QUANTUM AND COULOMB EFFECTS IN NANODEVICES[*]

D. VASILESKA, H.R. KHAN, S.S. AHMED[†]

IRA A. Fulton School of Engineering, Department of EE
Arizona State University, Tempe, AZ 85287-5706

C. RINGHOFER, C. HEITZINGER
Department of Mathematics
Arizona State University, Tempe, AZ 85287-5706

In state of the art devices, it is well known that quantum and Coulomb effects play significant role on the device operation. In this paper we demonstrate that a novel effective potential approach in conjunction with a Monte Carlo device simulation scheme can accurately capture the quantum-mechanical size quantization effects. We also demonstrate, via proper treatment of the short-range Coulomb interactions, that there will be significant variation in device design parameters for devices fabricated on the same chip due to the presence of unintentional dopant atoms at random locations within the channel.

Keywords: nanodevice modeling; quantum effects; discrete impurity effects.

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 with the fabrication of bipolar devices on slabs of polycrystalline germanium (Ge) [2]. 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/SiO_2 materials system, necessitating a new generation of crystal pullers with improved environmental controls to prevent SiO_2 formation. Later, the stability and low interface-state density of the Si/SiO_2 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 to lasers—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.

The 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.

1.1. *Device Scaling*

Device engineers are most concerned with the method of scaling introduced in the previous section. 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 [3] 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 [4], 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 metal-oxide-semiconductor field-effect transistor (MOSFET) from larger to smaller physical dimensions. Scaling of CMOS technology has progressed relentlessly from a line width of 1 μm to the current 90-nm line width. Two key features characterize this era. First, slavish devotion to scaling by constant improvements in lithography, as described by Dennard *et al*. [5]. 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 have 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.

Regarding conventional silicon MOSFETs, the device size is scaled in all dimensions, resulting in smaller oxide thickness, junction depth, channel length, channel width, and isolation spacing. Currently, 90 nm (with a physical gate length of 50 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 2016, when devices with 20 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 [5,6] and thus demonstrated that these feature sizes are feasible. An ultrasmall MOSFET with a channel length of 15 nm has been demonstrated in 2001 [6]. Conventional Silicon MOS transistors with physical gate length of 10 nm have recently been demonstrated by the Intel Corporation [7]. 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 already are 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 approaching 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 inno-

vations to circumvent barriers due to the fundamental physics that constrains the conventional MOSFET. The limits most often cited [8] 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; control of threshold voltage over the die is another major scaling challenge; (4) voltage-related effects such as subthreshold swing, built-in voltage and minimum logic voltage swing; (5) Short-channel effects (SCEs), such as drain-induced barrier lowering (DIBL) that degrade the device performance; (6) Hot carriers that degrade device reliability, and (7) 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. 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/cm^2. 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.

1.2. *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 [8] 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: new materials and 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 [9] and metal gates), band engineering methods (using strained Si [10,11,12] 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. 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 [13]. Intel has adopted strained silicon technology for its 90 nm process [14]. 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 65 nm node.

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. 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 [15], 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 [16] summarizes the advantages and challenges of some of the above-mentioned device structures.

Table 1. Non-classical CMOS devices.

Device	Ultrathin Body (UTB) SOI	Band-Engineered Transistor	Vertical Transistor	FinFET	Double-Gate Transistor
Concept	Fully-depleted SOI	SiGe or Strained Si Channel; 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); stacked NAND	
Scaling Issues	Si film thickness, gate stack; worse SCE than bulk CMOS	High mobility film thickness (SOI); gate stack; integratability	Si film thiness; 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		

1.3. *Nanoscale Device Simulation*

Standard sequence that one follows when modeling device structures of interest involves process simulation step that is followed by a device simulation and is finalized with a 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 transistor. 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. Simulation is playing key role in device development today.

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. 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.

The main components of semiconductor device simulation at any level of approximation are illustrated in Fig. 1 [17]. 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 Fig. 2 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.
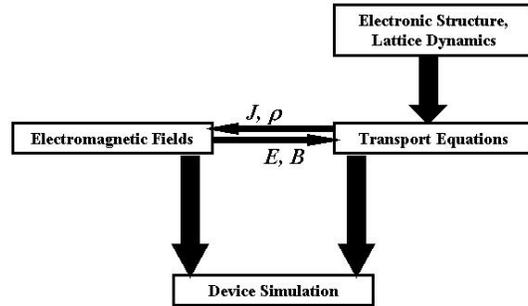
Fig. 1. A schematic description of the device simulation sequence.

