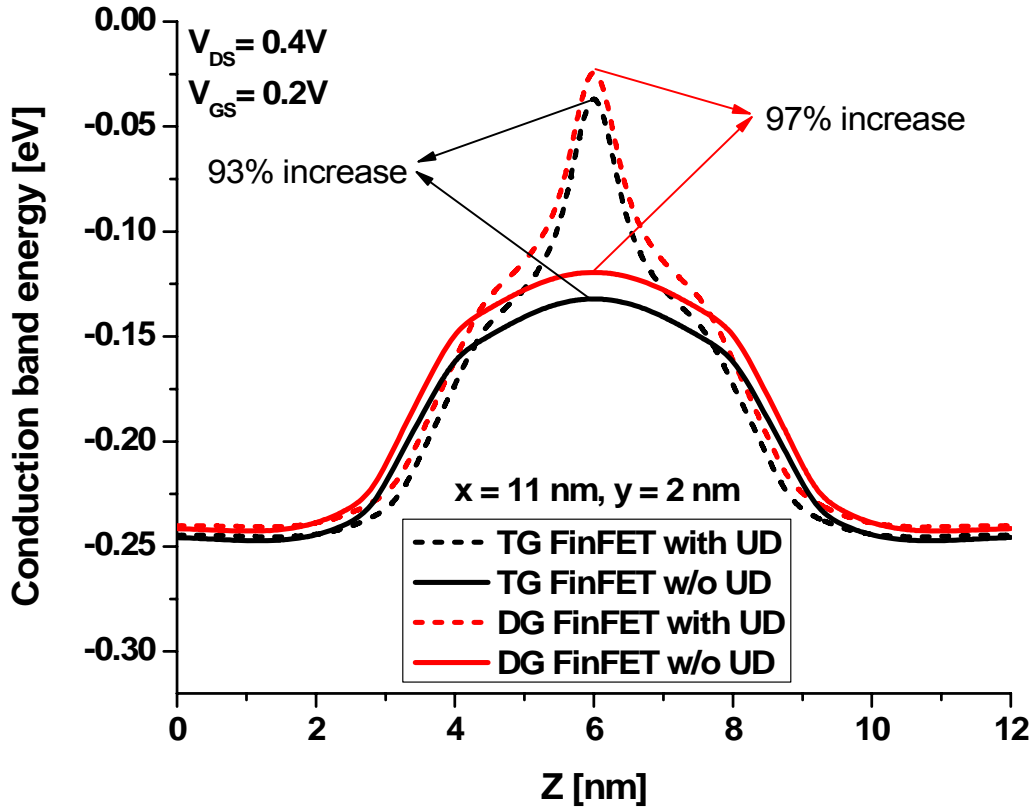


Figure 4-28 One-dimensional conduction band energy profile along Z direction crossing the position of the unintentional dopant (case (a)) at $V_{GS} = 0.2V$, $V_{DS} = 0.4V$. Also shown are the conduction band energy profiles without unintentional dopant.

We, finally, return to the 2D simulation of the experimental FinFET presented in section 4.6.1. It has been conjectured there that the presence of unintentional dopant near the source side could be one of the reasons for the remaining discrepancy between the experiment and simulations. The 3D transport analysis for DG and TG FinFETs presented in this section (e.g., Figure 4-27) shows that for the impurity located near the source side, the reduction in current is at maximum around the threshold voltage. The same feature was observed when comparing the discrepancy between experimental and 2-D simulation (that did not include the impurity effects) results, thus implying that the conjecture made in section 4.6.1 is plausible. Also the trend of the reduction in drain current in the experimental FinFET when compared to the 2-D simulation is similar to the one obtained from 3D simulation of an UD sitting near the source end of the channel for DG FinFET as shown in Figure 4-29.

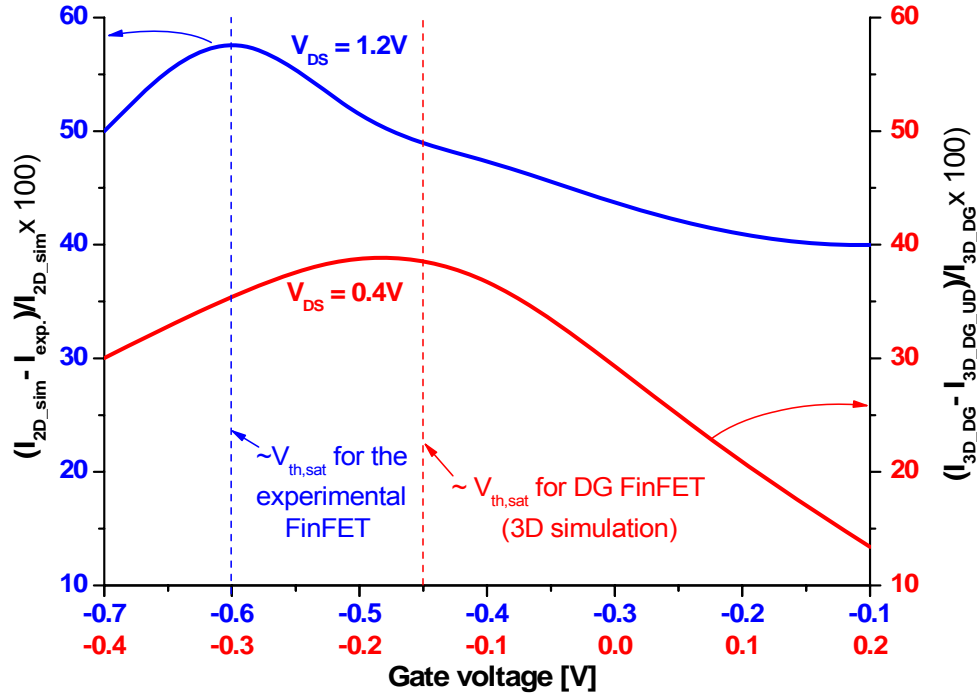


Figure 4-29 Blue curve- percentage reduction in drain current in the experimental device when compared to the corresponding 2-D simulation at $V_{DS} = 1.2\text{V}$. Red curve- percentage reduction in drain current to an UD (case (a)) obtained from 3-D simulation of DG FinFET at $V_{DS} = 0.4\text{V}$.

However, in addition to the position of the unintentional dopant, the exact nature and magnitude of the reduction would also heavily depend on device geometry and the value of the biases (gate voltage and drain voltage).

5. Reduced Density Matrix Formalism and its Application to Modeling RTDs

As already pointed out in the description of the nonequilibrium Green's function formalism in section 3, the current flow through the active region of a semiconductor device or a nanostructure is related to the efficiency of electron injection from the contacts. In disproving the Pauli master equation implementation, Frensley [37] argued that the injection from the contacts implies that density matrix approaches must include off-diagonal elements. Another view of the same problems is given by Pötzt [149], who explicitly worded the need to include the contact-active region interaction as arising from the fact that the active region is a quantum-mechanical open system, and a subsystem of a larger, (almost) closed system. ('Almost' predominantly appears here to caution us that we must not forget the battery, which re-injects electrons that have completed the transition through the structure. However, the inclusion of the battery is by no means a trivial question.) Recently, many authors have emphasized the importance of the proper treatment of contacts [150,151,152,153]. At this point, it is well established that the active region of a semiconductor device or a mesoscopic structure should be treated as an *open system*. In order to attribute mathematical terms to this notion, let us be a bit more precise: the active region is open [dynamically coupled to the contacts (reservoirs, leads)] in the sense that:

- *It exchanges particles with the contacts.* Electrons are injected from/into the leads. Therefore, a formalism that is to be used in the description of the evolution of the current-limiting active region should be able to capture a varying number of electrons. Also, in the case of multiple electrons in the active region, it should in principle be able to account for electron-electron interaction within the active region.
- *It exchanges information with the contacts.* The microscopic state of the contacts is important to what happens in the active region. For instance, if a bias is applied across a semiconductor structure, certain

single-particle states are full in the left and empty in the right contact, and this information should be present in the time evolution of the active region. Therefore, a theoretical description should account for the so-called *memory terms* that explicitly incorporate the microstate of the contacts in the evolution of the active region.

- *It is dynamically coupled to the contacts.* There exists an interaction Hamiltonian between the active region and the contacts, and a correction to the energy spectrum of the active region occurs due to the interaction. Also, the levels of an isolated active region are known to become broadened (the lifetime of these states becomes finite) due to the coupling. Generally, electron-electron interaction between the electrons in the contacts and those in the active region should also be considered [154,155].

These three points are not mutually independent. On the contrary, they are just different aspects of the fact that the active region is a dynamically open many-body system. While there are established formalisms adequate for treatment of small, few-level systems [156], it is very difficult to use the same techniques directly to treat relaxation in mesoscopic structures in a numerically tractable fashion. Nonequilibrium Green's functions (NEGF) and the Wigner function formalisms address parts of the problem, but have their limitations. Basically, to fully describe the relaxation in many-body open systems, a completely general quantum transport formalism still needs to:

- *Not disregard the feedback that a nanostructure active region's evolution has on the contacts or (nonequilibrium) phonons.* This feature requires going beyond the open-boundary approach of conventional nonequilibrium Green's functions [153]. There is a well-known and widely exploited approach to open systems, the approach of *reduced density matrices*, but this approach generally works for small (few-particle) systems, such as molecules. [A (many-body) density matrix (or statistical operator) is a quantity more general than the quantum-mechanical wavefunction (state). For instance, it is capable of capturing information about population of different (many-body) states at finite temperatures. Just like the Hamiltonian, the density matrix (statistical operator) is an operator, acting on the Hilbert space inhabited by wavevectors. Henceforth, *by the density matrix we will assume the generally many-body statistical operator, not the density matrix as defined through single-particle correlation functions*].

- *Regard the full many-body nature of evolution, especially on short timescales.* It has been pointed out by several authors [157,158] that, if we are to have a reasonable chance of understanding sub-femtosecond processes (and consequently be able to go to upper THz frequencies), we need to put some thought into dealing with more than single-particle Green's functions. Namely, the information about buildup and destruction of correlations is stored in multi-particle correlation functions [159]; ultimately, all correlation functions can be deduced from the many-body density matrix, a quantity rich in information but virtually impossible to compute in a real-world electronic systems.

On the other hand, computing the correlation function $-iG^<$ usually offers sufficient information for engineering transport calculations, but the NEGF formalism is formulated for closed systems, so it cannot capture the so-called memory effects. Even so, a rigorous equation of motion for $-iG^<$ includes two-body correlations, the equations for which include three-body correlations, and so on (the Martin-Schwinger hierarchy, see [160]), re-emphasizing the need for multi-particle correlations at short times [113].

- *Does not assume the thermodynamic limit.* Physicists have long been occupied with highly nonequilibrium processes, and the ultimate test of a kinetic approach is how well it describes relaxation towards equilibrium. However, we often cite thermodynamic laws or use certain kinetic equations without even thinking twice that *they are based on the thermodynamic limit* (limit of the system consisting of N particles, $N \rightarrow \infty$, and occupying a volume V , $V \rightarrow \infty$, so that N/V is finite). So it is not true that no one has thought of devising ways to treat real-time relaxation in many-body systems [161], openness and all, but many of the most remarkable treatments are in the field of plasma physics, within the thermodynamic limit.

The problem with mesoscopic systems is that they contain neither very few, nor very many particles. For mesoscopic systems, not even the laws of thermodynamics hold in the form we are used to (see a very insightful paper by Gross [162], on the validity of the second law of thermodynamics in mesoscopic structures). This is precisely why short-time relaxation in these structures is so challenging: you can neither make common approximations, nor can you treat it with brute computational force without a significant loss of physics.

In section 5.1, we demonstrate how the approach of the reduced density matrices can be adapted to help build a bridge towards nonequilibrium transport calculations in mesoscopic systems. As an example, in section 5.2 we will derive a Landauer-like formula for a ballistic resonant-tunneling diode from first principles: by starting with a model Hamiltonian and using the reduced density matrix formalism. In section 5.3, we will also point out how the NEGF formalism can be adapted to include the memory effects important in truly open systems.

5.1 Partial-trace-free approach to open systems. Equations with memory dressing

5.1.1 Reduced density matrix (statistical operator)

Describing the evolution of a quantum-mechanical system, coupled to its environment and influenced by external driving fields, is one of the oldest and most important problems in quantum mechanics (see, e.g., Ref. [161]). The problem actually lies in the attempt to describe the system's time development, without necessarily having to collect too much information about the environment.

Open systems are traditionally described by a *reduced density matrix* (a.k.a. *reduced statistical operator*) formalism. The density matrix is a quantity more general than the quantum-mechanical wavefunction. Namely, real world quantum-mechanical systems can rarely be described by just a single wavefunction (i.e., be in a pure state). For instance, a free gas of electrons at zero temperature can be described by a single many-body wavefunction—a single Slater determinant that corresponds to all one-particle levels below the Fermi level being full, and those above it being empty. However, at a temperature greater than zero, there is a non-zero probability of the electron gas occupying other many-body states, such as that with one hole below the Fermi level and one electron above it. The (many-body) density matrix is a quantity capable of capturing information about the population of (many-body) states. Just like the Hamiltonian, the density matrix (i.e., statistical operator) is an operator, acting on the Hilbert space inhabited by wavevectors. If the quantum-mechanical system is closed, meaning that its dynamics are not influenced by the state of any external system, its density matrix ρ obeys the *quantum Liouville equation*,

$$\frac{d\rho}{dt} = -i[h, \rho],$$

where h is the total (many-body) Hamiltonian.

However, a real quantum-mechanical system is seldom closed. Rather, it is a part of a larger, approximately closed system. We therefore speak of the (open) *system*, while the rest is termed the *environment*. In the case of electronic transport, the electrons in the active region constitute the system, while the environment includes contacts, phonons, etc. The information about the system (our active region electrons) is obtained by tracing out the environmental states, which produces the *reduced system density matrix* $\rho_S = \text{Tr}_E \rho$, a quantity that contains all the quantum-mechanical information about the system (active region). The equation of motion for ρ_S will indeed contain information about the environment (the memory terms).

The reduced density matrix approaches are very popular among physical chemists, and for generally small systems with at most a few particles. The problem is that we wish to extract the information about the open system without gathering too much information about the environment (even though the measurements are actually made in the environment). Many approaches have been devised to perform this task (see, for instance, a comprehensive recent book by Breuer and Petruccione [156]). A widely used approach to obtaining the equation of motion for the reduced system density matrix, is by using a *projection-operator technique*, introduced by Nakajima [163], Zwanzig [164] and Mori [165]. Variants of the projection operator technique have successfully been used in many fields. For example, Argyes and Kelley [166] presented a theory of linear response in spin-systems, Barker and Ferry [167] treated quantum transport in very small semiconductor devices, Kassner [168] analyzed relaxation in systems with initial system-bath coupling, Sparpaglione and Mukamel [169] presented a theory for electron transfer in polar media, followed by analyses of condensed phase electron transfer by Hu and Mukamel [170], and Romero-Rochin and Oppenheim [171] addressed relaxation of a two-level systems weakly coupled to a bath. Essentially, all variations of the technique rely upon the use of two complementary projection operators to generate the equations of motion for two mutually orthogonal projections of the total 'system+environment' density matrix. Equations of motion for the two projections are coupled, and the

equation for the relevant projection, i.e., the one yielding the reduced density matrix after a partial trace is taken over the environment states, can be obtained in a closed form. This requires incorporating a formal solution of the equation for the orthogonal projection. The resulting equation of motion for the reduced density matrix typically exhibits non-Markovian (or ‘time-convolution’) behavior. The non-Markovian nature and the need for full knowledge of the ‘system+environment’ (in order to take the partial trace over environmental states) are significant constraints on this approach. An equation of motion containing a term with a memory-kernel is particularly difficult to solve self-consistently, and one must often be satisfied with the fast-modulation (Markov) limit.

In response to the difficulties arising from memory kernels in nonequilibrium statistical mechanics, Tokuyama and Mori [172] first proposed a time-convolutionless (also known as ‘memoryless’) equation of motion in the Heisenberg picture. Soon afterwards, derivations of Shibata *et al.* [173,174] in the Schrödinger picture appeared, and this approach is the basis for much work that followed. This includes Saeki’s analysis of linear response of an externally-driven system coupled to a heat bath [175] and work on systems coupled to a stochastic reservoir [176,177]. The latter was extended by Ahn to formulate the quantum kinetic equations for semiconductors [178] and arrive at a theory of optical gain in quantum-well lasers [179]. Chang and Skinner [180] applied the time-convolutionless approach to analyze relaxation of a two-level system strongly coupled to a harmonic bath. More recently, Ahn *et al.* treated noisy quantum channels [181] and quantum information processing [182], and Golosov and Reichmann [183] analyzed condensed-phase charge-transfer processes. Both the time-convolutionless equation of motion [173-183], and the non-Markovian equations described in the previous paragraph [163-171], are based on projection-operator techniques. The difference is that, in the time-convolutionless approach, the memory effects are taken into account by evaluating particular evolution operators, which couple states of the system with the environment states, rather than through a term with a memory-kernel. However, the entire approach is based on the assumption of invertibility of one of the evolution operators, and this is an important and potentially problematic issue. Once a time-convolutionless equation for the relevant projection is obtained, a partial trace with respect to the environment states is performed, as before, to uncover the evolution of the reduced density matrix.

The resulting time-convolutionless equation of motion for the reduced density matrix still has two major shortcomings. First, it has an explicit dependence on the choice of the projection operator (or, on the environment density matrix that induces the projection operator). This is unphysical, as the projection operator is just a tool, an external assumption, and in the end the equation of motion for the reduced density matrix should not depend on such assumptions. This does not mean that the time-convolutionless approach is incorrect, it just means that *all the equations for different projection operators eventually must reduce to one and the same equation*. The second shortcoming is that the partial trace used to obtain the evolution of the reduced density matrix (from the equation of motion for the relevant projection) requires one to evaluate large and unpleasant evolution matrices, only to extract significantly less information after the partial trace. In other words, one must obtain much more information than actually necessary. With increasing size of the ‘system+environment’, and inclusion of a time dependence due to external forces, this obstacle becomes more serious.

In this Section, two main results are presented. First, we introduce a very useful and simple isomorphism between the space of operators acting on the system’s Hilbert space and the unit-eigenspace of the projection operator induced by the uniform density matrix of the environment. This allows us to derive a time-convolutionless equation of motion for the reduced density matrix, which is free of the partial trace with respect to environment states. This isomorphism enables us to effectively perform a partial trace with respect to the environment states without actually performing it: what we perform instead is a well-defined basis transformation. Being partial-trace-free, this equation successfully addresses one of the shortcomings mentioned in the previous paragraph. It depends on submatrices considerably smaller than those in the conventional time-convolutionless approach [173-183], which makes the present approach especially interesting for numerical application in those cases where evaluation of large evolution matrices is particularly undesirable. Secondly, we point out that a time-convolutionless approach, conventional or partial-trace-free, is based on the assumption of invertibility of a particular ‘entanglement operator’, and show that this assumption is equivalent to the assumption of reversibility of the system’s evolution. This puts constraints on the applicability of time-convolutionless approaches in general. In particular, we investigate the application of the approach to the description of a steady state in far-from-equilibrium situations.

5.1.2 Basic definitions

Consider a system S , interacting with its environment E , so that the ‘system+environment’ ($S+E$) is either closed, or influenced by external driving fields that are assumed known and unaffected by the feedback from $S+E$. The Hilbert spaces of both the environment and the system, \mathcal{H}_E and \mathcal{H}_S respectively, are assumed to be finite-dimensional, $d_E = \dim \mathcal{H}_E$, $d_S = \dim \mathcal{H}_S$. These two spaces form a tensor product Hilbert space of the ‘system+environment’, $\mathcal{H}_{S+E} = \mathcal{H}_E \otimes \mathcal{H}_S$, with dimensionality $d = \dim \mathcal{H}_{S+E} = d_E d_S$. The spaces of operators acting on \mathcal{H}_E , \mathcal{H}_S and \mathcal{H}_{S+E} will be denoted by \mathcal{H}_E^2 , \mathcal{H}_S^2 and \mathcal{H}_{S+E}^2 , respectively, whereas the elements of \mathcal{H}_E^2 , \mathcal{H}_S^2 and \mathcal{H}_{S+E}^2 (i.e., operators on \mathcal{H}_E , \mathcal{H}_S and \mathcal{H}_{S+E}) will be denoted by lowercase letters, Greek or Roman. Moreover, operators acting on \mathcal{H}_E^2 , \mathcal{H}_S^2 and \mathcal{H}_{S+E}^2 , sometimes called superoperators, will be denoted by capital Roman letters. When there is no risk of confusion, we will simply refer to them as operators.

Let us choose a basis $\{|i\rangle \mid i = 1, \dots, d_E\}$ in \mathcal{H}_E , and a basis $\{|\alpha\rangle \mid \alpha = 1, \dots, d_S\}$ in \mathcal{H}_S , which induce a tensor-product basis $\{|i\alpha\rangle \equiv |i\rangle \otimes |\alpha\rangle \mid i = 1, \dots, d_E; \alpha = 1, \dots, d_S\}$ in \mathcal{H}_{S+E} . These naturally give rise to the following expanded bases:

$$\begin{aligned} \{|ij\rangle \equiv |i\rangle \otimes |j\rangle \mid i, j = 1, \dots, d_E\} &\text{ in } \mathcal{H}_E^2, \quad \{|\alpha\beta\rangle \equiv |\alpha\rangle \otimes |\beta\rangle \mid \alpha, \beta = 1, \dots, d_S\} &\text{ in } \mathcal{H}_S^2, \\ \{|i\alpha, j\beta\rangle \equiv |i\alpha\rangle \otimes |j\beta\rangle \mid i, j = 1, \dots, d_E; \alpha, \beta = 1, \dots, d_S\} &\text{ in } \mathcal{H}_{S+E}^2 \end{aligned} \quad (5-1)$$

For an operator x on \mathcal{H}_{S+E} (i.e., an element of \mathcal{H}_{S+E}^2), one can write

$$x = |x\rangle = \underbrace{\sum_{i,j=1}^{d_E} \sum_{\alpha,\beta=1}^{d_S} x_{j\beta}^{i\alpha} |i\alpha\rangle \langle j\beta|}_{\text{as operator on } \mathcal{H}_{S+E}} = \underbrace{\sum_{i,j=1}^{d_E} \sum_{\alpha,\beta=1}^{d_S} x^{i\alpha, j\beta} |i\alpha, j\beta\rangle}_{\text{as vector in } \mathcal{H}_{S+E}^2}, \quad x_{j\beta}^{i\alpha} = x^{i\alpha, j\beta}, \quad (5-2)$$

whereas for an operator A that acts on \mathcal{H}_{S+E}^2 , the following form is valid

$$A = \sum_{i,j,p,q=1}^{d_E} \sum_{\alpha,\beta,\sigma,\gamma=1}^{d_S} A_{p\sigma, q\gamma}^{i\alpha, j\beta} |i\alpha, j\beta\rangle \langle p\sigma, q\gamma|. \quad (5-3)$$

Analogous expressions are easily obtained for elements of \mathcal{H}_E^2 , \mathcal{H}_S^2 , and operators on \mathcal{H}_E^2 , \mathcal{H}_S^2 .

The total Hamiltonian h , acting on \mathcal{H}_{S+E} , consists of the system part $h_{\text{sys}} = 1_E \otimes h_S$, the environment part $h_{\text{env}} = h_E \otimes 1_S$ and the interaction part h_{int} , so that

$$h = h_{\text{sys}} + h_{\text{env}} + h_{\text{int}}. \quad (5-4)$$

It is important to note that driving fields, provided by applied potentials, are included as part of the system. These are assumed to arise for $t > t_0$, and provide an explicit time dependence for h . The evolution of the total $S+E$ density matrix ρ is given by the quantum Liouville equation (we use units $\hbar = 1$)

$$\frac{d\rho}{dt} = -i[h, \rho] \equiv -iL\rho, \quad (5-5)$$

where L is the Liouville superoperator (superoperators will be denoted by capital roman letters) corresponding to h , and is of the form $L = L_{\text{sys}} + L_{\text{env}} + L_{\text{int}}$. Equation (5) actually represents a system of $d_E^2 d_S^2$ linear first-order differential equations on the time interval $t \in [t_0, \infty)$. Since L is continuous throughout this entire interval, we are guaranteed to have a unique solution of Eq. (5-5) on $t \in [t_0, \infty)$ for a given initial condition $\rho(t_0)$ [184]. With T^c (T^a) denoting the chronological (anti-chronological) time-ordering, Θ being the Heaviside step function, and t_0 being the initial time at which we assume we know $\rho(t_0)$, the formal solution of Eq. (5-5) is given by

$$\rho(t) = U(t, t_0) \rho(t_0),$$

$$U(t, t') = \Theta(t - t') T^c \exp\left(-i \int_{t'}^t d\tau L(\tau)\right) + \Theta(t' - t) T^a \exp\left(i \int_t^{t'} d\tau L(\tau)\right). \quad (5-6)$$

The quantity that describes the evolution of the system S is the reduced density matrix ρ_S , defined by

$$\rho_S = \text{Tr}_E(\rho), \quad (5-7)$$

where $\text{Tr}_E(\dots)$ denotes the partial trace over the environment states. The goal is to deduce how ρ_S evolves, without having to gather too much information about the environment.

5.1.3 Projection-operator technique. Conventional time-convolutionless equation of motion

A class of approaches based on projection operators yield the so-called *time-convolutionless equation* (TCE) of motion for the projection $P\rho$ of the total $S+E$ density matrix ρ , where P is a projector generated by an arbitrary environment density matrix [156,172-183]. The term “time-convolutionless” means that the differential equation for $P\rho(t)$ at a time t formally depends only on $P\rho(t)$ and on $Q\rho(t_0)$ at the initial time t_0 , where $Q = 1 - P$. Instead of the notorious integral with the memory kernel, which requires knowing $Q\rho(\tau)$ at all times τ in the past ($t_0 \leq \tau \leq t$), special (invertible) operators are introduced, which account for the system’s memory of the environment’s evolution.

Let us now choose an arbitrary environmental density matrix $\tilde{\rho}_E$, which is constant in time. We introduce time-independent projection operators \tilde{P} and \tilde{Q} , which are associated with $\tilde{\rho}_E$ and act on \mathcal{H}_{S+E}^2 , as

$$\tilde{P}x = \tilde{\rho}_E \otimes \text{Tr}_E x, \quad \tilde{Q} = 1 - \tilde{P}, \quad x \in \mathcal{H}_{S+E}^2. \quad (5-8)$$

In particular, since $\text{Tr}_E(\tilde{\rho}_E) = 1$ ($\tilde{\rho}_E$ is a proper density matrix), it follows that

$$\text{Tr}_E(\tilde{P}\rho) = \text{Tr}_E(\tilde{\rho}_E) \cdot \text{Tr}_E \rho = \text{Tr}_E \rho = \rho_S. \quad (5-9)$$

Therefore, the evolution of $\text{Tr}_E(\tilde{P}\rho)$ should be described by an equation that is independent of $\tilde{\rho}_E$ (or, equivalently, \tilde{P}).

Since \tilde{P} and \tilde{Q} are time independent, they commute with the time derivative. Therefore, for the equations of motion of $\tilde{P}\rho$ and $\tilde{Q}\rho$, we obtain from Eq. (5-5)

$$\frac{d\tilde{P}\rho(t)}{dt} = -i\tilde{P}L(t)\rho(t) = -i\tilde{P}L(t)\tilde{P}\rho(t) - i\tilde{P}L(t)\tilde{Q}\rho(t), \quad (5-10a)$$

$$\frac{d\tilde{Q}\rho(t)}{dt} = -i\tilde{Q}L(t)\rho(t) = -i\tilde{Q}L(t)\tilde{Q}\rho(t) - i\tilde{Q}L(t)\tilde{P}\rho(t). \quad (5-10b)$$

A formal solution of Eq. (5-10b) is of the form

$$\tilde{Q}\rho(t) = -i \int_{t_0}^t dt' H(t, t') \tilde{Q}L(t') \tilde{P}U(t', t)\rho(t) + H(t, t_0) \tilde{Q}\rho(t_0), \quad (5-11)$$

where for $t > t'$

$$H(t, t') = T^c \exp\left(-i \int_{t'}^t ds \tilde{Q}L(s)\tilde{Q}\right). \quad (5-12)$$

Upon introducing

$$K(t; t_0) = 1 + i \int_{t_0}^t dt' H(t, t') \tilde{Q}L(t') \tilde{Q}U(t', t), \quad (5-13)$$

Eq. (5-11) can be rearranged to give

$$K(t; t_0) \tilde{Q}\rho(t) = [1 - K(t; t_0)] \tilde{P}\rho(t) + H(t, t_0) \tilde{Q}\rho(t_0). \quad (5-14)$$

The second time variable in the argument of $K(t; t_0)$ is not exactly a variable. Namely, t_0 denotes the fixed initial time, which defines the initial conditions for K , and is separated from the first time variable t by a semi-colon instead of a comma to indicate that it is not a variable, but a fixed parameter. If $K(t; t_0)$ is invertible, which is an important question to which we will return, Eq. (5-14) becomes

$$\tilde{Q}\rho(t) = [K(t; t_0)^{-1} - 1] \tilde{P}\rho(t) + K(t; t_0)^{-1} H(t, t_0) \tilde{Q}\rho(t_0), \quad (5-15)$$

and (5-10a) can be rewritten as

$$\frac{d\tilde{P}\rho(t)}{dt} = -i\tilde{P}L(t)K(t; t_0)^{-1} \tilde{P}\rho(t) - i\tilde{P}L(t)K(t; t_0)^{-1} H(t, t_0) \tilde{Q}\rho(t_0). \quad (5-16)$$

After taking a partial trace with respect to the environment states, Eq. (5-16) will give us the equation of motion for ρ_S as

$$\begin{aligned} \frac{d\rho_S(t)}{dt} &= -i \text{Tr}_E \left[\tilde{P}L(t)K(t; t_0)^{-1} \tilde{P}\rho(t) \right] - i \text{Tr}_E \left[\tilde{P}L(t)K(t; t_0)^{-1} H(t, t_0) \tilde{Q}\rho(t_0) \right] \\ &= -i \text{Tr}_E \left[L(t)K(t; t_0)^{-1} \tilde{\rho}_E \otimes \rho_S(t) \right] - i \text{Tr}_E \left[L(t)K(t; t_0)^{-1} H(t, t_0) \tilde{Q}\rho(t_0) \right] \\ &= -i \left\langle L(t)K(t; t_0)^{-1} \right\rangle_{\tilde{\rho}_E} \rho_S(t) - i \text{Tr}_E \left[L(t)K(t; t_0)^{-1} H(t, t_0) \tilde{Q}\rho(t_0) \right]. \end{aligned} \quad (5-17)$$

The last line was obtained by introducing $\langle \dots \rangle_{\tilde{\rho}_E}$, which maps operators acting on \mathcal{H}_{S+E}^2 onto operators

acting on \mathcal{H}_S^2 : for any given A , $\langle A \rangle_{\tilde{\rho}_E}$ is defined as $\left[\langle A \rangle_{\tilde{\rho}_E} \right]_{\sigma\gamma}^{\alpha\beta} = \sum_{i,p,q=1}^{d_E} A_{p\sigma,q\gamma}^{i\alpha,i\beta} (\tilde{\rho}_E)^{pq}$ [see Eq. (5-3)].

Eq. (5-16) is what we will refer to as *the conventional time-convolutionless equation of motion for $\rho_S(t)$* [173-183].

There are two features of Eq. (5-17) that need attention. First, within the partial trace in both terms on the right-hand-side of Eq. (5-17), there is an explicit dependence on the choice of the projection operator \tilde{P} (or, equivalently, on the environment density matrix $\tilde{\rho}_E$ that induces the projection operator), so one must make a choice of $\tilde{\rho}_E$ to actually be able to use Eq. (5-17). As $\tilde{\rho}_E$ is just an external arbitrary tool, in the end the equation of motion for $\rho_S(t)$ should not depend on it. This does not mean that the time-convolutionless approach is incorrect, it just means that all the equations for different projection operators ought to, eventually, reduce to one and the same equation. What this equation is, or how to rigorously prove mutual equivalence of equations obtained using different projection operators, is not presently known. Secondly, due to the partial trace, one must evaluate generally large and unpleasant evolution matrices U , H and K , only to extract significantly less information after the partial trace. Much more information than actually necessary has to be obtained. With increasing size of the ‘system+environment’ and inclusion of a time-dependence due to external driving, this obstacle becomes increasingly serious when numerical implementation of Eq. (5-17) is considered. Thirdly, it is obvious that the entire approach is based on invertibility of K , which may be a requirement hard to fulfill with some approximations. These are the most prominent problems with the applicability of the conventional time-convolutionless approach, and they are addressed in the following sections.

5.1.4 Eigenproblem of a projection operator. Partial-trace-free approach

As noted previously, a serious constraint on the application of the conventional time-convolutionless approach (5-17) in large systems is evaluation of large matrices, when not all the information contained in them is needed. In this section, we derive a partial-trace-free equation of motion for ρ_S , which successfully addresses this issue.

Some features of the eigenvalue problem of \tilde{P} are virtually obvious: \tilde{P} is Hermitian, and it is easily verified that \tilde{P} is idempotent, i.e., $\tilde{P}^2 = \tilde{P}$. Hence, the eigenvalues are 0 and 1. On the other hand, by construction (5-8), the image space of \tilde{P} corresponds with \mathcal{H}_S^2 , so one sees that the eigenspace of \tilde{P} , corresponding to the eigenvalue 1, ought to be isomorphic to \mathcal{H}_S^2 . An important and useful point that we need to recognize is the decomposition of \mathcal{H}_{S+E}^2 into a direct sum of eigenspaces of \tilde{P}

$$\mathcal{H}_{S+E}^2 = \left(\mathcal{H}_{S+E}^2 \right)_{\tilde{P}=1} \oplus \left(\mathcal{H}_{S+E}^2 \right)_{\tilde{P}=0}, \quad (5-18)$$

where $\left(\mathcal{H}_{S+E}^2 \right)_{\tilde{P}=1}$ is the d_S^2 -dimensional unit eigenspace, and $\left(\mathcal{H}_{S+E}^2 \right)_{\tilde{P}=0}$ is the $d_S^2(d_E^2 - 1)$ -dimensional zero eigenspace.

If we choose an orthonormal eigenbasis of \tilde{P} , $\left\{ |\tilde{n}\rangle \mid \tilde{n} = 1, \dots, d_E^2 d_S^2 \right\}$, so that, according to decomposition (5-18), the first d_S^2 basis vectors span $\left(\mathcal{H}_{S+E}^2 \right)_{\tilde{P}=1}$, we will be able to write

$$\tilde{P} = \sum_{\tilde{n}=1}^{d_S^2} |\tilde{n}\rangle \langle \tilde{n}|. \quad (5-19)$$

The crucial question that we ask is: is it possible not to mix original basis vectors $|i\alpha, j\beta\rangle$ with different α, β to obtain a given $|\tilde{n}\rangle$? The answer is: yes, if and only if the density matrix inducing the projection operator is the uniform density matrix [185], i.e.,

$$\bar{\rho}_E \equiv d_E^{-1} \cdot 1_{d_E \times d_E}, \quad (5-20)$$

with the associated projection operator denoted by \bar{P} . One indeed finds that the vectors defined as

$$|\overline{\alpha\beta}\rangle \equiv \frac{1}{\sqrt{d_E}} \sum_{i=1}^{d_E} |i\alpha, i\beta\rangle \quad (5-21)$$

constitute an orthonormal basis within the unit-eigenspace of \bar{P} , namely

$$\bar{P}|\overline{\alpha\beta}\rangle = |\overline{\alpha\beta}\rangle, \quad \langle \overline{\alpha\beta} | \overline{\sigma\gamma} \rangle = \delta_{\alpha\sigma} \delta_{\beta\gamma}, \quad (\forall \alpha, \beta, \sigma, \gamma) . \quad (5-22)$$

\bar{P} can therefore be written as

$$\bar{P} = \sum_{\alpha, \beta=1}^{d_S} |\overline{\alpha\beta}\rangle \langle \overline{\alpha\beta}|, \quad (5-23)$$

and it follows that

$$\bar{P}x = \sum_{\alpha, \beta=1}^{d_S} (\bar{P}x)^{\overline{\alpha\beta}} |\overline{\alpha\beta}\rangle, \quad (5-24)$$

where

$$(\bar{P}x)^{\overline{\alpha\beta}} = \langle \overline{\alpha\beta} | x \rangle = \frac{1}{\sqrt{d_E}} \sum_{i=1}^{d_E} \langle i\alpha, i\beta | x \rangle = \frac{1}{\sqrt{d_E}} \sum_{i=1}^{d_B} x^{i\alpha, i\beta}, \quad (5-25)$$

(we identified $x = |x\rangle$ to take advantage of the compactness of the Dirac notation). Using the fact that $\text{Tr}_E x$, being a vector in \mathcal{H}_S^2 , is written in terms of the basis $\{|\alpha\beta\rangle\}$ as

$$\text{Tr}_E x = \sum_{\alpha, \beta=1}^{d_S} (\text{Tr}_E x)^{\alpha\beta} |\alpha\beta\rangle = \sum_{\alpha, \beta=1}^{d_S} \left(\sum_{i=1}^{d_E} x^{i\alpha, i\beta} \right) |\alpha\beta\rangle, \quad (5-26)$$

from Eq. (5-25), we obtain the crucial equation for the rest of this work:

$$(\bar{P}x)^{\overline{\alpha\beta}} = (\text{Tr}_E x)^{\alpha\beta} / \sqrt{d_E}. \quad (5-27)$$

Eq. (5-27) represents a very useful isomorphism between $(\mathcal{H}_{S+E}^2)_{\bar{P}=1}$, (the unit-eigenspace of \bar{P}) and \mathcal{H}_S^2 . We have managed to ‘preserve’ the identity of basis vectors $\{|\alpha\beta\rangle\}$ from \mathcal{H}_S^2 when mapping to $(\mathcal{H}_{S+E}^2)_{\bar{P}=1}$. This isomorphism is basis-induced, and enables us to effectively perform a partial trace with respect to the environment states without actually performing it: instead, what we do perform is a basis transformation in \mathcal{H}_{S+E}^2 .