Fig. 2. 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 semi-conductor 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. Table 2 highlights some of the basic physical effects that are viewed as important in nanoelectronics research [17].

Table 2. Important effects in ultrasmall electronics.

1. Transport Effects

   (a) Drift
   - Velocity overshoot
   - Ballistic transport
   - Oxide polar phonons decreasing channel mobility
   - Hot-electron effects (scattering in high electric field, injection into oxide)
   - Hot-phonon effects

   (b) Diffusion
   - Hot-electron diffusion (invalidation of Einstein relation)
   - Anisotropy of diffusion
   - Diffusion and reduced dimensionality

2. Size Effects
   - 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

The following section reviews the two *fundamental* problems encountered in modeling and simulations of current nanoscale Si MOSFETs: (1) the quantum effects and (2)

the effects associated with the proper treatment of the short-range and the long-range Coulomb potentials of electrons and impurities which are closely related to the discreteness of the carriers and impurities.

1.4. *Fundamental Problems in Nanoscale Device Modeling*

1.4.1. *Problems in Quantum Transport*

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 [18].

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. 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. 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. 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 [19]. 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. 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. The present work focuses mainly on the modeling and simulations of the size-quantization effect within a semiclassical transport framework.

On the other hand, the dynamical quantum effects in nanoscale MOSFETs, associated with energy dissipating scattering in electron transport are physically much more involved [20]. 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 (which is still much beyond the present computational power). 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 [21]. This is, of course, one of the most difficult problems to solve in quantum physics.

There is another fundamental problem associated with quantum transport. Since one is mainly concerned with devices operated at room temperature, phase-breaking 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, 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.

Table 3. Quantum Effects.

---

1. Static Quantum Effects
   - 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

---

2. Dynamical Quantum Effects
   - Collisional broadening
   - Intra-collisional field effects
   - Temperature dependence
   - Electron-electron scattering
   - Dynamical screening
   - Many-body effects
   - Pauli exclusion principle

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 effects, 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 3. 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.

1.4.2. *Problems with Accurate Treatment of the Coulomb Potential*

Statistical fluctuations of the channel dopant number were predicted by Keyes [22] as a fundamental physical limitation of MOSFET down-scaling. Entering into the nanometer regime results in a decreasing number of channel impurities whose random distribution leads to significant fluctuations of the threshold voltage and off-state leakage current. These effects are likely to induce serious problems on the operation and performances of logical and analog circuits. It has been experimentally verified by Mizuno and co-workers [23] that threshold voltage fluctuations are mainly caused by random fluctuations of the number of dopant atoms and that other contributions such as fluctuations of the oxide thickness are comparably very small. It follows from these remarks that impurities cannot be considered anymore using the continuum doping model in advanced semiconductor device modeling but the precise location of each individual impurity within a full Coulomb interaction picture must be taken into account.

With the down-scaling of MOSFET devices, electrons in the conducting channel are in ever closer proximity to the high-density electron gases present in the source and drain regions—separated from each other by as little as tens of nanometers—and in the polycrystalline Si gate—separated from the channel by as little as 1.5 nm of SiO_2. As studied in Ref. [24], the role of these long-range Coulomb interactions is twofold: (1) The interaction between electrons in the channel and the high-density electron gases in the source and drain regions can be pictured classically as a reshaping of the electron distribution in the channel caused by the potential-fluctuations, associated with plasma oscillations in the source and drain regions, *leaking* into the low-density channel. Quantum-mechanically, this corresponds to emission and absorption of plasmons by the channel electrons. While these processes do not subtract directly momentum from the electron gas, their net effect is a *thermalization* of the hot electrons energy distribution in the channel, the resulting higher energy tail being affected by additional momentum-relaxation processes (phonons, ionized impurities). This causes, *indirectly*, a reduction of the effective electron velocity in the channel, and so, a depression of the transconductance as the channel length is reduced below about 4 nm. Such a disappointing behavior of aggressively scaled devices has been recently emphasized by the MIT group. (2) On the other hand, the interaction between channel-electrons and electrons in the gate (*Coulomb drag* across the very thin insulator) results in a *direct* loss of momentum of the electrons in the channel. Semiclassically, this interaction—also plasmon-mediated—has been studied by a group at IBM predicting a significant depression of the electron velocity for SiO_2 layers thinner than about 2–3 nm. This behavior has also been observed experimentally [25],