This basis in the unit eigenspace is complemented by an orthonormal basis in the zero eigenspace [185]. According to the decomposition (5-18), a vector $x \in \mathcal{H}_{S+E}^2$ can be represented in the complete eigenbasis of \bar{P} by a column

$$x = \begin{bmatrix} x_1 \\ x_2 \end{bmatrix}, \quad (5-28a)$$

and the projectors are represented by

$$\bar{P} = \begin{bmatrix} 1 & 0 \\ 0 & 0 \end{bmatrix}, \quad \bar{Q} = \begin{bmatrix} 0 & 0 \\ 0 & 1 \end{bmatrix}. \quad (5-28b)$$

Taking the partial trace with respect to E of any $S+E$ observable x gives what the system sees of this observable. It is important to stress that, from now on, we will make no distinction between a system variable $x_S = \text{Tr}_E x$ and its representation column in the basis $\{|\alpha\beta\rangle\}$ of \mathcal{H}_S^2 . Therefore, for x_S being the representation column of $x_S = \text{Tr}_E x$ in the basis $\{|\alpha\beta\rangle\}$ of \mathcal{H}_S^2 , according to Eq. (5-27) we obtain

$$x_S = x_1 \cdot \sqrt{d_E}. \quad (5-29)$$

In the eigenbasis of \bar{P} , a superoperator A that acts on \mathcal{H}_{S+E}^2 is, in general, represented by a block-matrix form

$$A = \begin{bmatrix} A_{11} & A_{12} \\ A_{21} & A_{22} \end{bmatrix}. \quad (5-30a)$$

If the operator A_{sys} is a system operator, i.e., an operator of the form $A_{\text{sys}} = I_E \otimes A_S$, where A_S acts on \mathcal{H}_S^2 , then A_{sys} commutes with \bar{P} , and is therefore represented by a block-diagonal form in the eigenbasis of \bar{P} . Furthermore, the block-diagonal form is such that the upper-left block-matrix is exactly the one representing A_S in the basis $\{|\alpha\beta\rangle\}$ (see the Appendix B of Ref. [186]), namely,

$$A_{\text{sys}} = I_E \otimes A_S = \begin{bmatrix} A_S & 0 \\ 0 & A_2 \end{bmatrix}. \quad (5-30b)$$

5.1.5 Partial-trace-free time-convolutionless equation of motion for the reduced density matrix

According to Eq. (5-27), there is a very simple correspondence between how $\bar{P}\rho$ looks in terms of the eigenbasis $\{|\overline{\alpha\beta}\rangle\}$, (4.21), and how ρ_S looks in terms of the basis $\{|\alpha\beta\rangle\}$ in \mathcal{H}_S^2 . Using this result, we conclude that the evolution of the representation matrix of ρ_S can be tracked in $(\mathcal{H}_{S+E}^2)_{\bar{P}=1}$ directly, by following the evolution $\bar{P}\rho$ without taking the partial trace with respect to the environment states (which would mean going back to \mathcal{H}_S^2). However, that everything must be written in the eigenbasis of \bar{P} (Figure 5-1), whose first d_S^2 vectors are $\{|\overline{\alpha\beta}\rangle\}$, Eq. (5-21), and the rest is straightforwardly constructed. In the eigenbasis of \bar{P} , the total density matrix is given by

$$\rho = \begin{bmatrix} \rho_1 \\ \rho_2 \end{bmatrix}, \quad \rho_S = \rho_1 \sqrt{d_E}. \quad (5-31a)$$

The Liouville operator and the evolution operator from Eq. (5-6) are given by the block forms

$$L(t) = \begin{bmatrix} L_{11}(t) & L_{12}(t) \\ L_{21}(t) & L_{22}(t) \end{bmatrix}, \quad U(t, t') = \begin{bmatrix} U_{11}(t, t') & U_{12}(t, t') \\ U_{21}(t, t') & U_{22}(t, t') \end{bmatrix}, \quad (5-31b)$$

where $L_{21} = (L_{12})^+$ (L is Hermitian), and U is unitary. When equations (5-5) and (5-6) are written out in their matrix representations, we obtain

$$\begin{aligned} \frac{d\rho_1}{dt} &= -iL_{11}(t)\rho_1(t) - iL_{12}(t)\rho_2(t), \\ \frac{d\rho_2}{dt} &= -iL_{21}(t)\rho_1(t) - iL_{22}(t)\rho_2(t), \end{aligned} \quad (5-32a)$$

and

$$\begin{aligned} \rho_1(t) &= U_{11}(t, t')\rho_1(t') + U_{12}(t, t')\rho_2(t'), \\ \rho_2(t) &= U_{21}(t, t')\rho_1(t') + U_{22}(t, t')\rho_2(t'). \end{aligned} \quad (5-32b)$$

The block forms of H and K from Eqs. (5-12) and (5-13) are readily written as

$$\begin{aligned} H(t, t') &= T^c \exp\left(-i \int_{t'}^t ds \bar{Q} L(s) \bar{Q}\right) = \begin{bmatrix} 1 & 0 \\ 0 & T^c \exp\left(-i \int_{t'}^t ds L_{22}(s)\right) \end{bmatrix} \\ &= \begin{bmatrix} 1 & 0 \\ 0 & H_{22}(t, t') \end{bmatrix}, \quad t > t' \end{aligned} \quad (5-33a)$$

$$\begin{aligned} K(t; t_0) &= 1 + i \int_{t_0}^t dt' \begin{bmatrix} 1 & 0 \\ 0 & H_{22}(t, t') \end{bmatrix} \begin{bmatrix} 0 & 0 \\ L_{21}(t') & 0 \end{bmatrix} \begin{bmatrix} U_{11}(t', t) & U_{12}(t', t) \\ U_{21}(t', t) & U_{22}(t', t) \end{bmatrix} \\ &= \begin{bmatrix} 1 & 0 \\ i \int_{t_0}^t dt' H_{22}(t, t') L_{21}(t') U_{11}(t', t) & 1 + i \int_{t_0}^t dt' H_{22}(t, t') L_{21}(t') U_{12}(t', t) \end{bmatrix} \\ &= \begin{bmatrix} 1 & 0 \\ K_{21}(t; t_0) & K_{22}(t; t_0) \end{bmatrix}. \end{aligned} \quad (5-33b)$$

We see that $K(t; t_0)$ is invertible if

$$\det K(t; t_0) = \det K_{22}(t; t_0) = \det \left[1 + i \int_{t_0}^t dt' H_{22}(t, t') L_{21}(t') U_{12}(t', t) \right] \neq 0. \quad (5-34)$$

If Eq. (5-34) is fulfilled, from Eq. (5-33b) we obtain

$$K^{-1}(t; t_0) = \begin{bmatrix} 1 & 0 \\ -K_{22}^{-1}(t; t_0)K_{21}(t; t_0) & K_{22}^{-1}(t; t_0) \end{bmatrix}. \quad (5-35)$$

Using the block forms Eqs. (5-28)-(5-35), the equation of motion for $\bar{P}\rho$ (4.16), and the isomorphism Eq. (5-27), we obtain

$$\begin{aligned} \frac{d\rho_S(t)}{dt} = & -i \left[L_{11}(t) - L_{12}(t) K_{22}^{-1}(t; t_0) K_{21}(t; t_0) \right] \rho_S(t) \\ & - i \sqrt{d_E} L_{12}(t) K_{22}^{-1}(t; t_0) H_{22}(t, t_0) \rho_S(t_0). \end{aligned} \quad (5-36)$$

Equation (5-36) is the so-called *partial-trace-free time-convolutionless equation of motion for ρ_S* . It actually describes the evolution of the representation matrix of ρ_S in the basis $\{|\alpha\beta\rangle\}$. Dealing with representation matrices in a given basis, rather than a representation-independent form, is not a downside in itself, particularly if one has numerical implementation in mind. In this partial-trace-free form, multiplication of generally non-square submatrices (e.g., $L_{12}K_{22}^{-1}$) is what mimics the partial trace. Moreover, since no explicit dependence on the projection operator used is present, the partial-trace-free equation (5-36) offers a clearer picture as to what elements of the Hamiltonian h , through the evolution submatrices, actually come into play. This increased transparency will prove useful when choosing and employing different approximations.

5.1.6 “Purely system states” and “entangled states”

We have established that there are two classes of states in \mathcal{H}_{S+E}^2 : those from $(\mathcal{H}_{S+E}^2)_{\bar{P}=1}$, and those from its orthocomplement $(\mathcal{H}_{S+E}^2)_{\bar{P}=0}$. Due to the isomorphism given by Eq. (5-27), which reveals that the states from $(\mathcal{H}_{S+E}^2)_{\bar{P}=1}$ faithfully represent (within the total Liouville space) what goes on with the system, we will call the states from $(\mathcal{H}_{S+E}^2)_{\bar{P}=1}$ *the purely system states*. Also, we see [Eqs. (5-30a,b)] that the upper-left block matrix of any $S+E$ superoperator can be dubbed the “purely system part” of that superoperator. For instance, the upper-left block matrix L_{11} of the Liouvillian is of the commutator-generated form, i.e., it corresponds to an effective system Hamiltonian $h_{S,eff}$ given by

$$h_{S,eff} = h_S + Tr_E(h_{int})/d_E, \quad (5-37)$$

which accounts for the well-known first-order correction to the system energy spectrum [175-177], due to the coupling with the environment (see Appendix C of Ref. [186]).

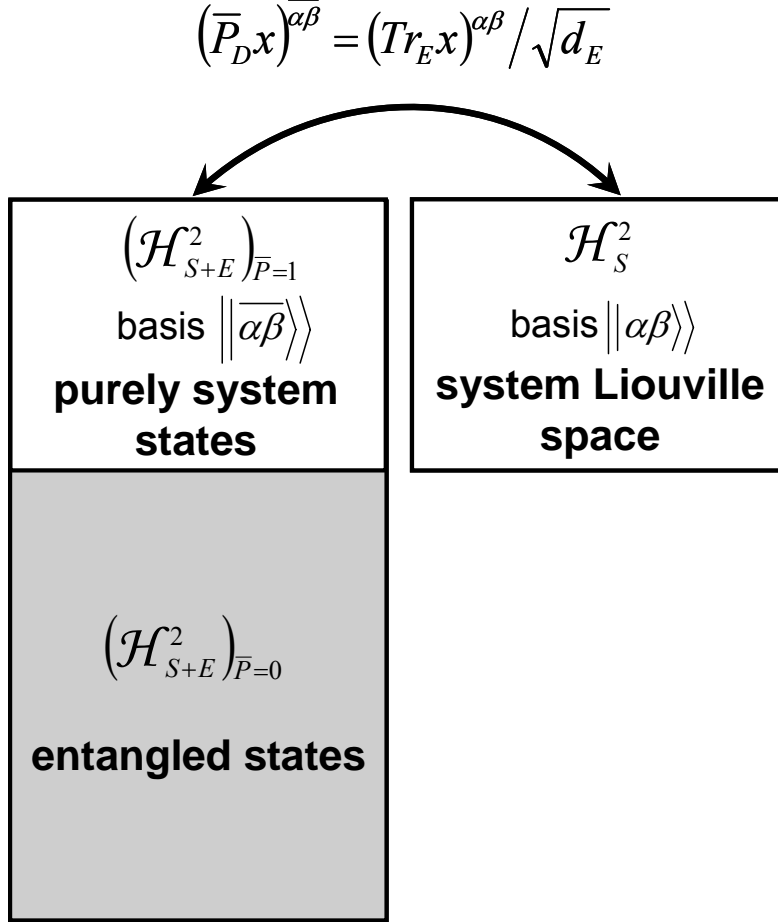


Figure 5-1. Decomposition of the total Liouville space into the eigenspaces of the projection superoperator \bar{P} , and identification of the *purely system states* and the *entangled states*. The purely system states and the states from the system Liouville space are related via the isomorphism given in Eq (5-27).

States from $(\mathcal{H}^2_{S+E})_{\bar{P}=0}$, the orthocomplement to $(\mathcal{H}^2_{S+E})_{\bar{P}=1}$, we call *the entangled states*, in keeping with the general nomenclature of quantum information [187]. The decomposition of \mathcal{H}^2_{S+E} and the isomorphism (5-27) are depicted in Figure 5-1. It is crucial here to clarify what we mean by purely system states and by entangled states. Namely, in quantum information theory, entanglement has a precise definition (see, e.g., Ref. [187]): a composite system ($S+E$) is said to be in a *non-entangled (separable) state*, if its density matrix can be written as a linear combination of tensor products of the subsystems' density matrices (here, density matrices of the system and the environment). Otherwise, the composite system is said to be in an *entangled state*.

In the present approach, the purely system states are the $S+E$ vectors of the form $x = \bar{\rho}_E \otimes Tr_E x$, i.e., those which are of a particular separable form. This form is special because all density matrices that are among the purely system states are of such a form that the *information entropy of the environment in such states is maximal*: in other words, since its density matrix is uniform, *the environment has no information to transmit to the system*. Therefore

$$s(\rho) \leq s(\rho_S) + s(\rho_E) \leq s(\rho_S) + s(\bar{\rho}_E), \quad (5-38)$$

where s denotes the information (von-Neumann) entropy. Equation (5-38) means that, among all the $S+E$ states ρ that yield a given reduced density matrix ρ_S , the purely system state corresponding to that ρ_S is the one with the largest entropy [188]. So, we can think of the purely system states as *the states depleted*

of environmental information. In a similar fashion, what we call the entangled states should alternatively be thought of as the *states rich in environmental information*. In that sense, since the information exchange between the system and the environment is another way of (loosely) stating that the system and environmental states become entangled, we believe that choosing the name “entangled states” for the vectors in $(\mathcal{H}_{S+E}^2)_{\bar{P}=1}$ is fairly appropriate, and certainly rolls off the tongue better than, for instance, “states rich in environmental information”. We are, however, aware that the name “entangled states” can be considered as somewhat of a misnomer in the present context.

5.1.7 Memory dressing and the reduced density matrix

Within the partial-trace-free approach, i.e., in the eigenbasis of \bar{P} , the equations of motion for the projections ρ_1 and ρ_2 of the density matrix [see Eqs. (5-16) and (5-15), respectively], can be written as

$$\begin{aligned} \frac{d\rho_1(t)}{dt} &= -i \left[L_{11}(t) - L_{12}(t) K_{22}^{-1}(t; t_0) K_{21}(t; t_0) \right] \rho_1(t) - i L_{12}(t) K_{22}^{-1}(t; t_0) H_{22}(t, t_0) \rho_2(t_0), \\ \rho_2(t) &= -K_{22}^{-1}(t; t_0) K_{21}(t; t_0) \rho_1(t) + K_{22}^{-1}(t; t_0) H_{22}(t, t_0) \rho_2(t_0), \end{aligned} \quad (5-39)$$

where H_{22} , K_{21} and K_{22} , defined in Eqs. (5-33a,b), also satisfy the following equations with the corresponding initial conditions

$$\begin{aligned} H_{22}(t, t_0) &= T^c \exp \left(-i \int_{t_0}^t d\tau L_{22}(\tau) \right), \\ \frac{dK_{21}(t; t_0)}{dt} &= -i L_{22}(t) K_{21}(t; t_0) + i K_{21}(t; t_0) L_{11}(t) + i K_{22}(t; t_0) L_{21}(t), \\ \frac{dK_{22}(t; t_0)}{dt} &= -i L_{22}(t) K_{22}(t; t_0) + i K_{22}(t; t_0) L_{22}(t) + i K_{21}(t; t_0) L_{12}(t), \\ K_{21}(t_0; t_0) &= 0, \quad K_{22}(t_0; t_0) = 1. \end{aligned} \quad (5-40)$$

As before, the second time variable in the argument of submatrices $K_{12}(t; t_0)$ and $K_{22}(t; t_0)$ is not exactly a variable. Namely, t_0 denotes the fixed initial time, which defines the initial conditions for K 's, and is separated from the first time variable t by a semi-colon instead of a comma.

According to Eq. (5-39), the equations of motion for the evolution submatrices [see Eqs. (5-31b) and (5-32b)], with the second time argument fixed to t_0 , are given by

$$\begin{aligned} \frac{dU_{11}(t, t_0)}{dt} &= -i \left[L_{11}(t) - L_{12}(t) K_{22}^{-1}(t; t_0) K_{21}(t; t_0) \right] U_{11}(t, t_0), \\ \frac{dU_{12}(t, t_0)}{dt} &= -i \left[L_{11}(t) - L_{12}(t) K_{22}^{-1}(t; t_0) K_{21}(t; t_0) \right] U_{12}(t, t_0) \\ &\quad - i L_{12}(t) K_{22}^{-1}(t; t_0) H_{22}(t, t_0), \\ U_{21}(t, t_0) &= -K_{22}^{-1}(t; t_0) K_{21}(t; t_0) U_{11}(t, t_0), \\ U_{22}(t, t_0) &= K_{22}^{-1}(t; t_0) \left[H_{22}(t, t_0) - K_{21}(t; t_0) U_{12}(t, t_0) \right]. \end{aligned} \quad (5-41)$$

The above are generic time-convolutionless equations of motion, and are generally well known (maybe not necessarily in this particular form, which is valid for any choice of the projection operator as long as we are writing the equations in the projection operator's eigenbasis). Note that U_{21} and U_{22} can be written in terms of U_{11} and U_{12} , which is a typical feature of time-convolutionless approaches. If we are to solve equations (5-41), we first have to solve the very unpleasant equations (5-40), among which the equations for K_{21} and K_{22} are coupled. Even if we were to solve (5-40), in order to obtain a solution to (5-41) we need to perform an additional unappealing task of inverting K_{22} . With the increasing dimension of the $S+E$ Hilbert spaces, K_{22} becomes large [the largest matrices we deal with are H_{22}, K_{22}, U_{22} , of dimension $d_S^2(d_E^2 - 1) \times d_S^2(d_E^2 - 1)$; needless to say, one would not want to unnecessarily invert one of them].

Luckily, there is a way out of this unpleasant situation. If we examine equations (5-41) more closely, we will notice that we do not actually need all three matrices H_{22}, K_{21}, K_{22} from Eqs. (5-40). The quantities we do need for the equations of motion (5-41) are actually

$$\begin{aligned} R(t) &= K_{22}^{-1}(t; t_0) K_{21}(t; t_0), \\ S(t; t_0) &= K_{22}^{-1}(t; t_0) H_{22}(t, t_0). \end{aligned} \quad (5-42)$$

Notice that we have omitted the time quasi-variable t_0 in defining R . It does not mean that this information is not present or important; on the contrary, choosing t_0 defines the initial condition for R . But, it will become apparent below why keeping R with one time-variable only is in the service of transparency and intuitive plausibility.

It is easily shown, by using equations (5-40), that R and S satisfy the following equations of motion:

$$\begin{aligned} \frac{dR(t)}{dt} &= -iL_{22}(t)R(t) - iR(t)L_{12}(t)R(t) + iR(t)L_{11}(t) + iL_{21}(t), \quad R(t_0) = 0; \\ \frac{dS(t; t_0)}{dt} &= -i[L_{22}(t) + iR(t)L_{12}(t)]S(t; t_0), \quad S(t_0; t_0) = 1. \end{aligned} \quad (5-43)$$

Therefore, if one is to solve for the evolution of the reduced density matrix ρ_S , starting from a given initial $S+E$ density matrix $\rho(t_0) = [\rho_1(t_0) \quad \rho_2(t_0)]^T$, one only needs to solve the following equations (with corresponding initial conditions)

$$\begin{aligned} \frac{dR(t)}{dt} &= -iL_{22}(t)R(t) - iR(t)L_{12}(t)R(t) + iR(t)L_{11}(t) + iL_{21}(t), \quad R(t_0) = 0; \\ \frac{dS(t; t_0)}{dt} &= -i[L_{22}(t) + R(t)L_{12}(t)]S(t; t_0), \quad S(t_0; t_0) = 1; \\ \frac{dU_{11}(t, t_0)}{dt} &= -i[L_{11}(t) - L_{12}(t)R(t)]U_{11}(t, t_0), \quad U_{11}(t_0, t_0) = 1; \\ \frac{dU_{12}(t, t_0)}{dt} &= -i[L_{11}(t) - L_{12}(t)R(t)]U_{12}(t, t_0) - iL_{12}(t)S(t; t_0), \quad U_{12}(t_0, t_0) = 0. \end{aligned} \quad (5-44)$$

(Again, these equations hold for any projection operator, provided that we are writing equations in that projection operator's eigenbasis.) Also, the remaining two submatrices evolve according to

$$U_{21}(t, t_0) = -R(t)U_{11}(t, t_0), \quad U_{22}(t, t_0) = S(t; t_0) - R(t)U_{12}(t, t_0),$$

but we have no interest in them at present. Their evolution is of interest when calculating, for instance, two-time correlation functions in electronic transport [189], which requires $U(t, t')$, $t' \neq t_0$. We are interested here only in deducing how the evolution of the reduced density matrix proceeds after t_0 , for which we need the initial $S+E$ density matrix $\rho(t_0)$, and submatrices $U_{11}(t, t_0)$ and $U_{12}(t, t_0)$.

At this point, it is useful to note several features of equations (5-44): First, R has a *self-contained nonlinear equation of motion*, which should obviously be the starting point of a calculation. This equation is a matrix Riccati equation, whose general properties and solutions are the subject of active research, especially in control systems theory [190]. We will solve this equation of motion below. Secondly, evolution of U_{11} obeys a Liouville-like equation, with a generally non-Hermitian “quasi-Liouvillian” $L_{11}(t) - L_{12}(t)R(t)$, with $L_{12}R$ obviously playing the role of an effective, memory-containing interaction. Since we have already identified L_{12} as the term accounting for the real physical interaction, it is then clear that $R(t)$ plays the role of a *memory dressing* of the interaction, and this is what we will call it henceforth. In the evolution of U_{12} , we see an effective driving term, $-iL_{12}S$, in addition to the quasi-Liouvillian $L_{11}(t) - L_{12}(t)R(t)$.