recent results being in quantitative agreement with early theoretical estimates [24]. The conclusion one can draw from these results is that shrinking devices in the nanoscale regimes is not likely to result in much performance gain as often predicted by the researchers. In other words, the *ballistic limit* may remain a pipe dream: shorter channels are required for ballistic transport to occur. But as this is done, while also scaling the gate-insulator thickness, unavoidable Coulomb interactions gain strength, killing the concept of *ballistic transport* at its very onset.

In the past, the effect of discrete dopant random distribution in MOSFET channel has been assessed by analytical or drift-diffusion (DD) approaches. The first DD study consisted in using a stochastically fluctuating dopant distribution obeying Poisson statistics [26]. 3D *atomistic* simulators have also been developed for studying threshold voltage fluctuations [27,28]. Even though the DD/HD methods are very useful because of their simplicity and fast computing times, it is not at all clear whether such macroscopic simulation schemes can be exploited into the atomistic regime. In fact, it is not at all clear how such discrete electrons and impurities are modeled in macroscopic device simulations due to the long-range nature of the Coulomb potential.

Three-dimensional (3D) Monte Carlo (MC) simulations should provide a more realistic transport description in ultra-short MOSFET. The MC procedure gives an exact solution of the Boltzmann transport equation. It, thus, correctly describes the nonstationary transport conditions. Even if microscopic simulations such as the MC method are concerned, the treatment of the electrons and impurities is not straightforward due to, again, the long-range nature of the Coulomb potential. The incorporation of the long-range Coulomb potential in the MC method has been a long-standing issue [29,30]. This problem is, in general, avoided by assuming that the electrons and the impurities are always screened by the other carriers so that the long-range part of the Coulomb interaction is effectively suppressed. The complexity of the MC simulation increases as one takes into account more complicated screening processes by using the dynamical and wave-vector dependent dielectric function obtained from, for example, the random phase approximation. However, the screening is a very complicated many-body matter [31].

This situation is also complicated in the MC *device simulations* in which the BTE is self-consistently coupled with the Poisson equation [32]. The Coulomb potential due to electrons and impurities is then separated into the long-range and the short-range parts. The long-range part is taken into account by the solution of the Poisson equation, whereas the short-range part is usually included in the BTE through the scattering kernel. In other words, the Coulomb potential is separated into the long-range and short-range parts by the size of the mesh employed in the Poisson equation. However, the choice of the mesh size is not trivial. For example, the mesh cannot be arbitrarily small, otherwise the Coulomb potential would be double-counted by the Poisson equation and the BTE. Since the long-range part of the Coulomb potential is responsible for the many-body effects, the mesh size has to be determined consistently with, say, the renormalized electron (kinetic) energy calculated from the many-body theory [33]. This is, of course, not an easy task, especially, for the case of small device structures. On the other hand, since the size of localized electrons in the MC device simulations is roughly given by the size of the mesh, this is not consistent with the concept of the electron wave packet. The BTE (or equivalently, the microscopic simulation) assumes that the electrons are localized and described by the wave packet whose size is comparable to the de Broglie wavelength. However, the size of the active device region is now comparable with the size of the

wave packet in nanoscale MOSFETs and so, it is not clear how the localized electrons in the channel should be interpreted in such microscopic simulations.

2. The Role of Quantum-Mechanical Space Quantization Effects on Nano-Scale Devices Operation

Successful scaling of MOSFETs towards shorter channel lengths requires thinner gate oxides and higher doping levels to achieve high drive currents and minimized short-channel effects [34,35]. For these nanometer devices it was demonstrated a long time ago that, as the oxide thickness is scaled to 10 nm and below, the total gate capacitance is smaller than the oxide capacitance due to the comparable values of the oxide and the inversion layer capacitances. As a consequence, the device transconductance is degraded relative to the expectations of the scaling theory [36]. The inversion layer capacitance was also identified as being the main cause of the second-order thickness dependence of MOSFET's *IV*-characteristics [37]. The finite inversion layer thickness was estimated experimentally by Hartstein and Albert [38]. The high levels of substrate doping, needed in nano-devices to prevent the punch-through effect, that lead to quasi-two-dimensional (Q2D) nature of the carrier transport, were found responsible for the increased threshold voltage and decreased channel mobility, and a simple analytical model that accounts for this effect was proposed by van Dort and co-workers [39,40]. Later on, Vasileska and Ferry [41] confirmed these findings by investigating the doping dependence of the threshold voltage in MOS capacitors. The experimental data for the doping dependence of the threshold voltage shift and our simulation results from Ref. [41] are shown in Fig. 3.
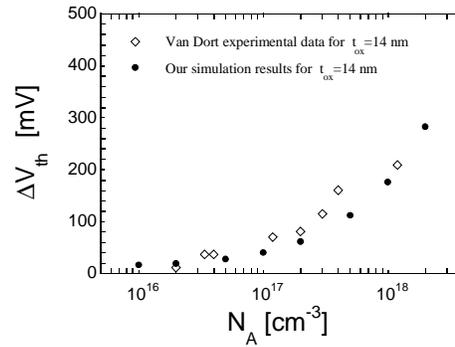

Fig. 3. SCHRED simulation data for the shift in the threshold voltage compared to the experimental values provided by van Dort and co-workers.

These results clearly demonstrate the influence of quantum-effects on the operation of nano-scale MOSFETs in both the off- and the on-state. The two physical origins of the inversion layer capacitance, due to the finite density of states and due to the finite inversion layer thickness, were demonstrated experimentally by Takagi and Toriumi [42]. A computationally efficient three-subband model, that predicts both the quantum-mechanical effects in the electron inversion layers and the electron distribution within the

inversion layer, was proposed and implemented into the PICSEC simulator [43]. The influence of the image and many-body exchange-correlation effects on the inversion layer and the total gate capacitance was studied by Vasileska *et al.* [44]. It was also pointed out that the depletion of the poly-silicon gates considerably affects the magnitude of the total gate capacitance [45].

The above examples outline the advances during the two decades of research on the influence of quantum-effects on the operation on nano-devices. The conclusion is that any state-of-the-art device simulator must take into consideration the quantum-mechanical nature of the carrier transport and the poly-depletion effects to correctly predict the device off- and on-state behavior. As noted by many of these authors, to account for the quantum-mechanical effects, one in principle has to solve the 2D/3D Schrödinger-Poisson problem in conjunction with an appropriate transport kernel. (For devices in which velocity overshoot is strongly pronounced, minimum that one can do is to solve the Boltzmann transport equation using the Ensemble Monte Carlo (EMC) technique.) Since the exact solution of the 2D/3D Schrödinger-Poisson problem is time-consuming even with present state-of-the-art computers, alternative paths have been sought for device simulators that utilize the effective potential approach.

The idea of quantum potentials originates from the hydrodynamic formulation of quantum mechanics, first introduced by de Broglie and Madelung [46,47,48], and later developed by Bohm [49,50]. In this picture, the wave function is written in complex form in terms of its amplitude $R(\mathbf{r},t)$ and phase $\psi(\mathbf{r},t) = R(\mathbf{r},t)\exp[iS(\mathbf{r},t)/\mathrm{h}]$. These are then substituted back into the Schrödinger equation to obtain the following coupled equations of motion for the density and phase

$$\frac{\partial \rho(\mathbf{r},t)}{\partial t} + \nabla \cdot \left(\rho(\mathbf{r},t)\frac{1}{m}\nabla S(\mathbf{r},t) \right) = 0 \, , \tag{1}$$

$$-\frac{\partial S(\mathbf{r},t)}{\partial t} = \frac{1}{2m}\left[\nabla S(\mathbf{r},t) \right]^2 + V(\mathbf{r},t) + Q(\rho,\mathbf{r},t) \, , \tag{2}$$

where $\rho(\mathbf{r},t) = R^2(\mathbf{r},t)$ is the probability density. By identifying the velocity as $\mathbf{v} = \nabla S / m$, and the flux as $\mathbf{j} = \rho\mathbf{v}$, Eq. (1) becomes the continuity equation. Hence, Eqs. (1) and (2) arising from this so-called *Madelung transformation* to the Schrödinger equation have the form of classical hydrodynamic equations with the addition of an extra potential, often referred to as the *quantum* or *Bohm potential*, written as

$$V_Q = -\frac{\mathrm{h}^2}{2mR}\nabla^2 R \rightarrow -\frac{\mathrm{h}^2}{2m\sqrt{n}}\nabla^2\sqrt{n} \tag{3}$$

where the density *n*, is related to the probability density as $n(\mathbf{r},t) = N\rho(\mathbf{r},t) = NR^2(\mathbf{r},t)$, where *N* is the total number of particles. The Bohm potential essentially represents a field through which the particle interacts with itself. It has been used, for example, in the study of wave packet tunneling through barriers [51], where the effect of the quantum potential is shown to lower or smoothen barriers, and hence *allow* for the particles to leak through.