(A) Evaluation of the memory dressing $R(t)$

Since the memory dressing R apparently plays an important role in calculation of the evolution operators time dependence, we will now solve its equation of motion

$$\frac{dR}{dt} = -iL_{22}R - iRL_{12}R + iRL_{11} + iL_{21}. \quad (5-45)$$

(We will omit the time arguments for brevity). Regardless of the interaction strength (i.e., the magnitude of L_{12}), this equation is complemented by the initial condition $R(t_0) = 0$. We will thus solve it by forming a power-expansion in terms of L_{12} (or L_{21} , which is the adjoint of L_{12}), i.e., by assuming the form

$$R(t) = \sum_{n=0}^{\infty} R^{(n)}(t), \quad R^{(n)} = (L_{12})^n \times \text{function of } (L_{11}, L_{22}). \quad (5-46a)$$

(The right-hand-side equation of (5-46a) is a symbolic expression, because L_{12} is not a square matrix, so its powers are not defined. Rather, $R^{(n)}$ is a product of generally non-square matrices, so that L_{12} and its adjoint appear a total of n times in the product, whereas the other matrices in the product depend only on L_{11} and/or L_{22} .) In addition, for all n , the following initial condition holds:

$$R^{(n)}(t_0) = 0, \quad n \geq 0. \quad (5-46b)$$

For the zeroth order term, the equation of motion is obtained as

$$\frac{dR^{(0)}}{dt} = -iL_{22}R^{(0)} + iR^{(0)}L_{11}. \quad (5-47a)$$

A general solution of Eq. (5-47a), for times $t \geq t_0$ (which are the only times meaningful to investigate, as we do not have sufficient knowledge about the system prior to t_0), is given by

$$R^{(0)}(t) = T^c \exp\left(-i \int_{t_0}^t d\tau L_{22}(\tau)\right) R^{(0)}(t_0) T^a \exp\left(i \int_{t_0}^t d\tau L_{11}(\tau)\right). \quad (5-47b)$$

Since, $R^{(0)}(t_0) = 0$, for all times t we have

$$R^{(0)}(t) = 0. \quad (5-47c)$$

The equation of motion for the first-order term $R^{(1)}$ reads

$$\frac{dR^{(1)}}{dt} = -iL_{22}R^{(1)} + iR^{(1)}L_{11} + iL_{21}. \quad (5-48a)$$

This is an inhomogeneous linear equation, with the general solution

$$\begin{aligned} R^{(1)}(t) &= T^c \exp\left(-i \int_{t_0}^t d\tau L_{22}(\tau)\right) \\ &\times \left[R^{(1)}(t_0) + \int_{t_0}^t d\tau T^a \exp\left(i \int_{t_0}^{\tau} d\tau' L_{22}(\tau')\right) [iL_{21}(\tau)] T^c \exp\left(-i \int_{t_0}^{\tau} d\tau' L_{11}(\tau')\right) \right] \\ &\times T^a \exp\left(i \int_{t_0}^t d\tau L_{11}(\tau)\right). \end{aligned} \quad (5-48b)$$

With the initial condition $R^{(1)}(t_0) = 0$, we obtain

$$R^{(1)}(t) = \int_{t_0}^t d\tau T^c \exp\left(-i \int_{\tau}^t d\tau' L_{22}(\tau')\right) [iL_{21}(\tau)] T^a \exp\left(i \int_{\tau}^t d\tau' L_{11}(\tau')\right). \quad (5-48c)$$

At this point it is useful to recall that, in the absence of coupling (i.e., $L_{12} = 0$), the off-diagonal evolution submatrices become zero ($U_{21}^{(0)} = 0, U_{12}^{(0)} = 0$). Moreover, for $t, t' > t_0$, we obtain the diagonal evolution submatrices in the form:

$$\begin{aligned} U_{11}^{(0)}(t, t') &= \Theta(t - t') T^c \exp\left(-i \int_{t'}^t d\tau L_{11}(\tau)\right) + \Theta(t' - t) T^a \exp\left(i \int_t^{t'} d\tau L_{11}(\tau)\right), \\ U_{22}^{(0)}(t, t') &= \Theta(t - t') T^c \exp\left(-i \int_{t'}^t d\tau L_{22}(\tau)\right) + \Theta(t' - t) T^a \exp\left(i \int_t^{t'} d\tau L_{22}(\tau)\right). \end{aligned} \quad (5-49)$$

(As before, Θ denotes the Heaviside step function.) Each of the submatrices $U_{11}^{(0)}, U_{22}^{(0)}$ is unitary, and these evolution submatrices, for purely system states and entangled states separately, form groups [note that operator $H_{22}(t, t_0)$ from Eq. (5-40) is actually just $U_{22}^{(0)}(t, t_0)$]. We can now rewrite (5-48c) as

$$R^{(1)}(t) = \int_{t_0}^t d\tau U_{22}^{(0)}(t, \tau) [iL_{21}(\tau)] U_{11}^{(0)}(\tau, t), \quad (5-50)$$

which is the form we will use frequently later. It was important to evaluate the first-order-correction explicitly, as the only free term (i.e., term not containing R) in Eq. (5-45) is of the first order in L_{12} [the

iL_{21} term on the right-hand-side of Eq. (5-45)]. For all higher orders n , $n > 1$, it is straightforward to prove, using Eq. 5-45), that

$$\frac{dR^{(n)}}{dt} = -iL_{22}R^{(n)} - i\sum_{k=0}^{n-1} R^{(k)}L_{12}R^{(n-1-k)} + iR^{(n)}L_{11}. \quad (5-51)$$

Lemma 1: All even-order terms in R are zero, i.e.,

$$R^{(2k)}(t) = 0, \quad k \geq 0. \quad (5-52)$$

The detailed proof is given in Ref. [191].

Lemma 2: Odd-order terms, $R^{(2k+1)}$, $k \geq 1$, can be calculated according to

$$\begin{aligned} R^{(2k+1)}(t) = & \int_{t_0}^t d\tau_1 \int_{t_0}^{\tau_1} d\tau_2 \dots \int_{t_0}^{\tau_k} d\tau_k \mathcal{P}(t, \tau_1) [-iL_{12}(\tau_1)] \mathcal{P}(\tau_1, \tau_2) [-iL_{12}(\tau_2)] \mathcal{P}(\tau_2, \tau_3) \dots \\ & \times \dots \mathcal{P}(\tau_{k-1}, \tau_k) [-iL_{12}(\tau_k)] \mathcal{P}(\tau_k, t) \end{aligned} \quad (5-53)$$

where

$$\mathcal{R}(t) = R^{(1)}(t) = \int_{t_0}^t d\tau U_{22}^{(0)}(t, \tau) [iL_{21}(\tau)] U_{11}^{(0)}(\tau, t), \quad t \geq t_0, \quad (5-54a)$$

$$\mathcal{P}(\tau, \tau') = \begin{cases} \tau > \tau', & U_{22}^{(0)}(\tau, \tau') \mathcal{R}(\tau') \\ \tau < \tau', & \mathcal{R}(\tau) U_{11}^{(0)}(\tau, \tau') \end{cases}, \quad \tau, \tau' \geq t_0. \quad (5-54b)$$

The quantity \mathcal{P} is called *the information-exchange propagator*. The detailed proof is given also in Ref. [191].

Theorem: The solution of equation (4.45), with the initial condition $R(t_0) = 0$, is given by

$$\begin{aligned} R(t) &= \sum_{k=0}^{\infty} R^{(2k+1)}(t) \\ &= \mathcal{R}(t) + \sum_{k=1}^{\infty} \int_{t_0}^t d\tau_1 \dots \int_{t_0}^{\tau_k} d\tau_k \mathcal{P}(t, \tau_1) [-iL_{12}(\tau_1)] \mathcal{P} \dots \mathcal{P}(\tau_{k-1}, \tau_k) [-iL_{12}(\tau_k)] \mathcal{P}(\tau_k, t) \end{aligned} \quad (5-54c)$$

The proof follows a straightforward implementation of Lemmas 1 and 2.

Now that we have formally solved the equation of motion for the memory dressing R , and introduced the information-exchange propagator \mathcal{P} , let us try to develop an intuitive feeling of what these results actually mean. In Figure 5-2, $\mathcal{R}(t)$ is depicted. We see that $\mathcal{R}(t)$ calculates a cumulative effect that the purely system states have on the entangled states, provided that at any given time the information exchange is due to the first order in coupling. Note that, to obtain $\mathcal{R}(t)$, we go backwards in time among the purely system states (which evolve as if there were no coupling, according to $U_{11}^{(0)}$), then the interaction occurs, and afterwards the entangled states evolve forward in time, again as if there were no coupling.

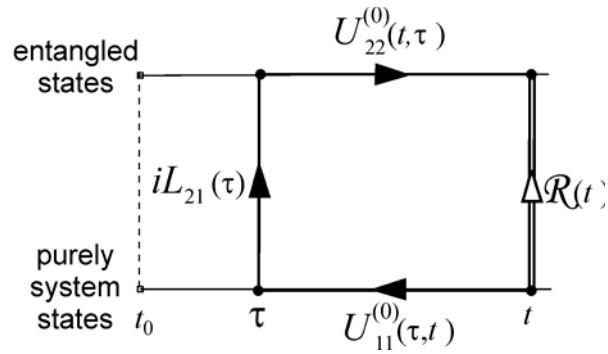


Figure 5-2. Diagrammatic representation of $\mathcal{R}(t) \equiv R^{(1)}(t)$, the first-order contribution to the memory dressing.

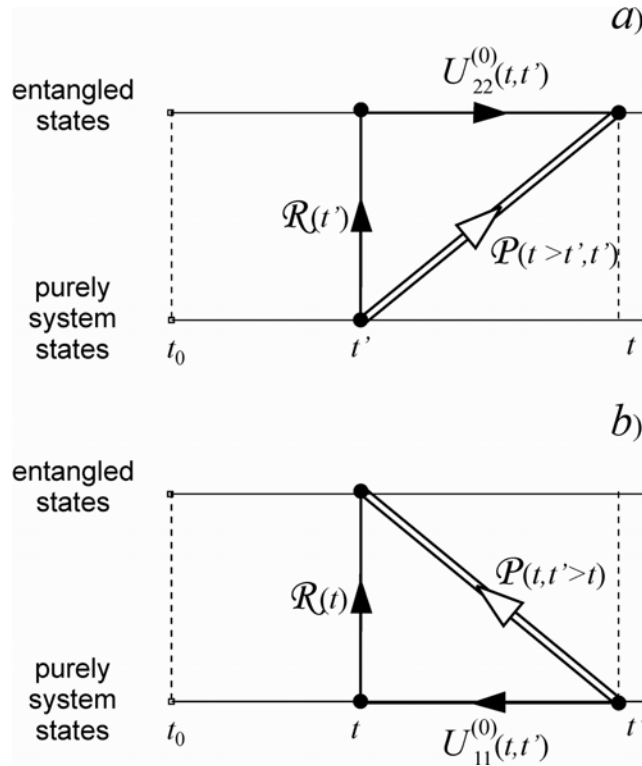


Figure 5-3. The information-exchange propagator $\mathcal{P}(t, t')$ for different orderings of t and t' . a) $t > t'$. b) $t' > t$.

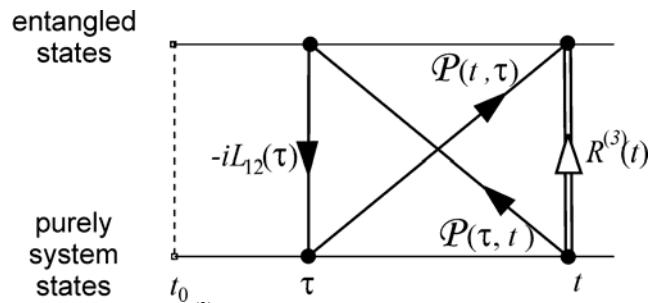


Figure 5-4. Diagrammatic representation of $R^{(3)}(t)$, the third-order contribution to the memory dressing.

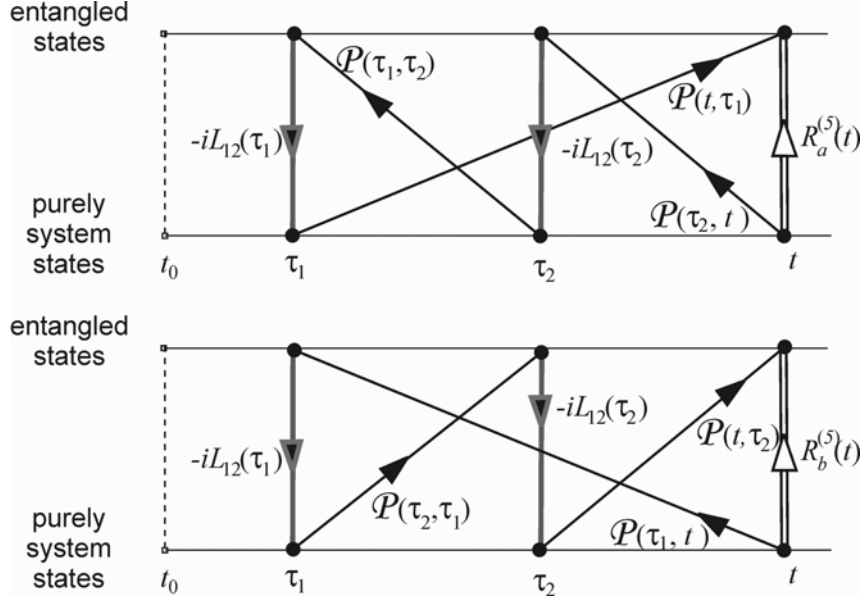


Figure 5-5. Diagrammatic representation of the two topologically nonequivalent contributions to $R^{(5)}(t)$, the fifth-order term in the memory dressing.

Figure 5-3 presents a diagrammatic representation of $\mathcal{P}(t, t')$ for two different orderings of t and t' . Note how $\mathcal{R}(t)$ is always present at the later of the times. In Figure 5-4, the third-order correction to the memory dressing R is depicted. Note how, if one uses the result for $\mathcal{P}(t, t')$ depicted in Figure 5-3, we see that this third-order correction contains a cumulative effect that the purely system states have on the entangled states, provided that at any given time the information exchange is due to the third order in coupling ($iL_{21}(\tau)L_{12}(\tau)L_{21}(\tau)$ at time τ in the figure.) In Figure 5-5, we see two topologically nonequivalent contributions to the fifth-order correction.

5.1.8 Short-time evolution in the case of initially uncorrelated system and environment

With the inclusion of the memory dressing, we have now managed to obtain the evolution in the form

$$\frac{d\rho_S}{dt} = -i[L_{11} - L_{12}R]\rho_S - i\sqrt{d_E}L_{12}S(t; t_0)\rho_2(t_0)$$

In the case when the system and the environment start from an initially uncorrelated state, i.e., the initial density matrix is of the form $\rho(t_0) = \rho_E(t_0) \otimes \rho_S(t_0)$, it is possible to completely reduce the problem to subspace 1, namely it is possible to write

$$\rho_2(t_0) = M\rho_1(t_0) = (d_E)^{-1/2}M\rho_S(t_0), \quad (5-56)$$

where the mapping M is completely determined by the components of $\rho_E(t_0)$, the initial environmental density matrix [192]. The above equation embodies the argument made by Lindblad [193] that a subdynamics exists only for an initially uncorrelated state, as it enables one to write

$$\frac{d\rho_S}{dt} = -i[L_{11} - L_{12}R]\rho_S - iL_{12}S(t, t_0)M\rho_S(t_0) \quad (5-57a)$$

or in the integral form

$$\rho_S(t) = [U_{11}(t, t_0) + U_{12}(t, t_0)M]\rho_S(t_0). \quad (5-57b)$$

Without the loss of generality, we can assume that there exists a time-dependent generator of the above map, i.e., that the above map can be rewritten in the form

$$\rho_S(t) = T^c \exp \left[-i \int_{t_0}^t L_{eff}(t) dt - \Gamma(t) \right] \rho_S(t_0), \quad (5-58)$$

where $L_{eff}(t)$ is an undetermined effective Liouvillian and $\Gamma(t)$ is the decoherence exponent. In the Markovian approximation, $-iL_{eff} - \frac{d\Gamma}{dt} = const.$ is of the well-known Lindblad form [194] and generates a semigroup of completely positive maps. Complete positivity of a dynamical map means that the unit trace and positivity of the density matrix are preserved at all times, but also that the evolution is robust enough to support tensor products - for example, two noninteracting systems evolving according to completely positive maps can together be described by a tensor product of those maps, itself a completely positive map.

When we perform the short-time Taylor expansion of the exact equation (5-58) up to the second order in time around $t_0 = 0$, we obtain an approximate completely positive map in the form

$$\rho_S(t) = \exp[-L_{eff}t - \Lambda t^2] \rho_S(0) + o(t^3), \quad (5-59)$$

where

$$L_{eff} = [h_S + \langle h_{int} \rangle, \dots] = L_S + [\langle h_{int} \rangle, \dots] \quad (5-60a)$$

and Λ is a superoperator acting on the system Liouville space, whose matrix elements in a basis $\alpha\beta$ in this space are given by

$$\begin{aligned} \Lambda_{\alpha\beta}^{\alpha\beta} = & \frac{1}{2} \left\{ \langle h_{int}^2 \rangle_{\alpha'}^{\alpha} \delta_{\beta}^{\beta'} + \langle h_{int}^2 \rangle_{\beta}^{\beta'} \delta_{\alpha'}^{\alpha} - 2 \sum_{i,j} (h_{int})_{j\alpha'}^i \rho_E^j (h_{int})_{j\beta}^{i\beta'} \right\} \\ & - \frac{1}{2} \left\{ \langle h_{int} \rangle_{\alpha'}^{\alpha} \delta_{\beta}^{\beta'} + \langle h_{int} \rangle_{\beta}^{\beta'} \delta_{\alpha'}^{\alpha} - 2 \langle h_{int} \rangle_{\alpha'}^{\alpha} \langle h_{int} \rangle_{\beta}^{\beta'} \right\}. \end{aligned} \quad (5-60b)$$

Operator Λ contains essential information on the directions of coherence loss in both non-Markovian and Markovian systems [192].

5.1.9 Coarse-grained Markovian evolution

There are generally two rigorous ways to obtain Markovian evolution from non-Markovian ones: one is the weak-coupling limit in conjunction with the long-time limit on appropriately rescaled timescale (the so-called van Hove limit, where the time $t \rightarrow \infty$ and the interaction $\|h_{int}\| \rightarrow 0$, so that the product $\|h_{int}\|^2 t$ is constant and finite); the other is the so-called singular coupling limit, which basically embodies the fact that the environment has a mechanism through which its internal state can be reset after a short, finite time (this resetting mechanism basically means that SE is not truly closed, but open as well, so there is an outermost environment responsible for resetting of environment's state, but with which our system does not interact directly). The notion of the environmental memory being reset is at the heart of what we can do to construct Markovian maps for nanoscale electronic systems, because we generally know that dephasing in the leads (also known as contacts or reservoirs) typically occurs on the timescales of at most a few hundred femtoseconds [195], even at low temperatures, due to efficient electron-electron scattering. Electron-electron scattering forces the distribution function in the leads back into a (possibly displaced) Fermi-Dirac distribution function; so if we are not interested in resolving timescales shorter than about a picosecond, the Markovian approximation should be fine due to this built-in resetting mechanism.

So how do we obtain this Markovian map from the exact non-Markovian one? One approach would be to coarse-grain the exact evolution over an effective memory retention time τ of the environment (the energy relaxation time in the bulk-like contacts). This approach is satisfactory in cases when we care the most about describing the nonequilibrium steady state, and not necessarily how long it exactly takes to reach this state. So the coarse-grained Markovian map is derived as [201,192]

$$\frac{d\rho_s}{dt} = (-iL_{eff} - \Lambda\tau)\rho_s. \quad (5-61)$$

Clearly, the steady state must belong to the common null-space of Λ and L_{eff} . Next, on the example of a resonant-tunneling diode, we demonstrate how the above Markovian map straightforwardly yields the well-known I-V curve that manifests the prominent negative differential resistance features.

5.2 Decoherence in the active region of a resonant-tunneling diode

A resonant-tunneling diode (RTD) is a two-barrier tunneling structure, fabricated on a high mobility two-dimensional (2D) electron gas. Typical dimensions of both the barriers and the well range from 2 to 5 nanometers. The barriers are usually AlGaAs, while the well is GaAs, all grown by molecular-beam epitaxy. The interaction between the active region and the contacts has so far been addressed within the model of the RTD as a two-level system, where the electron can either be in the RTD bound state (“+” state) or absent from the RTD (“-” state) [196]. However, since the RTD naturally has open boundaries and continuous spectrum, the two-level model must usually be supplemented with a phenomenological resonance width to account for the fact that quite a few plane waves from the contacts contribute to the making of the bound state. Moreover, the continuum-state increase after the valley cannot be captured by employing a two-level model for the RTD. The openness of the RTD active region has also been addressed in the works of Frenley [37] and Pötz [149, 197], as well as in the works of the Purdue group [198,199,200] where the contacts are accounted for through a special injection self-energy term.