An alternate form of the quantum potential was proposed by Iafrate, Grubin and Ferry [52], who derived a form of the quantum potential based on moments of the

Wigner-Boltzmann equation, the kinetic equation describing the time evolution of the Wigner distribution function [53]. Their form, based on moments of the Wigner function in the pure state, and involving an expansion of order $O(\eta^2)$, is given by

$$V_Q = -\frac{h^2}{8m}\nabla^2\left(\ln n\right),\tag{4}$$

which is sometimes referred to as the *Wigner potential*, or as the density gradient correction. Such quantum potentials have been extensively used in *density-gradient* and *quantum-hydrodynamic* methods. Their use in particle-based simulation schemes becomes questionable due to the presence of statistical noise in the representation of the electron density and the considerable difficulty to calculate the second derivative of the density on a completely unstructured mesh given by the particle discretization.

To avoid this problem, Ferry and Zhou derived a form for a smooth quantum potential [54], based on the effective classical partition function of Feynman and Kleinert [55]. More recently, Gardner and Ringhofer [56] derived a smooth quantum potential for hydrodynamic modeling, valid to all orders of $\hbar^2$, which involves a smoothing integration of the classical potential over space and temperature. There, it was shown that, close to the equilibrium regime, the influence of the potential on the ensemble can be replaced by the classical influence of a smoothed non-local barrier potential. While this effective potential depends non-locally on the density, it does not directly depend on its derivatives. Through this effective quantum potential, the influence of the barriers on an electron is felt at quite some distance from the barrier. The smoothed effective quantum potential has been used successfully in quantum-hydrodynamic simulations of resonant tunneling effects in one dimensional double-barrier structures [57].

In analogy to the smoothed potential representations discussed above for the quantum hydrodynamic models, it is desirable to define a smooth quantum potential for use in quantum particle-based simulations. Ferry [59] has suggested an *effective potential scheme* that emerges from a wave packet description of the particle motion, where the extent of the wave packet spread is obtained from the range of wavevectors in the thermal distribution function (characterized by an electron temperature). The effective potential, V_{eff}, is related to the self-consistent Hartree potential, V, obtained from the Poisson equation, through an integral smoothing relation

$$V_{\mathit{eff}}(\mathbf{x}) = \int V(\mathbf{x}+\mathbf{y})\,G(\mathbf{y},a_0)\,d\mathbf{y}\,,\tag{5}$$

where G is a Gaussian with standard deviation a_0. The effective potential V_{eff} is then used to calculate the electric field that accelerates the carriers in the transport kernel of the Monte Carlo particle-based device simulator discussed in Ref. [58]. The calculation of V_{eff} has a fairly low computational cost, but the requirement that the electric field is updated every 0.01 fs to get physically accurate particle trajectories and to eliminate the artificial heating of the carriers in the vicinity of the Si/SiO_2 interface (where the fields are the strongest), adds to the computational cost. Note also that within this approach the parameter a_0 has to be adjusted in the initial stages of the simulation via comparisons of the sheet/line density of the Q2D/Q1D structure being investigated using the effective potential approach and the 1D/2D Schrödinger-Poisson simulations.

In this review paper, in addition to the effective potential approach due to Ferry [59], we present a new form of the effective quantum potential for use in Monte Carlo device

simulators. The proposed approach is based on perturbation theory around thermodynamic equilibrium and leads to an effective potential which depends on the energy and wavevector of each individual electron, thus effectively lowering step-function barriers for high-energy carriers [60]. The quantum potential is derived from the idea that the Wigner and the Boltzmann equation with the quantum corrected potential should possess the same steady state. The resultant quantum potential is, in general, two-degrees smoother than the original Coulomb and barrier potentials, i.e. possesses two more classical derivatives, which essentially eliminates the problem of statistical noise. The computation of the quantum potential involves only the evaluation of pseudo-differential operators and can, therefore, be effectively facilitated using Fast Fourier Transform (FFT) algorithms. The approach is quite general and can easily be modified to modeling of, for example, triangular quantum wells. The above-described approach has been used in simulation of 25 nm MOSFET device with oxide thickness of 1.2 nm.

2.1. *Thermodynamic Effective Potential*

The basic idea of the thermodynamic approach to effective quantum potentials is that the resulting semiclassical transport picture should yield the correct thermalized equilibrium quantum state. Using quantum potentials, one generally replaces the quantum Liouville equation

$$\partial_t \rho + \frac{i}{\hbar}[H, \rho] = 0 \tag{6}$$

for the density matrix $\rho(x,y)$ by the classical Liouville equation

$$\partial_t f + \frac{\hbar}{2m^*} k \cdot \nabla_x f - \frac{1}{\hbar} \nabla_x V \cdot \nabla_k f = 0 \ , \tag{7}$$

for the classical density function $f(x,k)$. Here, the relation between the density matrix and the density function f is given by the Weyl quantization

$$f(x,k) = W[\rho] = \int \rho(x + y/2, x - y/2) \exp(ik \cdot y) dy . \tag{8}$$

The thermal equilibrium density matrix in the quantum mechanical setting is given by $\rho^{eq} = e^{-\beta H}$, where $\beta = 1/k_B T$ is the inverse energy, and the exponential is understood as a matrix exponential, i.e. $\rho^{eq}(x,y) = \sum_\lambda \psi_\lambda(x) \exp(-\beta\lambda)\psi_\lambda(y)^*$ holds, with $\{\psi_\lambda\}$ the orthonormal eigensystem of the Hamiltonian H. In the semiclassical transport picture, on the other hand, the thermodynamic equilibrium density function f_{eq} is given by the Maxwellian $f_{eq}(x,k) = \exp\left(-\frac{\beta\hbar^2|k|^2}{2m^*} - \beta V\right)$. Consequently, to obtain the quantum mechanically correct equilibrium states in the semiclassical Liouville equation with the effective quantum potential V^Q, we set

$$f_{eq}(x,k) = \exp\left(-\frac{\beta\hbar^2|k|^2}{2m^*} - \beta V^Q\right) = W[\rho^{eq}] = \int e^{-\beta H} \rho(x + y/2, x - y/2) \exp(ik \cdot y) dy . \tag{9}$$

This basic concept was originally introduced by Feynman and Kleinert [55]. Different forms of the effective quantum potential arise from different approaches to approximate the matrix exponential $e^{-\beta H}$.

In the approach presented in this paper, we represent $e^{-\beta H}$ as the Green's function of the semigroup generated by the exponential. Introducing an artificial dimensionless parameter γ and defining $\rho(x,y,\gamma) = \sum_\lambda \psi_\lambda(x)\exp(-\gamma\beta\lambda)\psi_\lambda(y)^*$, we obtain a heat equation for ρ by differentiating ρ w.r.t. γ and using the eigenfunction property of the wave functions ψ_λ. This heat equation is referred to as the Bloch equation

$$\partial_\gamma \rho = -\frac{\beta}{2}\left(H\cdot\rho+\rho\cdot H\right), \quad \rho(x,y,\gamma=0)=\delta(x-y) \; , \tag{10}$$

and $\rho^{eq}(x,y)$ is given by $\rho(x,y,\gamma=1)$. Under the Weyl quantization this becomes, with the usual Hamiltonian $H = -\frac{h^2}{2m^*}\Delta_x + V$ and defining the effective energy E by $f = W[\rho]=e^{-\beta E}$,

$$\partial_\gamma E = \frac{\beta h^2}{8m^*}\left(\Delta_x E - \beta|\nabla_x E|^2\right) + \frac{h^2|k|^2}{2m^*} + $$
$$\frac{1}{2(2\pi)^3}\sum_{\nu=\pm 1}\int V(x+\nu y/2)\exp\left[\beta E(x,k,\gamma)-\beta E(x,q,\gamma)+iy(k-q)\right]dqdy \; , \tag{11}$$
$$E(x,k,\gamma=0)=0.$$