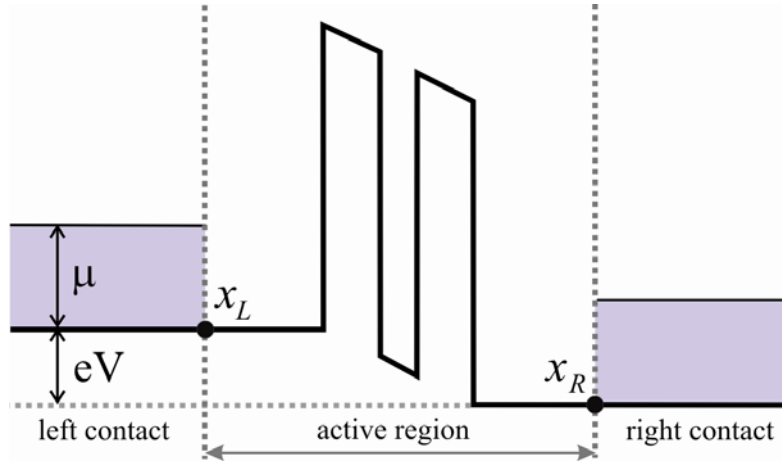


Figure 5-6: Schematic of a resonant-tunneling diode under bias $-V$ applied to the left contact.

A schematic of the simulated RTD under bias is given in Figure 5-6. The RTD well width is 3 nm, each barrier's thickness is 5 nm, and the barrier height is 0.3 eV. The Fermi level in each contact is at 0.1 eV, below the equilibrium bound state, which is at 0.15 eV. A negative bias V is applied to the left contact. While still focusing on the active region/contacts interaction, we will offer an RTD model different from those previously published. Any given energy $E_k = \hbar^2 k^2 / 2m^*$ above the bottom of the higher, left contact, is a doubly degenerate eigenvalue of the single particle Hamiltonian of the RTD; to this value correspond a forward propagating solution Ψ_k and a backward propagating solution Ψ_{-k} . These functions are not plane waves, but rather correspond to the plane waves being injected from the left and the right contacts, respectively, and have the following behavior to the far left (x_L , the left boundary of the active region) and right (x_R , the right boundary of the active region) of the heterostructure:

$$\begin{aligned}
\Psi_k(x \rightarrow x_L) &= e^{ikx} + r_{-k,L} e^{-ikx}, \\
\Psi_k(x \rightarrow x_R) &= t_{k',R} e^{ik'x}, \\
\Psi_{-k}(x \rightarrow x_L) &= t_{-k,L} e^{-ikx}, \\
\Psi_{-k}(x \rightarrow x_R) &= e^{-ik'x} + r_{k',R} e^{ik'x},
\end{aligned} \tag{5-62}$$

where k and k' are the wavevectors in the left and right contact, respectively, that correspond to the same energy: $k^2 = 2m^* E_k / \hbar^2 = k'^2 - 2m^* eV / \hbar^2$. We will use k and $-k$ associated with this energy to label the forward and backward propagating states, respectively, even though the backward propagating states are actually injected with $-k'$. The subscripts in the transmission and reflection coefficient denote the final wavevector and contact (for instance, the reflected wave that originally incident from the left with k , has the wavevector $-k$ and remains in the left contact, thus notation $r_{-k,L}$). What is important is that we adopt a potential profile (or, in numerical calculation, couple with a Poisson solver) that will enable us to solve for Ψ_k and Ψ_{-k} and find the transmission and reflection coefficients of the forward and backward propagating plane waves. Here, for simplicity, a slinear potential drop across the well and barriers is adopted, but in general a self-consistent solution (described in more detail below) should be sought. Associated with Ψ_k (Ψ_{-k}) in the active region are the creation and destruction operators d_k^+ and d_k (d_{-k}^+ and d_{-k}), so that the active region Hamiltonian is described by

$$h_S = \sum \omega_k (d_k^+ d_k + d_{-k}^+ d_{-k}), \tag{5-63}$$

where $\omega_k = E_k / \hbar$ (all Hamiltonians and Liouvillians are defined in the units of frequency). Moreover, the discussion in this section disregards the spin quantum number, which only adds to the degeneracy; the resulting currents at the end of this section should be considered “per spin orientation”.

The active region communicates with the reservoirs of charge - the contacts. We introduce a model interaction for the coupling between the eigenfunctions Ψ_k from the active region and the forward propagating plane waves, injected from the left contact, as well as the resulting transmitted and reflected waves:

$$h_{\text{int},+} = \sum_k \Delta_k \left[(c_{k,L}^+ + R_k c_{-k,L}^+) d_k + d_k^+ (c_{k,L} + r_{-k,L} c_{-k,L}) + T_k (c_{k',R}^+ d_k + d_k^+ c_{k',R}) \right]. \tag{5-64a}$$

$c_{k,L}^+$ ($c_{k,L}$) and $c_{k',R}^+$ ($c_{k',R}$) create (destroy) an electron with a wavevector k in the left and k' in the right contact, respectively, while R_k and T_k are the reflection and transmission coefficient at a given energy, satisfying $R_k + T_k = 1$. The model interaction captures injection from the contacts: namely, the coupling coefficient Δ_k in (5-64a) is the rate of injection of carriers with momentum k from the left contact into the active region i.e., it is proportional to the current injected into the state Ψ_k by the hopping of one electron:

$\Delta_k = \hbar k / m \|\Psi_k\|^2$, where $\|\Psi_k\|^2 = \int_0^W dx |\Psi_k(x)|^2$ is the norm squared of Ψ_k over the active region of width W . Similarly, the coefficient multiplying $c_{-k,L}^+$ is the rate of reflection, proportional to the reflected current and thus equal to $R_k \Delta_k$ while the coefficient multiplying $c_{k',R}^+$ is the rate of transmission, equal to $T_k \Delta_k$. The Hamiltonian for the backward propagating states can be written in an analogous fashion

$$h_{\text{int},-} = \sum_k \Delta_{-k} \left[(c_{-k',R}^+ + R_{-k} c_{k',R}^+) d_{-k} + T_{-k} c_{-k,L}^+ d_{-k} \right] + h.c. \tag{5-64b}$$

Here, $\Delta_{-k} = \hbar k' / m \|\Psi_{-k}\|^2$ and $T_{-k} = T_k$, $R_{-k} = R_k$. Furthermore, we will assume that bias is swept slowly (so that between two bias points the system is allowed to relax) and in small increments (so that the

transmission and reflection coefficients and the barrier profile do not change much between two bias points, and can be regarded constant during each transient).

When we put it all together, we have for the interaction Hamiltonian of the active region with the left/right contact:

$$\begin{aligned} h_{\text{int},L} &= \sum_k \Delta_k \left[(c_{k,L}^+ + R_{-k} c_{-k,L}^+) d_k + d_k^+ (c_{k,L} + R_{-k} c_{-k,L}) \right] + \Delta_{-k} T_{-k} (c_{-k,L}^+ d_{-k} + d_{-k}^+ c_{-k,L}), \\ h_{\text{int},R} &= \sum_k \Delta_{-k} \left[(c_{-k,R}^+ + R_{-k} c_{k,R}^+) d_{-k} + d_{-k}^+ (c_{-k,R} + R_{-k} c_{k,R}) \right] + \Delta_k T_k (c_{k,R}^+ d_k + d_k^+ c_{k,R}). \end{aligned} \quad (5-65)$$

Since all the elements of the interaction Hamiltonians are linear in the contact creation/destruction operators, and we assume that the initial states of the contacts are thermal equilibrium grand-canonical ensembles, $\langle h_{\text{int},L/R} \rangle = 0$. This means that $L_S = L_{\text{eff}}$ and only leaves us with the first three terms in the equation for Λ to calculate. One can show that $\Lambda = \Lambda_L + \Lambda_R$, where

$$(\Lambda_{L/R})_{\alpha\beta}^{\alpha\beta} = \frac{1}{2} \left(\langle h_{\text{int},L/R}^2 \rangle_{\alpha'}^{\alpha} \delta_{\beta}^{\beta'} + \langle h_{\text{int},L/R}^2 \rangle_{\beta'}^{\beta} \delta_{\alpha}^{\alpha'} - 2 \sum_{i,j} \rho_{L/R}^i (h_{\text{int},L/R})_{i\alpha'}^{j\alpha} (h_{\text{int},L/R})_{ij\beta}^{\beta'} \right). \quad (5-66)$$

The first and the second terms give general contributions of the form $\Lambda_{\alpha\beta}^{\alpha\beta}$, since $\langle h_{\text{int},L/R}^2 \rangle$ preserves the filling of states. The third term gives a contribution of the form $\Lambda_{\beta\beta}^{\alpha\alpha}$ [201]. Each term in Λ attacks only single-particle states with a given k , so in reality we have a multitude of two-level problems [201], where the two levels are a particle being in Ψ_k (“+”) and the particle being absent from Ψ_k (“-”). In each of the four-dimensional Liouville spaces associated with each of the two-level problems, the relaxation occurs according to

$$\frac{d\rho_k}{dt} = (-iL_{S,k} - \Lambda_k \tau) \rho_k, \quad (5-67)$$

where $\rho_k = [\rho_k^{++}, \rho_k^{+-}, \rho_k^{-+}, \rho_k^{--}]^T$ is the component of the reduced density matrix that describes the occupation of Ψ_k , and

$$L_{S,k} = \begin{bmatrix} 0 & 0 & 0 & 0 \\ 0 & 2\omega_k & 0 & 0 \\ 0 & 0 & -2\omega_k & 0 \\ 0 & 0 & 0 & 0 \end{bmatrix}, \quad \Lambda_k = \begin{bmatrix} A_k & 0 & 0 & -B_k \\ 0 & C_k & 0 & 0 \\ 0 & 0 & C\omega_k & 0 \\ -A_k & 0 & 0 & B_k \end{bmatrix}, \quad (5-68)$$

and

$$\begin{aligned} A_k &= \Delta_k^2 \left[(1 + R_k^2)(1 - f_k^L) + T_k^2(1 - f_{k'}^R) \right], \\ B_k &= \Delta_k^2 \left[(1 + R_k^2)f_k^L + T_k^2 f_{k'}^R \right], \\ C_k &= \Delta_k^2 (1 + R_k^2 + T_k^2)/2. \end{aligned} \quad (5-69)$$

Here, f_k^L and $f_{k'}^R$ are the Fermi distribution functions in the left and right contact, respectively, corresponding to the energy E_k with respect to the conduction band bottom in the left contact (and $E_k + eV$ with respect to the conduction band bottom in the right contact). Clearly, the off-diagonal elements of the density matrix are zero in the steady state. The equations for the diagonal components are coupled as

$$\begin{aligned} \frac{d\rho_k^{++}}{dt} &= -\tau A_k \rho_k^{++} + \tau B_k \rho_k^{--}, \\ \frac{d\rho_k^{--}}{dt} &= \tau A_k \rho_k^{++} - \tau B_k \rho_k^{--} \end{aligned} \quad (5-70)$$

Clearly, $\frac{d\rho_k^{++}}{dt} + \frac{d\rho_k^{--}}{dt} = 0$, as it should be, because $\rho_k^{++} = f_k$ and $\rho_k^{--} = 1 - f_k$, where f_k is the distribution function for the state Ψ_k of the active region. The two equations (5-70) are basically one and the same, yielding

$$\begin{aligned}\frac{df_k}{dt} &= -\tau A_k f_k + \tau B_k (1 - f_k) = -\tau(A_k + B_k)f_k + \tau B_k, \\ f_k(t) &= f_k(0)e^{-\tau(A_k+B_k)t} + \frac{B_k}{A_k + B_k}.\end{aligned}\quad (5-71)$$

Clearly, the steady state solution $f_k^\infty = B_k / (A_k + B_k)$ (for $-k$, by analogy), so finally we have

$$\begin{aligned}f_k^\infty &= \frac{f_k^L(1 + R_k^2) + f_{k'}^R T_k^2}{1 + R_k^2 + T_k^2}, \\ f_{-k}^\infty &= \frac{f_{k'}^R(1 + R_{-k}^2) + f_k^L T_{-k}^2}{1 + R_{-k}^2 + T_{-k}^2}.\end{aligned}\quad (5-72)$$

In equilibrium ($V=0$ so $k'=k$), we automatically obtain $f_k^\infty = f_{-k}^\infty = f_k^L = f_k^R$. Note how, only if the transmission is low do we actually get $f_k^\infty \approx f_k^L$ and $f_{-k}^\infty \approx f_{k'}^R$, which is the assumption most often used in mesoscopic calculations. We see this assumption may not be entirely justified: if there is appreciable transmission, the active region feels the distribution functions in both contacts, not just the injecting contact!

The resulting current per spin orientation can be calculated from the quantum-mechanical relationship

$$\begin{aligned}j(x) &= \frac{e\hbar}{2m^*i} \sum_{k>0} f_k^\infty \left[\frac{d\Psi_k^*(x)}{dx} \Psi_k(x) - \Psi_k^*(x) \frac{d\Psi_k(x)}{dx} \right] \|\Psi_k\|^{-2} \\ &\quad + f_{-k}^\infty \left[\frac{d\Psi_{-k}^*(x)}{dx} \Psi_{-k}(x) - \Psi_{-k}^*(x) \frac{d\Psi_{-k}(x)}{dx} \right] \|\Psi_{-k}\|^{-2} \\ &= \sum_{k>0} f_k^\infty j_k(x) + f_{-k}^\infty j_{-k}(x).\end{aligned}\quad (5-73)$$

Each of the two current components is constant across the structure. The simplest way to calculate them is by focusing on the regions far from the barriers, where the wavefunctions are plane waves. So to the far right of the heterostructure,

$$\begin{aligned}j^+ &= \frac{e\hbar}{2m^*i} \sum_{k>0} f_k^\infty \frac{|t_{k',R}|^2 (2ik')}{\|\Psi_k\|^2} = \frac{e\hbar}{m^*} \frac{W}{2\pi} \int dk f_k^\infty \frac{(|t_{k',R}|^2 k'/k)k}{\|\Psi_k\|^2} \\ &= \frac{eW}{h} \int dE_k f_k^\infty \frac{T_k}{\|\Psi_k\|^2}.\end{aligned}\quad (5-74)$$

where $T_k = |t_{k',R}|^2 k'/k$ is the elastic transmission coefficient (the same for both directions and dependent on energy only), and $W = x_R - x_L$ is the width of the active region (the well, the barriers, and a large enough portion of the contacts to arrive at the flatband conditions and also reasonably ensure a quasicontinuum of k 's). Similarly, we find the current carried by backward propagating states to be

$$j^- = \frac{e\hbar}{2m^*i} \sum_{k>0} f_k^\infty \frac{|t_{-k,L}|^2 (-2ik)}{\|\Psi_{-k}\|^2} = -\frac{eW}{h} \int dE_k f_{-k}^\infty \frac{T_k}{\|\Psi_{-k}\|^2}, \quad (5-75)$$

so the total current per spin orientation can be found as

$$j = \frac{eW}{h} \int dE_k T_k \left\{ \frac{f_k^\infty}{\|\Psi_k\|^2} - \frac{f_{-k}^\infty}{\|\Psi_{-k}\|^2} \right\}, \quad (5-76)$$

Figure 5-7 shows the I-V curve for the RTD of Fig. 4.6, as calculated according to the exact expression (5-76), and the Landauer formula for the current per spin orientation

$$j^{Lan} = \frac{e}{h} \int dE_k T_k \{f_k^L - f_{k'}^R\} \quad (5-77)$$

The width of each contact in the calculation is 20 times the width of the well and barriers combined. The Landauer formula predicts the peak at a lower voltage and overestimates the peak current with respect to the exact solution. The reason is that the distribution functions in the active region coincide with the distribution functions in the injecting contacts only for energies for which transmission is not high, i.e., away from the resonance; otherwise, a transmitting RTD feels the distribution functions in both contacts. All three formulas describe ballistic transport, so no crossing of the curves typical for the inclusion of inelastic scattering should be expected. The only difference is that j (5-76) captures the strongly nonlocal nature of tunneling.

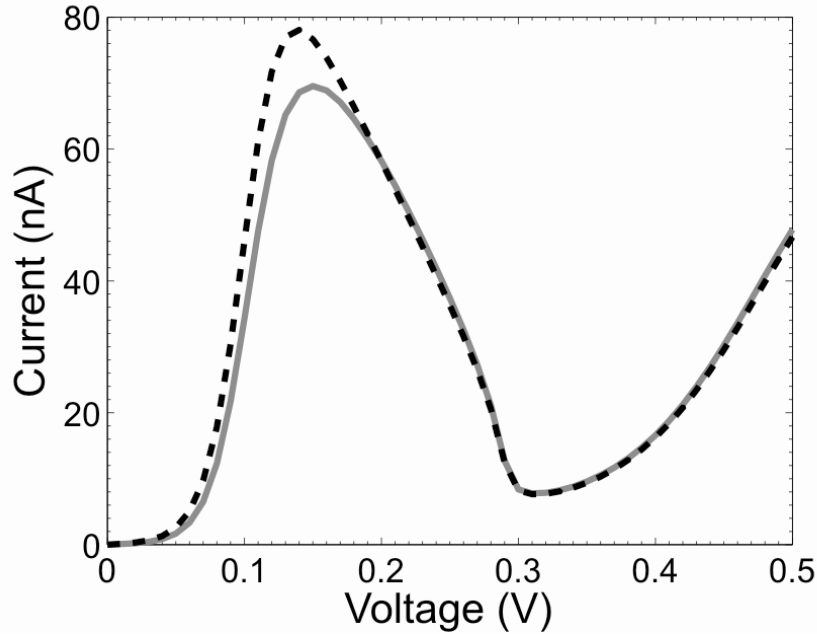


Figure 5-7: Steady-state I-V curve for the RTD of Fig. 4.6, according to the exact expression (4.76) (solid curve), and the Landauer formula (4.77) (dashed curve), at 77 K.

Our results for f_k^∞ and f_{-k}^∞ basically give the explicit, exact, ballistic limit of the active region's reduced density matrix away from equilibrium. These results can further be built on to include the electron-electron interaction between the active region and contacts or within the active region itself, or to include scattering with phonons. Starting from this ballistic nonequilibrium many-body density matrix, Green's functions formalism could be generalized to include scattering in open systems: the ballistic open-system nonequilibrium many-body density matrix will be the starting point, as opposed to the closed-system noninteracting density matrix [in which the injection from the contacts also has to be put in and is done through the injection self-energy terms]. The next Section outlines how, by starting from the open system's techniques basis, one can generalize the single-particle Green's functions.

Finally, the Markovian map (5-61) is valid for $\|L_{eff}\| \gg \|\Lambda\|\tau$, which basically yields $\Delta_k^2 \tau \ll \omega_k$.

By approximating $\|\Psi_k\|^2 \approx W$, we obtain, which will be satisfied for $\tau \approx 10$ fs (appropriate for contact doping of about 10^{20}cm^{-3}), in GaAs-based structures whose active region is longer than a few nanometers.

5.3 Generalizing Green's function for a full treatment of dynamically open systems

The seminal work by Kadanoff and Baym [107] (and independently by Keldysh [110]), laid the foundation for much of the many-body nonequilibrium transport theory that is used in mesoscopic transport today. Actually, these authors were not the first to employ Green's functions, that have been, till then, exclusively field-theoretical tools. The pioneering works on the field-theoretical method implementation to condensed matter physics go as far back as 1940's, with the famous Bogoliubov-Born-Green-Kirkwood-Yvon (BBGKY) hierarchy (for a comprehensive overview, see Ref. [202] and references therein) for the cluster expansion of the many-body density matrix. This latter work sprung off a whole movement in derivations of kinetic master equations, with notable contributions certainly being due to Progogine [203], Resibois [204], Zwanzig [205], Balescu [206], and others. Another line of work is based on the contributions by Martin and Schwinger [105,106], who introduced functional differentiation as a means of systematically constructing non-perturbative approximations for Green's functions, and these ideas are direct predecessors of Kadanoff and Baym's book [107]. Also, a lot of work on Green's functions in solid-state physics was done independently in the former Soviet Union (an example being the evergreen book by Abrikosov, Gorkov, and Dzyaloshinskii [111]) and Japan (Matubara's application of Green's functions to finite temperatures [207], and Kubo's introduction of periodic boundary conditions for two-time correlation functions [60]). This early work on nonequilibrium Green's functions pointed out that the quantum-mechanical counterpart of the Boltzmann distribution function is $-iG^<$, where $G^<$ is one of the so-called *two-time correlation functions*. It was also demonstrated that the effect of scattering lies in the broadening of the so-called spectral density (i.e., density of states), causing it to transform from a sharply peaked Dirac's δ -function to a broadened Lorentzian-type shape (in the long-time limit). Kadanoff and Baym derived equations of motion for the two-time correlation functions in the presence of an external potential (Kadanoff-Baym equations) [107]. On the other hand, Keldysh introduced time ordering on the so-called Keldysh contour, with the contour-ordered Green's function being a matrix that contains all four two-time Green's functions, and possessing a perturbation expansion in the form typical for the zero-temperature formalism. However, the nonequilibrium Green's function formalism is inherently defined for closed systems that obey Hamiltonian dynamics. The evolution is perfectly reversible, with the well-defined Heisenberg and interaction pictures, and consequently, the definitions of the Green's functions and the perturbation expansion.