The effective quantum potential is in this formulation given by $E(x,k,\gamma=1) = V^Q + \frac{h^2|k|^2}{2m^*}$. The logarithmic Bloch equation is now solved 'asymptotically', using the *Born approximation*, i.e. by iteratively inverting the highest order differential operator (the Laplacian). This involves successive solution of a heat equation for which the Green's function is well known, giving (see Ref. [61] for the details)

$$V^Q(x,k) = \frac{1}{(2\pi)^3}\int \frac{2m^*}{\beta h^2 k\cdot\xi}\sinh\left(\frac{\beta h^2 k\cdot\xi}{2m^*}\right)\exp\left(-\frac{\beta h^2}{8m^*}|\xi|^2\right)V(y)e^{i\xi\cdot(x-y)}dyd\xi \; . \tag{12}$$

Note that the effective quantum potential V^Q now depends on the wave vector k. For electrons at rest, i.e. for $k = 0$, the effective potential V^Q reduces to the Gaussian smoothing given in Eq. (5) and Ref. [59]. Also note that there are no fitting parameters in this approach, i.e. the size of the wavepacket is determined by the particle's energy.

The potential $V(y)$ that appears in the integral of Eq. (12) can be represented as a sum of two potentials: the barrier potential $V_B(x)$, which takes into account the discontinuity at the Si/SiO$_2$ interface due to the difference in the semiconductor and the oxide affinities, and the Hartree potential $V_H(x)$ that results from the solution of the Poisson equation. Note that the barrier potential is 1D and independent of time and needs to be computed only once in the initialization stage of the code. On the other hand, the Hartree potential is 2D and time-dependent as it describes the evolution of charge from quasi-equilibrium to a non-equilibrium state. Since the evaluation of the effective Hartree potential, as given by Eq. (12), is very time consuming and CPU intensive, approximate solution methods have been pursued to resolve this term within a certain level of error tolerance.

We recall from the above discussion that the barrier potential is just a step-function. Under these circumstances $e\nabla_x V_B(x) = B(1,0,0)^T \delta(x_1)$, where B is the barrier height (on the order of 3.2 eV) and x_1 is a vector perpendicular to the interface. We actually need

only the gradient of the potential so that, using the pseudo-differential operators (see Appendix A), we compute

$$\nabla_x V_B^Q(x,p) = \exp\left[\frac{\beta h^2 |\nabla_x|^2}{8m^*}\right] \frac{2m^* \sin\left(\frac{\beta h p \cdot \nabla_x}{2m^*}\right)}{\beta h p \cdot \nabla_x} \nabla_x V_B(x) \qquad (13)$$

This gives

$$e\nabla_x V_B^Q(x,p) = \frac{B}{2\pi}(1,0,0)^T \int \exp\left[-\beta \frac{h^2 |\xi_1|^2}{8m^*}\right] \frac{2m^* \sinh\left(\frac{\beta h p_1 \cdot \xi_1}{2m^*}\right)}{\beta h p_1 \cdot \xi_1} e^{i\xi_1 \cdot x_1} d\xi_1 \qquad (14)$$

Note that V_B^Q is only a function of (x_1, p_1), i.e. it remains to be strictly one-dimensional, where x_1 and p_1 are the position and the momentum vector perpendicular to the interface. This, when combined with the fact that we have to calculate this integral only once, is a reason why we have decided to tabulate the result given by Eq. (14) on a mesh

The Hartree potential, as computed by solving the d-dimensional Poisson equation, depends in general upon d particle coordinates. For example, on a rectangular mesh the 2D Hartree potential is given by $V_H(x_1,x_2,t)$, and one has to evaluate $V_H^Q(x_1, x_2, p_1, p_2, t)$ using Eq. (12) N times each time step for all particles position and momenta: $x^n, p^n, n = 1,$ $\dots, N$ (where N is the number of electrons, which is large). This is, of course, an impossible task to be accomplished in finite time on present state-of-the-art computers. We, therefore, suggest the following scheme. According to (12), we evaluate the quantum potential by multiplying the Hartree potential by a function of $h\nabla_x$, or by multiplying the Fourier transform of the Hartree potential by a function of $h\xi$. We factor the expression in Eq. (12) into

$$V_H^Q(x,k) = \frac{2im^*}{\beta h^2 k \cdot \nabla_x} \sinh\left(\frac{\beta h^2 k \cdot \nabla_x}{2im^*}\right) \exp\left(\frac{\beta h^2}{8m^*}|\nabla_x|^2\right) V_H(x) = \frac{2im^*}{\beta h^2 k \cdot \nabla_x} \sinh\left(\frac{\beta h^2 k \cdot \nabla_x}{2im^*}\right) V_H^0(x), \qquad (15)$$

with

$$V_H^0(x) = \exp\left(\frac{\beta h^2}{8m^*}|\nabla_x|^2\right) V_H(x) \quad . \qquad (16)$$