The pioneering efforts in the application of nonequilibrium Green's functions (NEGF) to semiconductors are due to Jauho and co-workers [196], but the approach most cited in recent years, especially for ballistic structures, is that of Datta and co-workers [198,199,200]. A notable feature about this approach is the fact that, although the active region is treated as dynamically closed, the contact/active region boundaries are considered to be open, in the sense that the contacts constantly inject electrons. There exists a *self-energy* contribution due to injection from the contacts, which bears information about the changes in the energy spectrum and the level broadening due to the coupling with the contacts.

In this Section, we will point out how nonequilibrium Green's functions (NEGF), described previously in this Chapter, can be adapted to account fully for the dynamical openness of the nanostructure's active region. In a nutshell, one would construct the diagrammatic technique by starting not with the Hamiltonian (unitary) dynamics of the active region+contacts full density matrix, but instead with the generally non-Markovian and certainly non-unitary evolution of the active region's reduced density matrix alone. The first task is to generalize the closed system's two-time correlation functions, i.e., functions of the type $\text{Tr}[\rho_H a_H(t') b_H(t)]$, where ρ is the (closed) system density matrix, and a and b are two system operators, all given in the Heisenberg picture, as indicated by the subscript H (no subscript indicates the Schrödinger picture). A definition of the correlation functions has been attempted previously by Haake [208], but, in his work, one was required to use the full evolution of the open system+environment, which is precisely what we wish to avoid as best we can. A generalization of the two-time correlation functions will eventually enable us to define the core transport variables: the "greater-than" and "less-than" single-particle Green's functions, given for closed systems by

$$\begin{aligned} iG^>(1,1') &= \text{Tr}[\rho_H \psi_H(t) \psi_H^\dagger(t')], \\ \pm iG^<(1,1') &= \text{Tr}[\rho_H \psi_H^\dagger(t') \psi_H(t)], \end{aligned} \quad (5-78)$$

where the upper sign refers to bosons, and the lower sign to fermions. The field operators, at this point, can be the creation or annihilation operators at a point in space, or in a single-particle state. The following approach can be easily generalized to higher-order Green's functions.

We have already seen that a closed system with Hamiltonian $h(t)$, which may be time-dependent due to external driving forces, the density matrix ρ in the Schrödinger picture obeys the quantum Liouville equation,

$$\frac{d\rho(t)}{dt} = -i[h(t), \rho(t)] \equiv -iL(t)\rho(t),$$

with the solution given by

$$\begin{aligned} \rho(t) &= U(t, t_0)\rho(t_0), \\ U(t, t') &= \Theta(t-t') T^c \exp\left(-i \int_{t'}^t d\tau L(\tau)\right) + \Theta(t'-t) T^a \exp\left(i \int_t^{t'} d\tau L(\tau)\right). \end{aligned}$$

If a and b are time-independent operators in the Schrödinger picture, it is easily shown that the two-time correlation functions can be written as

$$\text{Tr}[\rho_H a_H(t') b_H(t)] = \text{Tr}\{a U(t', t)[b \rho(t)]\}. \quad (5-79)$$

Namely, when written in the Schrödinger picture, the desired expectation value actually means that, at time t , b acts on ρ (actually, b acts on the 'ket' part of ρ), then $b\rho(t)$ evolves under the quantum Liouville equation (i.e., is the argument of $U(t', t)$), until a acts on the result at time t' . The form on the right-hand-side of (5-79) is actually the form that allows for a generalization to open systems, while keeping the proper physical meaning.

5.3.1 Two-time correlation functions for open systems

In the system Liouville space \mathcal{H}_S^2 , multiplication of the reduced system density matrix ρ_S by the system creation and annihilation operators, ψ_S^\dagger and ψ_S , can be described by superoperators Ψ^+ and Ψ^- acting on ρ_S as a vector in \mathcal{H}_S^2 . It is understood that the Liouville space is constructed to allow for this action, i.e., that, in addition to the states corresponding to a given number of particles (e.g., electrons), which are used to construct ρ_S , at least the states with ± 1 electron are included. This is a computational rather than a theoretical requirement, and it suffices when one is interested only in two-time correlation functions, meaning that at most one particle is created (annihilated). Accommodation of higher-order Green's functions will require further augmentation of the system Liouville space during computation.

Within the total system+environment, annihilation of a system particle can be described by the superoperator Ψ^- , of the form $\Psi^- = I_E \otimes \Psi_S^-$, which, according to Eq. (5-30b), has a following block-diagonal form in the eigenbasis of \bar{P} :

$$\Psi = \begin{bmatrix} \Psi_S & 0 \\ 0 & \Psi_2 \end{bmatrix}. \quad (5-80)$$

Obviously, the creation operator ψ_S^+ will be associated with Ψ^+ .

We are now almost fully equipped to define the two-time correlation functions for open systems. Following the discussion of the introductory part of this Chapter, a definition of $G^>(1,1')$ for the open system requires that, first, a system particle is created at time t' (i.e., Ψ^+ acts on $\rho(t')$), then $\Psi^+ \rho$ evolves until t , when a particle is annihilated. However, *we only need the system point of view of this action*, not the full system+environment. So, we define two auxiliary variables $\rho^{\Psi^+,t'}(t)$ and $\rho^{\Psi,t'}(t)$ such that

$$\begin{aligned} \rho^{\Psi^+,t'}(t) &= \Psi^+ \rho(t'), \quad \rho^{\Psi^+,t'}(t) = U(t,t') \rho^{\Psi^+,t'}(t'), \\ \rho^{\Psi,t'}(t) &= \Psi \rho(t'), \quad \rho^{\Psi,t'}(t) = U(t,t') \rho^{\Psi,t'}(t'). \end{aligned} \quad (5-81)$$

We then define the open system's "greater-than" and "less-than" Green's functions as

$$\begin{aligned} iG_S^>(1,1') &\equiv \text{Tr}_S [\Psi_S \rho_S^{\Psi^+,t'}(t)], \\ \pm iG_S^<(1,1') &\equiv \text{Tr}_S [\Psi_S^+ \rho_S^{\Psi,t'}(t)]. \end{aligned} \quad (5-82)$$

These definitions have exactly the required physical meaning. The information provided by these functions is clear after writing Eqs. (5-82) as

$$\begin{aligned} iG_S^>(1,1') &= \text{Tr}_S [\Psi_S U_{11}(t,t') \Psi_S^+ \rho_S(t')] + \sqrt{d_E} \cdot \text{Tr}_S [\Psi_S U_{12}(t,t') \Psi_2^+ \rho_2(t')], \\ \pm iG_S^<(1,1') &= \text{Tr}_S [\Psi_S^+ U_{11}(t',t) \Psi_S \rho_S(t)] + \sqrt{d_E} \cdot \text{Tr}_S [\Psi_S^+ U_{12}(t',t) \Psi_2 \rho_2(t)]. \end{aligned} \quad (5-83)$$

Before proceeding with the analysis of Eq. (5-83), it is important to recall that there are two classes of states in \mathcal{H}_{S+E}^2 : the purely system states, belonging to $(\mathcal{H}_{S+E}^2)_{\bar{P}=1}$, and the entangled states, belonging to $(\mathcal{H}_{S+E}^2)_{\bar{P}=0}$. We now see that each of the equations (5-83) contains two terms: one that describes propagation within the purely system states, and one that describes *the transfer of information between the entangled states and purely system states*. In Figure 5-8, we have depicted the contributions from the two terms to $G^<$. If we were to neglect the system-environment coupling, i.e., if the system were treated as closed, only the first term (Figure 5-8a) would survive, so propagation between the purely system states and purely system states would be *closed-system-like*.

An illustration of the solidity of our definition (5-83) is the form of the average density at a given point. Namely,

$$\begin{aligned} \langle n_S(\mathbf{r},t) \rangle &= \pm iG_S^<(\mathbf{r}t,\mathbf{r}t) = \text{Tr}_S [\Psi_S^+(\mathbf{r}) \Psi_S(\mathbf{r}) \rho_S(t)] \\ &= \text{Tr}_S [\psi_S^+(\mathbf{r}) \psi_S(\mathbf{r}) \rho_S(t)] \\ &= \text{Tr}_S [n(\mathbf{r}) \rho_S(t)], \end{aligned} \quad (5-84)$$

since $U(t,t) = 1$, so $U_{11}(t,t) = 1, U_{12}(t,t) = 0$, and the result above is exactly what is expected.

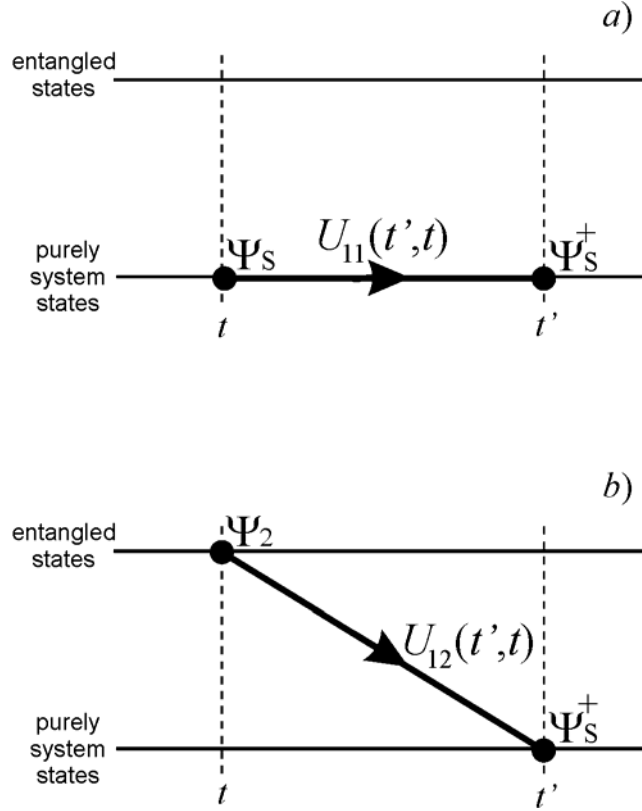


Figure 5-8. Illustration of the two terms in the definitions of $\pm iG^{\lessgtr}$, from Eq. (5-83). (a) A particle is destroyed among the purely system states at t , and created at t' among the purely system states. (b) A particle is destroyed among the entangled states at t , and due to this event the information about the state of the environment becomes apparent at t' , when the particle is created among the purely system states.

Having properly defined $G^>$ and $G^<$, we can now define the chronological and the antichronological Green's functions, G^c and G^a , as

$$\begin{aligned} G_S^c(1,1') &= \Theta(t-t')G_S^>(1,1') + \Theta(t'-t)G_S^<(1,1'), \\ G_S^a(1,1') &= \Theta(t'-t)G_S^>(1,1') + \Theta(t-t')G_S^<(1,1'). \end{aligned} \quad (5-85)$$

Writing down the equations of motion for these time-ordered Green's functions would be the next logical step. However, since we do not wish to specify anything about the interaction, and thus cannot *a priori* expect to obtain a Martin-Schwinger-type hierarchy, we will not proceed along these lines. Rather, we will focus our interest on the importance of the memory effects on transport.

5.3.2 Transport in the transient regime

Recently, transients have been receiving enhanced theoretical attention [157,159,160]. For example, much work on treatment of initial correlations in nonequilibrium Green's functions has been done recently [189]. Ladder-type diagrams have been obtained as a correction to the diagrammatic expansions due to the initial correlations. However, we believe that the introduction of initial correlations, which are undoubtedly very important, may be insufficient to describe the relaxation. Here, we will try to achieve an understanding of transient processes in general, within the present approach and without necessarily specifying the details of the Hamiltonians. We will investigate how exactly the memory effects influence the near-equilibrium transport, i.e., how the state of the environment and the entanglement between the system and the environment states are observed in the open system's evolution.

During transient processes, the state of $S+E$ can be tracked back to the initial state at time t_0 , which is assumed to be given by a known density matrix $\rho(t_0)$. There usually exists a typical relaxation time τ_{relax} , during which the transient may be considered to occur, and after which a steady state is achieved. Consider $G_S^<(1,1')$ in the transient regime, with the initial time set to t_0 , and $t_0 < t, t' \ll \tau_{\text{relax}}$. Using Eqs. (5-83) and (5-32b), we obtain

$$\begin{aligned} \pm iG_S^<(1,1') = & \text{Tr}_S [\Psi_S^+ U_{11}(t',t) \Psi_S U_{11}(t,t_0) \rho_S(t_0)] \\ & + \text{Tr}_S [\Psi_S^+ U_{12}(t',t) \Psi_2 U_{21}(t,t_0) \rho_S(t_0)] \\ & + \sqrt{d_E} \text{Tr}_S [\Psi_S^+ U_{11}(t',t) \Psi_S U_{12}(t,t_0) \rho_2(t_0)] \\ & + \sqrt{d_E} \text{Tr}_S [\Psi_S^+ U_{12}(t',t) \Psi_2 U_{22}(t,t_0) \rho_2(t_0)]. \end{aligned} \quad (5-86)$$

Of course, $G_S^<$ measures the probability of ending up in the same system state, after having annihilated a particle at time t at a given position \mathbf{r} , and then having created it at a (later) time t' at \mathbf{r}' . Apparently, there are four terms that contribute to this correlation function, and are depicted in Figure 5-9. The first term is the *closed-system-like term* (Figure 5-9a), the only one that survives if the coupling between S and E is turned off (i.e., if $L_{12} \rightarrow 0$), and the nickname given has to do with the fact that this term describes only the influence that the purely system states have on the purely systems states, whereas the *information on the state of the environment is never incorporated*. The last two terms (Figs. 5-9c,d) are the so-called *memory terms*, as they start with the entangled states, so they contain the information on the *initial state of the environment*. These terms are both of the first order in coupling L_{12} . The second term (Fig. 5-9b) is dubbed *the entanglement term*, as it does start and end with the purely system states, but meanwhile the information on the state of the environment is sampled. This term is of the second order in L_{12} . The classification of terms as closed-system-like, entanglement and memory terms retains its meaning even in multiple-time correlation functions (for entanglement terms, the requirement will be that the term starts and ends with purely system states, but meanwhile *at least once* the state of the environment is sampled). In the light of generalizing the Kadanoff-Baym-Keldysh equations for open systems, we believe that the entanglement terms will be possible to treat through a type of correction to the self-energy part, whereas the memory terms will unfortunately remain as additive terms in equations of motion. Evaluation of submatrices U_{ij} , needed to calculate the open system's $G^<$ (and $G^>$, in a similar fashion), becomes difficult with increasing size of the system and the environment, and direct computation is generally out of the question. However, within the time-convolutionless approach, U_{21} and U_{22} can always be written in terms of U_{11} and U_{12} , but formalization of these relationships requires establishing a time t_0 , so the resulting equations of motion are Eqs. (5-41).

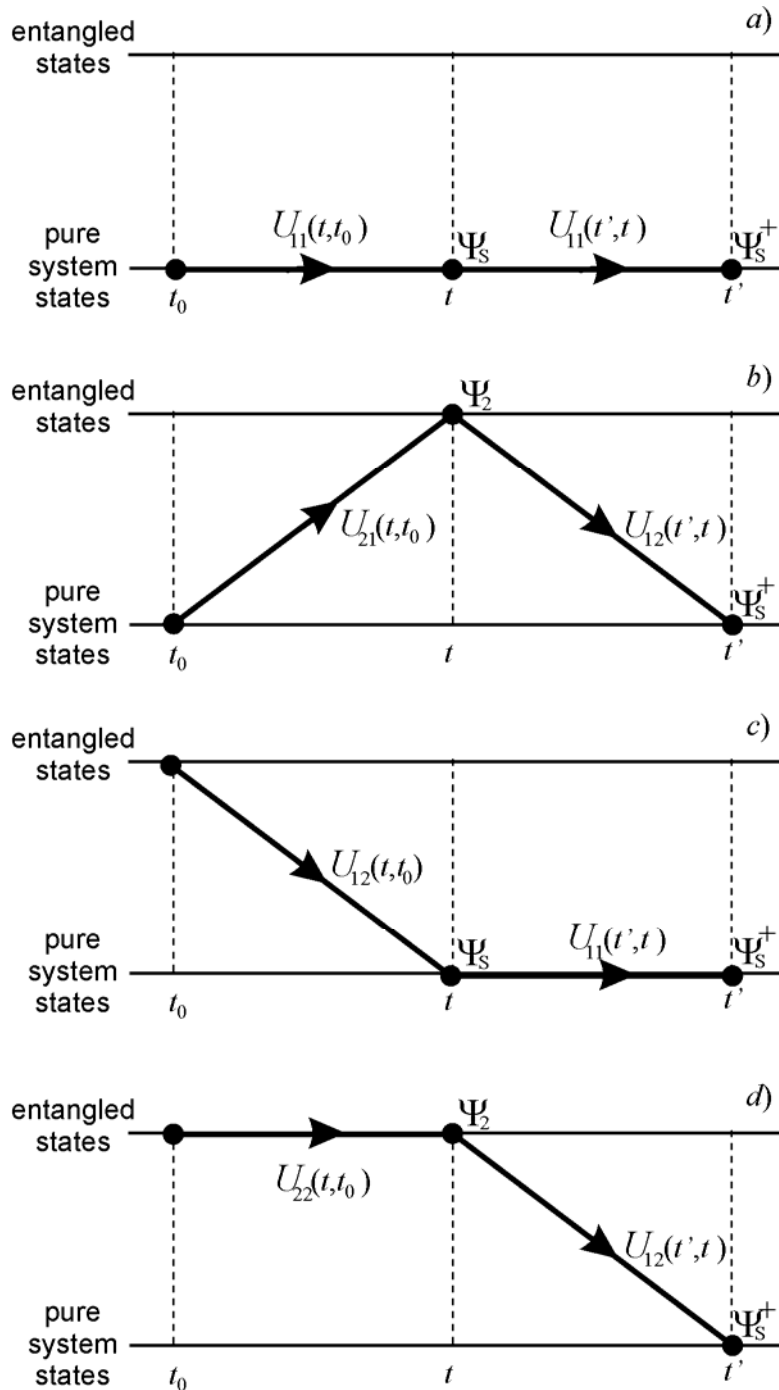


Figure 5-9. The four terms in $G_S^<$ for an open system in the transient regime, from Eq. (4.86). (a) The closed-system-like term; (b) the entanglement term; (c) and (d) the memory terms.

5.3.3 Transport in a far-from-equilibrium steady state

Reaching a well-controlled steady state, independent of initial conditions, is the goal of applying external driving forces in various systems, such as semiconductor devices. By a system in a far-from-equilibrium steady state we primarily consider a system driven by an external source (e.g., applied external bias), so that after a sufficiently long time *its relevant output quantities* (e.g., current, conductance, etc.) have reached values that no longer significantly vary with time (steady state), and are virtually insensitive to the exact preparation of the initial state, and independent of the values that these quantities would have if

evaluated in near equilibrium (far-from-equilibrium). Usually, there exists a characteristic relaxation time τ_{relax} , after which the system is considered to be stationary, at least with respect to a collection of measured average quantities and within a given margin of error. However, if we enhance the number of (independent) measured quantities, and all of them are to be constant in time, then it is a reasonable conclusion that the system density matrix ρ_S in a steady state may be considered a constant [209], i.e.,

$$\rho_S \approx \text{const.} \quad (t \gg \tau_{\text{relax}}). \quad (5-87)$$

The relaxation time is sufficiently long to destroy the information about the initial correlations, and build up new ones, in agreement with the external driving forces. If we have no interest in the fine details of the relaxation process, but instead just in the steady state properties, and the initially uncoupled system and environment is a reasonable approximation, then the evolution of the reduced density matrix in its approach to the steady state can be approximated by our coarse-grained semigroup map (5-61)

$$\frac{d\rho_S}{dt} = (-iL_{\text{eff}} - \Lambda\tau)\rho_S(t) \Rightarrow \rho_S(t) = V(t, t_0)\rho_S(t_0), \quad V(t, t_0) = \exp\{(-iL_{\text{eff}} - \Lambda\tau)t\} \quad (5-88a)$$

where

$$(-iL_{\text{eff}} - \Lambda\tau)\rho_S(t \gg \tau_{\text{relax}}) \approx 0 \rightarrow \rho_S(t \gg \tau_{\text{relax}}) \equiv \rho_S^\infty = \text{const.} \quad (5-88b)$$

As a consequence of the decoupling requirement, the far-from-equilibrium steady-state $G_S^<$ takes on the typical form for closed systems (for $t, t' \gg \tau_{\text{relax}}$)

$$\begin{aligned} \pm iG_S^<(1, 1') &= \text{Tr}_S[\Psi_S^\dagger V(t', t)(\Psi_S \rho_S(t))] \\ &\approx \text{Tr}_S[\Psi_S^\dagger V(t', t)(\Psi_S \rho_S^\infty)] \end{aligned} \quad (5-89)$$

The situation seems promising, because the commutator-generated L_{eff} can be taken as the the basis for defining the unperturbed Hamiltonian (basically the mean-free one), while the remainder of the evolution, generated by Λ can be treated perturbatively. Unfortunately, we do not usually know what the far-from-equilibrium steady-state ρ_S^∞ actually is, and clearly one obtains a single far-from-equilibrium state from multiple initial conditions. That is why a criterion such as finding ρ_S^∞ in the intersection of the null spaces of L_{eff} and Λ is particularly useful. One word of caution: the nonequilibrium steady state does not, in general, admit Wick's decomposition (this was shown for the so-called *relevant statistical operator* [210] that satisfies the self-consistency requirements for a given set of state parameters, and is basically a top-down approximation for ρ_S^∞). However, the Dyson equation can be recovered in a manner similar to that demonstrated for the so-called *mixed Green's functions* [211].

6. Conclusions

In this review article we have given a brief description of currently most important and most physically based semi-classical and quantum transport approaches. Particle-based device simulators can capture the essential physics up to ballistic transport regime and, when quantum interference effects start to dominate device behavior, quantum transport simulators based on either direct solution of the Schrödinger equation or its counterpart, the Green's functions, have been developed which, with the recent progress of state of the art computers, can simulate 3D nanoscale devices within a reasonable time-frame.

However, nanoelectronic device simulation of the future *must* ultimately include both, the sophisticated physics oriented electronic structure calculations and the engineering oriented transport simulations. Extensive scientific arguments have recently ensued regarding transport theory, basis representation, and practical implementation of a simulator capable of describing a realistic device. Starting from the field of molecular chemistry, Mujica, Kemp, Roitberg, Ratner²¹² applied tight-binding based approaches to the modeling of transport in molecular wires. Later, Derosa and Seminario²¹³ modeled

molecular charge transport using density functional theory and Green's functions. Further significant advances in the understanding of the electronic structure in technologically relevant devices were recently achieved through *ab initio* simulation of MOS devices by Demkov and Sankey²¹⁴. Ballistic transport through a thin dielectric barrier was evaluated using standard Green function techniques^{215,216} without scattering mechanisms. However, quantum mechanical simulations of electron transport through 3D confined structures, such as quantum dots, have not yet reached the maturity (it is important, for example, for simulating operation of the next generation quantum dot photodetectors). Early efforts of understanding the operation of coupled quantum dot structures were rate equation based^{217,218,219} where a simplified electronic structure was assumed.

Whereas traditional semiconductor device simulators are insufficiently equipped to describe quantum effects at atomic dimensions, most *ab-initio* methods from condensed matter physics are still computationally too demanding for application to practical devices, even as small as quantum dots. A number of intermediary methods have therefore been developed in recent years. The methods can be divided into two major theory categories: atomistic and non-atomistic. Atomistic approaches attempt to work directly with the electronic wave function of each individual atom. *Ab-initio* methods overcome the shortcomings of the effective mass approximation; however, additional approximations must be introduced to reduce computational costs. One of the critical questions is the choice of a basis set for the representation of the electronic wave function. Many approaches have been considered, ranging from traditional numerical methods, such as finite difference and finite elements, as well as plane wave expansions^{220,221,222}, to methods that exploit the natural properties of chemical bonding in condensed matter. Among these latter approaches, local orbital methods are particularly attractive. While the method of using atomic orbitals as a basis set has a long history in solid state physics, new basis sets with compact support have recently been developed^{223,224}, and, together with specific energy minimization schemes, these new basis sets result in computational costs which increase linearly with the number of atoms in the system without much accuracy degradation^{225,226}. However, even with such methods, only a few thousand atoms can be described with present day computational resources. NEMO3D uses an empirical tight-binding method^{227,228} that is conceptually related to the local orbital method and combines the advantages of an atomic level description with the intrinsic accuracy of empirical methods. It has already demonstrated considerable success^{229,230,229} in quantum mechanical modeling of electron transport as well as the electronic structure modeling of small quantum dots²³¹. NEMO3D typically uses sp^3s^* or $sp^3d^5s^*$ model that consists of five or ten spin degenerate basis states, respectively. Note that for the modeling of quantum dots, three main methods have been used in recent years: $k \cdot p$ ^{232,233}, pseudopotentials²²⁰, and empirical tight-binding²³¹.

As already discussed in Section 3, there are a number of methods developed by solid state theorists over the last several decades to address the issue of quantum transport in nano-devices. Among the most commonly used in nanostructure calculations schemes are the Wigner-function approach²³⁴, the Pauli master equation²³⁵, and the non-equilibrium Green's functions (NEGF)^{236,237}. The growing popularity of the latest (sometimes referred to as the Keldysh or the Kadanoff–Baym) formalism is conditioned by its sound conceptual basis for the development of the new class of quantum transport simulators²³⁸. Among its doubtless advantages are the clear physical conceptions, rigorous definitions, well-developed mathematical apparatus and flexibility of the algorithmization.

Thus, in our opinion, *the goal of any future simulation effort is to merge the electronic structure calculations with the quantum transport calculations and develop a NEGF technique that is numerically efficient and ready for engineering applications in 3D objects on the one hand (such as QDIP), and rigorously quantum-mechanical on the other hand so that it properly incorporates the electronic structure of, for example, regular or disordered quantum dots used in QDIPs.*

The groups from ASU and Purdue are currently working on the development of such simulator in order to be able to calculate all the properties of 3D open quantum systems, particularly QDIPs. The transport kernel of the simulator is based on the Contact Block Reduction (CBR) method^{127,128} and is discussed in more details in section 4 of this review article. As already noted, the CBR method is applicable to fully self-consistent quantum transport calculations in arbitrarily shaped 3D structures using either the effective mass approximation or the multi-band Hamiltonian description. The band-structure of the QDIP's will be calculated using NEMO3D simulation software.

In summary, from the discussion above it follows that *the ultimate goal of semiconductor transport calculation of future nanoscale devices will be to merge the 3D quantum transport approaches with ab-initio band structure calculations.* This will ensure the most accurate simulation and better understanding carrier transport and operation of novel nano-device structures.

Acknowledgements

This work has been supported in part by the Arizona Institute for Nano Electronics (AINE) and an NSF collaborative grant ECCS 0701926.

References

- [1] S. M. Sze and G. S. May, *Fundamentals of Semiconductor Fabrication* (John Wiley and Sons Inc., 04 April, 2003).
- [2] P. D. Agnello, *IBM J. Res. & Dev.*, Vol. 46, 317 (2002).
- [3] B. A. Kramer, R. J. Weber, *Electronics Letters*, Vol. 28, 1106 (1992).
- [4] In 1954, Charles Townes and Arthur Schawlow invented the maser. Theodore Maiman invented the ruby laser considered to be the first successful optical or light laser. Many historians claim that Theodore Maiman invented the first optical laser, however, there is some controversy that Gordon Gould was the first.
- [5] *International Technology Roadmap for Semiconductors*, 2006 Edition, Semiconductor Industry Association (SIA), Austin, Texas: SEMATECH, USA, 2706 Montopolis Drive, Austin, Texas 78741; <http://www.itrs.net/ntrs/publntrs.nsf>
- [6] G. Moore, *IEDM Tech. Digest*, 11 (1975)
- [7] R. Dennard, F. H. Gaensslen, H. N. Yu, L. Rideout, E. Bassous, and A. R. LeBlanc, *IEEE J. Solid State Circuits*. Vol. 9, 256 (1974).
- [8] M. Ono, M. Saito, T. Yoshitomi, C. Fiegna, T. Ohguro, and H. Iwai, *IEEE Trans. Electron Devices*, Vol. 42, 1822 (1995).
- [9] G. Timp, J. Bude, K.K. Bourdelle, J. Garno, A. Ghetti, H. Gossmann, M. Green, G. Forsyth, Y. Kim, R. Kleiman, F. Klemens, A. Kornblit, C. Lochstampfer, W. Mansfield, S. Moccio, T. Sorsch, D.M.Tennant, W.Timp, R.Tung., *IEDM Technical Digest*, pp. 55-58, 1999.
- [10] B. Yu *et al.*, *IEDM Tech. Dig.*, 937 (2001).
- [11] Robert Chau, B. Boyanov, B. Doyle, M. Doczy, S. Datta, S. Harelend, B. Jin, J. Kavalieros, and M. Metz, *4th Int. Symp. on Nanostructures and Mesoscopic Systems*, 17 (2003).
- [12] H. S. Wong, *IBM J., Res. & Dev.*, Vol. 46, 133 (2002).
- [13] W. Zhu, J. P. Han, T. P. Ma, *IEEE Trans. Electron Dev.*, Vol. 51, 98 (2004).
- [14] Welsler, J. L. Hoyt and J. F. Gibbons, *IEDM Tech. Dig.*, 1000 (1992).
- [15] G. Formicone, D. Vasileska, D.K. Ferry, *VLSI Design*, Vol. 6, 167 (1998).
- [16] D. Vasileska, G. Formicone and D.K. Ferry, *Nanotechnology*, Vol. 10, 147 (1999).
- [17] P. M. Garone, V. Venkataraman, and J. C. Sturm, *IEEE Electron Device Lett.*, Vol. 13, 56 (1992).
- [18] T. Ghani, M. Armstrong, C. Auth, M. Bost, P. Charvat, G. Glass, T. Hoffmann, K. Johnson, C. Kenyon, J. Klaus, B. McIntyre, K. Mistry, A. Murthy, J. Sandford, M. Silberstein, S. Sivakumar, P. Smith, K. Zawadzki, S. Thompson and M. Bohr, *IEDM Technical Dig.*, 978 (2003).
- [19] X. Huang, W.-C. Lee, C. Ku, D. Hisamoto, L. Chang, J. Kedzierski, E. Anderson, H. Takeuchi, Y.-K. Choi, K. Asano, V. Subramanian, T.-J. King, J. Bokor, and C. Hu, *IEDM Tech. Dig.*, 67 (1999).
- [20] L. Geppert, *IEEE Spectrum*, April 9, 2004.
- [21] D. K. Ferry and S. M. Goodnick, *Transport in Nanostructures* (Cambridge University Press, Cambridge, UK, 1997).
- [22] M. V. Fischetti, Z. Ren, P. M. Solomon, M. Yang, K. Rim, *Journal of Applied Physics*, Vol. 94, 1079 (2003).
- [23] S. Krishnan, D. Vasileska and M. Fischetti, *J. Vac. Sci. Technol. B*, Vol. 24(4), 1997 (2006)
- [24] N. Sano, A. Hiroki, K. Matsuzawa, *IEEE Trans. Nanotechnology*, Vol. 1, 63 (2002).
- [25] Irena Knezevic, *Ph. D. Dissertation*, Arizona State University, August 2004.
- [26] United States Patent 4573161.
- [27] S. M. Sze, *Modern Semiconductor Device Physics*, John Wiley and Sons, 1998. M. Shur, *Physics of Semiconductor Devices*, Prentice Hall Series in Solid State Electronics, ????
- [28] E. Bauser, "LPE-Grown Surfaces and Growth Mechanisms" in "Crystal Growth of Electronic Materials", E. Kaldis (Ed), North Holland, Amsterdam (1985).

-
- [29] R. Dingle, C. Weisbuch, H. L. Störmer, H. Morkoç, and A. Y. Cho, *Applied Physics Letters*, Vol. 40, Issue 6, pp. 507-510 (1982).
- [30] K. von Klitzing, G. Dorda, and M. Pepper, *Phys. Rev. Lett.*, Vol. 45, 494 (1980).
- [31] E. Calleja, P. M. Mooney, S. L. Wright, and M. Heiblum, *Appl. Phys. Lett.*, Vol. 49, 657 (1986).
- [32] F. Ali, and A. Gupta, *HEMTs and HBTs: Devices, Fabrication and Circuits*, Editors, Artech House, Boston, 1991.
- [33] L. Esaki, *Phys. Rev.*, Vol. 109, 603 (1958); R. Tsu, L. Esaki, *Appl. Phys. Lett.*, Vol. 22, 562 (1973).
- [34] K. K. Likharev, *FED Journal*, Vol. 6, 5 (1995).
- [35] L. L. Chang, L. Esaki and R. Tsu, *Appl. Phys. Lett.* 24, 593 (1974).; T. C. L. G Sollner,., W. D. Goodhue, P. E. Tannenwald, C. D. Parker, and D. D. Peck, *Appl. Phys. Lett.*, Vol. 43, 588 (1983).
- [36] D. Vasileska and S. M. Goodnick, *Materials Science and Engineering, Reports: A Review Journal* Vol. R38, 181 (2002).
- [37] W. R. Frensley, *Rev. Mod. Phys.*, Vol. 62, 745 (1990).
- [38] Selberherr, *vasileska, Goodnick*
- [39] G. Binnig and H. Rohrer, Nobel Prize in Physics for *their design of the scanning tunneling microscope*, 1986.
- [40] E. Merzbaher, *Quantum Mechanics* (Hamilton Printing Company, John Wiley & Sons, 1998).
- [41] J. von Neumann, *The Mathematical Foundations of Quantum Mechanics* (Princeton Univ. Press, Princeton NJ, 1932).
- [42] E. Schrödinger, *Ann. Phys.*, 489 (1926) p.79; E. Schrödinger, *Phys. Rev.*, Vol. 28, 1049 (1926).
- [43] Erwin Schrodinger, Albert Einstein, Max Planck, and H. A. Lorentz, *Letters on Wave Mechanics*, Philosophical Library Inc (March 1986).
- [44] L. de Broglie, *Recherches sur la théorie des quanta (Researches on the quantum theory)*, Thesis Paris, 1924.
- [45] D. K. Ferry, *Quantum Mechanics: An Introduction for Device Physicists and Electrical Engineers* (Institute of Physics Publishing, London, 2001).
- [46] D. Bohm, *Quantum Theory* (Dover Publication).
- [47] <http://www.eas.asu.edu/~vasilesk>
- [48] E. O. Kane, in *Tunneling Phenomena in Solids*, edited by E. Burstein and S. Lundqvist (Plenum, New York, 1969), p. 1
- [49] J. N. Schulman and Y. C. Chang, *Phys. Rev. B*, Vol. 27, 2346 (1983).
- [50] W. Frensley, <http://www.utdallas.edu/~frensley/technical/qtrans/qtrans.html>
- [51] C. Lent, D. Kirkner, *J. Appl. Phys.*, Vol. 67, 6353 (1990).
- [52] Z.-Y. Ting, E. T. Yu, and T. C. McGill, *Phys. Rev. B*, Vol. 45, 3583 (1992).
- [53] Y. X. Liu, D. Z.-Y. Ting, and T. C. McGill, *Phys. Rev. B*, Vol. 54, 5675 (1996).
- [54] C. Lent, D. Kirkner, *J. Appl. Phys.*, Vol. 67, 6353 (1990).
- [55] S.E. Laux, A. Kumar, and M. V. Fischetti, *J. Appl. Phys.*, Vol. 95, 5545 (2004).
- [56] R. Lake, G. Klimeck, R. C. Bowen, and D. Jovanovic, *J. Appl. Phys.*, Vol. 81, 7845 (1997).
- [57] A. Svizhenko, M. P. Anantram, T. R. Govindan, *J. Appl. Phys.*, Vol. 91, 2343 (2002).
- [58] R. Venugopal, Z. Ren, S. Datta, and M. S. Lundstrom, and D. Jovanovic, *J. Appl. Phys.*, Vol. 92, 3730 (2002).
- [59] C. Rivas, R. Lake, *Phys. Stat. Sol. (b)*, Vol. 239, 94 (2003).
- [60] R. Kubo, *J. Phys. Soc. Jpn.*, Vol. 12, 570 (1957).
- [61] R. Kubo, *Statistical Mechanics* (North Holland, Amsterdam, 1971).
- [62] D. K. Ferry, *Semiconductors* (MacMillan, New York, 1991).
- [63] G. D. Mahan, *Many-Particle Physics* (Plenum, New York, 1981).
- [64] A. L. Fetter and J. D. Walecka, *Quantum Theory of Many-Particle Systems* (McGraw-Hill, New York, 1971).
- [65] G. Rickayzen, *Green's Functions and Condensed Matter* (Academic Press, New York, 1980).
- [66] W. Jones and N. H. March, *Theoretical Solid State Physics* (Dover Publications, Inc., New York, 1985).
- [67] V. K. Thankappan, *Quantum Mechanics* (John Wiley & Sons, Inc., New York, 1985).
- [68] D. K. Ferry and C. Jacoboni, Eds., *Quantum Transport in Semiconductors* (Plenum, New York, 1992).
- [69] Dragica Vasileska, *PhD Thesis*, ASU, December 2005.

-
- [70] R. Kubo, *Many-Body Theory* (Syokabo, Tokyo and Benjamin, New York, 1966).
- [71] S. Datta, *Electronic transport in mesoscopic systems*, Cambridge University Press, Cambridge, 1995, pp.52-59. Sze S 1981 *Physics of Semiconductor Devices 2nd Edition* (New York : Wiley) p 270-9.
- [72] Y. Imry and R. Landauer, *Rev. Mod. Phys.*, Vol. 71, S306 (1999).
- [73] U. Fano, *Rev. Mod. Phys.*, Vol. 29, 74 (1957).
- [74] R. P. Feynman, *Statistical Mechanics, A Set of Lectures*, W. A. Benjamin, Reading, MA 1972, ch. 2.
- [75] I. Oppenheim, K. E. Shuler and G. H. Weiss, *Stochastic Processes in Chemical Physics: The Master Equation*, MIT Press, Cambridge, Mass. 1977.
- [76] H. J. Kreuzer, *Nonequilibrium Thermodynamics and its Statistical Foundations*, Oxford University Press, Oxford, 1981.
- [77] C. Jacoboni and L. Reggiani, *Rev. Mod. Phys.*, Vol. 55, pp. 645–705 (1983).
- [78] P.J. Price, *Semiconductors and Semimetals*, Vol. 14, pp. 249-334 (1979).
- [79] C. Jacoboni and P. Lugli, *The Monte Carlo Method for Semiconductor Device Simulation* (Springer-Verlag, Wien New York).
- [80] C. Kittel, *Introduction to Solid State Physics* (Wiley, New York, 1986, sixth edition).
- [81] N. W. Ashcroft and N. D. Mermin, *Solid State Physics* (Saunders College Publishing, 1976).
- [82] D. J. Griffiths, *Introduction to Quantum Mechanics* (Prentice Hall Inc., Englewood Cliffs, New Jersey, 1995).
- [83] R. F. Pierret, *Semiconductor Device Fundamentals* (Addison-Wesley, 1996).
- [84] S. M. Sze, *Physics of Semiconductor Devices* (John Wiley & Sons, Inc., 1981).
- [85] M. Lundstrom, *Fundamentals of Carrier Transport* (Cambridge University Press, Cambridge, 2000)
- [86] B. K. Ridley, *Quantum Processes in Semiconductors* (Oxford University Press, Oxford, 1988).
- [87] D. K. Ferry, *Semiconductor Transport* (Taylor & Francis, London, 2000).
- [88] J. P. McKelvey, *Solid State and Semiconductor Physics* (Krieger Pub. Co., 1982).
- [89] J. M. Ziman, *Electrons and Phonons: The Theory of Transport Phenomena in Solids* (Oxford University Press, New York, 2001).
- [90] L. I. Schiff, *Quantum Mechanics*, McGraw-Hill Inc., New York (1955).
- [91] J. R. Barker and D. K. Ferry, *Phys. Rev. Lett.*, Vol. 42, 1779 (1979).
- [92] Y.-C. Chang, D. Z.-Y. Ting, J. Y. Tang and K. Hess, *Appl. Phys. Lett.*, Vol. 42, 76 (1983).
- [93] L. Reggiani, P. Lugli and A. P. Jauho, *Phys. Rev. B*, Vol. 36, 6602 (1987).
- [94] D. K. Ferry, A. M. Kriman, H. Hida and S. Yamaguchi, *Phys. Rev. Lett.*, Vol. 67, 633 (1991).
- [95] P. Bordone, D. Vasileska and D. K. Ferry, *Phys. Rev. B*, Vol. 53, 3846 (1996).
- [96] D. Vasileska and S. M. Goodnick, *Computational Electronics* (Morgan and Claypool, 2006).
- [97] J. B. Gunn, *Solid State Commun.*, Vol. 1, 88 (1963).
- [98] E. Wigner, *Phys. Rev.*, Vol. 40, 749 (1932).
- [99] E. J. Heller, *J. Chem. Phys.*, Vol. 65, 1289 (1976).
- [100] W. R. Frensley, *Phys. Rev. B*, Vol. 36, 1570 (1987).
- [101] N.C. Kluksdahl, A.M. Kriman, D. K. Ferry and C. Ringhofer, *Phys. Rev. B*, Vol. 39, 7720 (1989).
- [102] W.R. Frensley, *Superlattices and Microstructures*, Vol. 4, 497 (1987).
- [103] H. Tsuchiya, M. Ogawa and T. Miyoshi, *IEEE Trans. Electron Devices*, Vol. 38, 1246 (1991).
- [104] J. von Neumann, *Mathematical Foundations of Quantum Mechanics, English Transl.* (Princeton University Press, Princeton, N. J., 1955).
- [105] P. C. Martin and J. Schwinger, *Phys. Rev.*, Vol. 115, 1342 (1959)
- [106] J. Schwinger, *J. Math. Phys.*, Vol. 2, 407 (1961).
- [107] L. P. Kadanoff and G. Baym, *Quantum Statistical Mechanics* (W. A. Benjamin, New York, 1962).
- [108] O. V. Konstantinov and V. I. Perel, *Sov. Phys.-JETP*, Vol. 12, 142 (1961).
- [109] I. E. Dzyaloshinski, *Sov. Phys.-JETP*, Vol. 15, 778 (1962).
- [110] L. V. Keldysh, *Zh. Eksp. Teor. Fiz.*, Vol. 47, 1515 (1964), [English Transl. *Sov. Phys. JETP*, Vol. 20, 1018 (1965)].
- [111] A. A. Abrikosov, L. P. Gorkov and I. E. Dzyaloshinski, *Quantum Field Theoretical Methods in Statistical Physics*, 2nd ed. (Pergamon, New York, 1965).
- [112] D. C. Langreth, in *Linear and Nonlinear Electron Transport in Solids*, ed. by J. T. Devreese and E. Van Doren (Plenum, New York, 1975).
- [113] P. Danielewicz, *Ann. Phys.*, Vol. 152, 239 (1984).

-
- [114] J. Rammer and H. Smith, *Rev. Mod. Phys.*, Vol. 58, 323 (1986).
- [115] R. D. Mattuck, *A Guide to Feynmann Diagrams in the Many-Body Problem* (Dover Publications, Inc., New York, 1992).
- [116] A. Blandin, A. Noutier and D. W. Hone, *J. Phys. (Paris)*, Vol. 37, 369 (1976).
- [117] R. A. Craig, *J. Math. Phys.*, Vol. 9, 605 (1968).
- [118] H. Haug and S. W. Koch, *Quantum Theory of the Optical and Electronic Properties of Semiconductors* (World Scientific Publishing Co., Singapore, 1990).
- [119] W. Hänsch and G. D. Mahan, *Phys. Rev. B*, Vol. 28, 1902 (1983).
- [120] F. Gámiz, J. Banqueri, I. Melchor, J. E. Craceller, P. Cartujo and J. A. López-Villanueva, *J. Appl. Phys.*, Vol. 74, 3289 (1993).
- [121] F. F. Fang and W. E. Howard, *Phys. Rev. Lett.*, Vol. 16, 797 (1966).
- [122] Y. Matsumoto and Y. Uemura, *Jpn. J. Appl. Phys. Suppl.* 2, 367 (1974).
- [123] S. Takagi, A. Toriumi, M. Iwase and H. Tango, *IEEE Trans. Electron Devices*, Vol. 41, 2363 (1994).
- [124] D. Scharfetter, private communication.
- [125] A.-P. Jauho, and J. W. Wilkins, *Phys. Rev. B*, Vol. 29, 1919 (1984)
- [126] M. Wagner, *Phys. Rev. B*, Vol. 44, 6104 (1991); *Phys. Rev. B*, Vol. 45, 11595 (1992).
- [127] D. Mamaluy, M. Sabathil, P. Vogl, *J. Appl. Phys.*, Vol. 93, 4628 (2003).
- [128] Hasanur R. Khan, Denis Mamaluy, Dragica Vasileska, *IEEE Trans El. Dev.*, Vol. 54, 784 (2007).
- [129] S. Datta, *Ftom Atom to Transistor* (Cambridge University Press, Cambridge, 1995).
- [130] A. Trellakis, A. T. Galick, A. Pacelli, U. Ravaioli, *J. Appl. Phys.*, Vol. 81, 7880 (1997).
- [131] R. Lake, G. Klimeck, R. C. Bowen, D. Jovanovic, D. Blanks, M. Swaminathan, *phys. stat. sol. (b)*, Vol. 204, 354 (1997).
- [132] L. Chand, C. Hu, *Superlattices and Microstructures*, Vol. 28, 351 (2000).
- [133] D. J. Frank, S. E. Laux and M. V. Fischetti, *IEDM Tech. Dig.*, 553 (1992).
- [134] H. Takato, K. Sunouchi, N. Okabe, A. Nitayama, K. Hieda, F. Horiguchi, F. Masuoka, *IEEE Trans El. Dev.* Vol. 38, 573 (1991).
- [135] C. Auth, J. Plummer, *IEEE Elec. Dev. Lett.*, Vol. 18, (1997).
- [136] H.-S. P. Wong, K. K. Chan, and Y. Taur, *IEDM Tech. Dig.*, 427 (1997).
- [137] J. Colinge, M. Gao, A. Romano-Rodriguez, H. Maes, and C. Claeys, *IEDM Tech. Dig.*, 595 (1990).
- [138] H.-S. Philip Wong, *Solid-State Electronics*, Vol. 49, 755 (2005).
- [139] D. Hisamoto, T. Kaga, Y. Kawamoto, and E. Takeda, *IEDM Tech. Dig.*, 833 (1989).
- [140] D. Hisamoto, W.-C. Lee, J. Kadzierski, H. Takeuchi, K. Asano, C. Kuo, T.-J King, J. Bokor and C. Hu, *IEEE Trans El. Dev.* Vol. 47, 2320 (2000).
- [141] Y. Choi, T. King and C. Hu, *IEEE Trans El. Dev.*, Vol. 23, 25 (2002).
- [142] Y.-K. Choi, N. Lindert, P. Xuan, S. Tang, D. Ha, E. Anderson, T.-J. King, J. Bokor, C. Hu, *IEEE International Electron Device Meeting Technical Digest*, 421 (2001).
- [143] X. Huang, W.-C. Lee, C. Kuo, D. Hisamoto, L. Chang, J. Kadzierski, E. Anderson, H. Takeuchi, Y.-K. Choi, K. Asano, V. Subramanian, T.-J King, J. Bokor and C. Hu, *IEEE Trans El. Dev.* Vol. 48, 880 (2001).
- [144] Bin Yu, Leland Chang, Shibly Ahmed, Haihong Wang, Scott Bell, Chih-Yuh Yang, Cyrus Tabery, Chau Ho, Qi Xiang, Tsu-Jae King, Jeffrey Bokor, Chenming Hu, Ming-Ren Lin, David Kyser, *IEDM Tech. Digest* (IEEE, Piscataway, NJ, 2002), pp. 251-254.
- [145] Y.Y. Chen, C.H. Chien, J.C. Lou, *Thin Solid Films*, Vol. 513, 264 (2006).
- [146] T. Mizuno, J. Okamura, A. Toriumi, *IEEE Trans El. Dev.* Vol. 41, 2216 (1994).
- [147] H. R. Khan, D. Vasileska and S. S. Ahmed, *J. Comp. El.* Vol. 3, 337 (2004).
- [148] H. Khan, D. Mamaluy and D. Vasileska, *IEEE Trans El. Dev.* (submitted May 2007).
- [149] W. Pötz, *J. Appl. Phys.*, Vol. 66, 2458 (1989).
- [150] H. L. Grubin and D. K. Ferry, *Semicond. Sci. Tech.*, Vol. 13 (Suppl. S), A44 (1998); D. K. Ferry *et al.*, *J. Vac. Sci. Technol. B*, Vol. 21, 1891 (2003).
- [151] G. Ferrari *et al.* *Semicond. Sci. Tech.*, Vol. 19, S254-S256 (2004).
- [152] A. Svizhenko and M. P. Anantram, *IEEE Trans. Electron. Dev.*, Vol. 50, 1459 (2003).
- [153] I. Knezevic and D. K. Ferry, *Superlatt. Microstruct. in press.*
- [154] W. J. Gross, D. Vasileska, and D. K. Ferry, *IEEE Trans. Electron. Dev.*, Vol. 47, 1831 (2000).

-
- [155] M. V. Fischetti and S. E. Laux, *Appl. Phys. Lett.*, Vol. 76, 2277 (2000); *J. Appl. Phys.*, Vol. 89, 1205 (2001).
- [156] H.-P. Breuer and F. Petruccione, *The Theory of Open Quantum Systems* (Oxford University Press, New York, 2002).
- [157] F. Rossi and T. Kuhn, *Rev. Mod. Phys.*, Vol. 74, 895 (2002).
- [158] V. M. Axt and T. Kuhn, *Rep. Prog. Phys.*, Vol. 67, 433 (2004).
- [159] *Nonequilibrium Physics on Short Time Scales*, edited by K. Morawetz (Springer, Berlin, 2004).
- [160] *Progress in Nonequilibrium Green's Functions*, edited by M. Bonitz (World Scientific, Singapore, 2000).
- [161] D. N. Zubarev, *Nonequilibrium Statistical Thermodynamics* (New York, Consultants Bureau, 1974).
- [162] D. H. E. Gross, in *Nonequilibrium Physics on Short Time Scales*, edited by K. Morawetz (Springer, Berlin, 2004), pp. 95-107.
- [163] S. Nakajima, *Prog. Theor. Phys.*, Vol. 20, 948 (1958).
- [164] R. Zwanzig, *J. Chem. Phys.*, Vol. 33, 1338 (1960).
- [165] H. Mori, *Prog. Theor. Phys.*, Vol. 33, 423 (1965).
- [166] P.N. Argyres and P.L. Kelley, *Phys. Rev.*, Vol. 134, A98 (1964).
- [167] J. R. Barker and D. K. Ferry, *Solid-State Electron.*, Vol. 23, 531 (1980).
- [168] K. Kassner, *Phys. Rev. A*, Vol. 36, 5381 (1987).
- [169] M. Sparpaglione and S. Mukamel, *J. Chem. Phys.*, Vol. 88, 3263 (1988).
- [170] Y. Hu and S. Mukamel, *J. Chem. Phys.*, Vol. 91, 6973 (1989).
- [171] V. Romero-Rochin and I. Oppenheim, *Physica A*, Vol. 155, 52 (1989).
- [172] M. Tokuyama and H. Mori, *Prog. Theor. Phys.*, Vol. 55, 411 (1975).
- [173] N. Hashitsume, F. Shibata and M. Shingu, *J. Stat. Phys.*, Vol. 17, 155 (1977).
- [174] F. Shibata, Y. Takahashi and N. Hashitsume, *J. Stat. Phys.*, Vol. 17, 171 (1977).
- [175] M. Saeki, *Prog. Theor. Phys.*, Vol. 67, 1313 (1982).
- [176] M. Saeki, *Prog. Theor. Phys.*, Vol. 79, 396 (1988).
- [177] M. Saeki, *Prog. Theor. Phys.*, Vol. 89, 607 (1993).
- [178] D. Ahn, *Phys. Rev. B*, Vol. 50, 8310 (1994).; Vol. 51, 2159 (1995).
- [179] D. Ahn, *Prog. Quantum Electron.*, Vol. 21, 249 (1997).
- [180] T. M. Chang and J. L. Skinner, *Physica A*, Vol. 193, 483 (1993).
- [181] D. Ahn, J. H. Oh, K. Kimm and S.W. Hwang, *Phys. Rev. A*, Vol. 61, 052310 (2000).
- [182] D. Ahn, J. Lee and S. W. Hwang, *Phys. Rev. A*, Vol. 66, 012302 (2002).
- [183] A. A. Golosov and D. R. Reichmann, *J. Chem. Phys.*, Vol. 115, 9849 (2001).
- [184] W. E. Boyce and R. C. DiPrima, *Elementary Differential Equations and Boundary Value Problems*, 4th ed. (Wiley, New York, 1986), p. 327.
- [185] I. Knezevic and D. K. Ferry, *Phys. Rev. E*, Vol. 66, 016131 (2002).
- [186] I. Knezevic and D. K. Ferry, *Phys. Rev. E*, Vol. 67, 066122 (2003).
- [187] W. H. Zurek, *Rev. Mod. Phys.*, Vol. 5, 715 (2003).
- [188] V. Vedral *et al.*, *Phys. Rev. Lett.*, Vol. 78, 2275 (1997).
- [189] D. Semkat, D. Kremp, and M. Bonitz, *Phys. Rev. E*, Vol. 59, 1557 (1999); *ibid*, *J. Math. Phys.*, Vol. 41, 7458 (2000); K. Morawetz *et al. Phys. Rev. E*, Vol. 63, 020102 (2001).
- [190] W. T. Reid, *Riccati Differential Equations* (Academic Press, New York, 1972); *The Riccati Equation*, edited by S. Bittanti, A. J. Laub, and J. C. Willems (Springer, Berlin, 1991).
- [191] I. Knezevic and D. K. Ferry, *Phys. Rev. A*, Vol. 69, 012104 (2004).
- [192] I. Knezevic, *unpublished*.
- [193] G. Lindblad, *J. Phys. A*, Vol. 29, 4197 (1996).
- [194] G. Lindblad, *Commun. Math. Phys.*, Vol. 48, 199 (1976).
- [195] M. A. Osman and D. K. Ferry, *Phys. Rev. B*, Vol. 36, 6018 (1987).
- [196] A.-P. Jauho, N. S. Wingreen, and Y. Meir, *Phys. Rev. B*, Vol. 50, 5528 (1994).
- [197] M. A. Talebian and W. Pötz, *Appl. Phys. Lett.*, Vol. 69, 1148 (1996).
- [198] R. Lake and S. Datta, *Phys. Rev. B*, Vol. 45, 670 (1992).
- [199] S. Datta and M. P. Anantram, *Phys. Rev. B*, Vol. 45, 13761 (1992).
- [200] S. Datta, *Phys. Rev. B*, Vol. 46, 9493 (1992).
- [201] I. Knezevic, *J. Comput. Theor. Nanosci.*, Vol. 4, 746 (2007).
- [202] M. Bonitz, *Quantum Kinetic Theory* (Teubner, Stuttgart, Leipzig, 1998).

-
- [203] I. Prigogine, *Non-Equilibrium Statistical Mechanics* (Interscience Publ., New York, 1963).
- [204] P. Resibois, *Physica*, Vol. 31, 645 (1965).
- [205] R. Zwanzig, *Phys. Rev.*, Vol. 129, 486 (1963).
- [206] R. Balescu, *Statistical Mechanics of Charged Particles* (Interscience Publ. New York, 1963)
- [207] T. Matsubara, *Prog. Theor. Phys.*, Vol. 14, 351 (1955).
- [208] F. Haake, *Phys. Rev. A*, Vol. 3, 1723 (1971).
- [209] I. Knezevic and D. K. Ferry, *Physica E*, Vol. 19, 71 (2003).
- [210] D. N. Zubarev, V. G. Morozov, and G. Röpke, *Statistical Mechanics of Nonequilibrium Processes*, Vol. 1 (Akademie Verlag, Berlin, 1996).
- [211] V. G. Morozov and G. Röpke, *Ann. Phys. (N.Y.)*, Vol. 278, 127 (1999).
- [212] V. Mujica, M. Kemp, A. Roitberg, and M. Ratner, *J. of Chem. Physics*, Vol. 104, 7296 (1996).
- [213] P. A. Derosa, J. M. Seminario, and P. B. Balbuena, *J. Phys. Chem. A*, Vol. 105, 7917 (2001)
- [214] A. Demkov and O. Sankey *Phys. Rev. Lett.*, Vol. 83, 2038 (1999).
- [215] A. Demkov, R. Liu, X. Zhang, and H. Loechelt, *Journal of Vacuum Science and Technology B*, Vol. 18, 2388 (2000).
- [216] A.A. Demkov, X. Zhang and D.A. Drabold, *Phys. Rev. B*, Vol. 64, 125306-1 (2001)
- [217] G. Klimeck, R. Lake, S. Datta, G. W. Bryant, *Phys. Rev. B*, Vol. 50, 5484 (1994).
- [218] G. Klimeck, G. Chen, S. Datta, *Phys. Rev. B*, Vol. 50, 2316 (1994).
- [219] G. Chen, G. Klimeck, S. Datta, G. Chen and W. A. Goddard, *Phys. Rev. B*, Vol. 50, 8035 (1994).
- [220] A. Canning, L.W. Wang, A. Williamson, A. Zunger, *J of Comp. Physics*, Vol. 160, 29 (2000).
- [221] L.W. Wang, J. N. Kim, A. Zunger, *Phys. Rev. B*, Vol. 59, 5678 (1999).
- [222] A.J. Williamson, L.W. Wang, A. Zunger, *Phys. Rev. B*, Vol. 62, 12963 (2000).
- [223] R. Martin, *Phys. Rev. B*, Vol. 1, 4005 (1970).
- [224] O. Sankey, D.J. Niklewski, *Phys. Rev. B*, Vol. 40, 3979 (1989).
- [225] P. Ordejón, D.A. Drabold, M.P. Grumbach, R.M. Martin, *Phys. Rev. B*, Vol. 48, 14646 (1993).
- [226] F. Ordejón, G. Galli, R. Car, *Phys. Rev. B*, Vol. 47, 9973 (1993).
- [227] P. Vogl, H.P. Hjalmarson, J.D. Dow, *J. Phys. Chem. Solids*, Vol. 44, 365 (1983).
- [228] J.M. Jancu, R. Scholz, F. Beltram, F. Bassani, *Phys. Rev. B*, Vol. 57, 6493 (1998).
- [229] R.C. Bowen, *IEDM 1997*, IEEE, New York, 869 (1997).
- [230] G. Klimeck, R. K. Lake, R. C. Bowen, C. L. Fernando and W. R. Frensley, *VLSI Design*, Vol. 6, p. 107 (1998).
- [231] S. Lee, L. Jönsson, J. W. Wilkins, G. Bryant, and G. Klimeck, *Phys. Rev. B*, Vol. 63, 195318 (2001).
- [232] C. Pryor, *Phys. Rev. B*, Vol. 57, 7190 (1998).
- [233] O. Stier, M. Grundmann, D. Bimberg, *Phys. Rev. B*, Vol. 59, 5688 (1999).
- [234] P. Brodono, M. Pascoli, R. Brunetti, A. Bertoni, and C. Jacoboni, *Phys. Rev. B*, Vol. 59, 3060 (1998).
- [235] M. V. Fischetti, *Phys. Rev. B*, Vol. 59, 4901 (1998).
- [236] A. Haque, A. N. Khondker, *J. Appl. Phys.*, Vol. 87, 2553 (2000).
- [237] D. Guan and U. Ravaioli, R.W. Giannetta, M. Hannan and I. Adesida, M. R. Melloch, *Phys. Rev. B*, Vol. 67, 205328 (2003).
- [238] Supriyo Datta, *Superlattices and Microstructures*, Vol. 28, 253 (2000).