The evaluation of the potential $V_H^0(x)$, which is a version of the Gaussian smoothed potential due to Ferry [59], is computationally inexpensive since it does not depend on the wavevector k. On the other hand, because of the Gaussian smoothing, $V_H^0(x)$ will be a smooth function of position, even if the Hartree potential $V_H(x)$ is computed via the Poisson equation where the electron density is given by a particle discretization. Therefore, the Fourier transform of the potential $V_H^0(x)$ will decay rapidly as a function of ξ, and it is admissible to use a Taylor expansion for small values of $h\xi$ in the rest of the operator. This gives

$$\frac{2im*}{\beta h^2 k \cdot \nabla_x} \sinh\left(\frac{\beta h^2 k \cdot \nabla_x}{2im*}\right) \approx 1 - \frac{\beta^2 h^4 \left(k \cdot \nabla_x\right)^2}{24\left(m*\right)^2} \quad , \tag{17}$$

or

$$\partial_{x_r} V_H^Q(x^n, p^n) = \partial_{x_r} V_H^0(x^n) - \frac{\beta^2 h^2}{24 m*^2} \sum_{j,k=1}^{2} p_j^n p_k^n \partial x_j \partial x_k \partial x_r V_H^0(x^n), \quad n = 1, K, N \tag{18}$$

for all particles. This is done simply by numerical differentiation of the sufficiently smooth grid function V_H^0 and interpolation. The evaluation of (18) is the price we have to pay when we compare the computational cost of this approach as opposed to the Ferry approach [59] which uses simple forward, backward or centered difference scheme for the calculation of the electric field. However, with this novel effective potential approach we avoid the use of adjustable parameters.

2.2.1. *Quantum Effects in a Conventional 25 nm MOSFET*

The parameters of the device structure being simulated are as follows: the average channel/substrate doping equals 10^{19} cm^{-3}, the doping of the source and drain regions is 10^{19} cm^{-3}, the junction depth is 30 nm, the oxide thickness is 1.2 nm and the gates are assumed to be metal gates with work-function equal to the semiconductor affinity. The gate/channel length is 25 nm. First in Fig. 4, the carrier confinement within the triangular potential well with and without the inclusion of the quantum-mechanical size-quantization effects is shown for the bias conditions $V_G = V_D = 1$V. From the results shown in this figure, it is evident that the low-energy electrons are displaced little more than the high-energy electrons; the reason being the fact that the high-energy electrons tend to behave as classical particles and hence are displaced relatively less. Also note that there is practically no carrier heating for the case when the effective potential is used in calculating the driving electric field. The carrier displacement from the interface proper is also seen from the results presented in Fig. 5. Notice that there is approximately 2 nm average shift of the electron density distribution near the source end of the channel when quantization effects are included in the model.
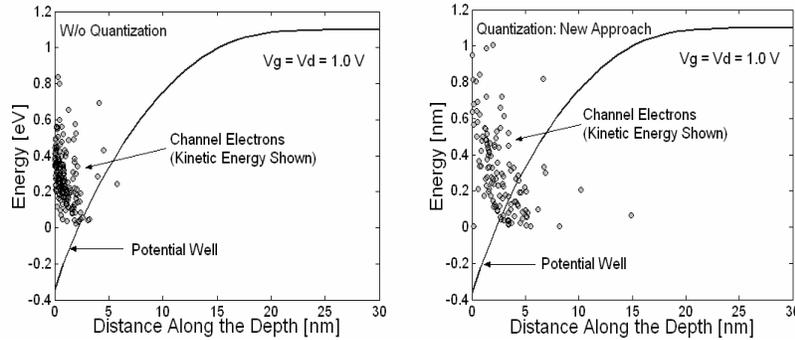

Fig. 4. Electron localization within the triangular potential barrier for the case when quantization effects are not included in the model (top panel) and for the case when we include quantum-mechanical space-quantization effects by using the effective potential approach presented in this paper (bottom panel).
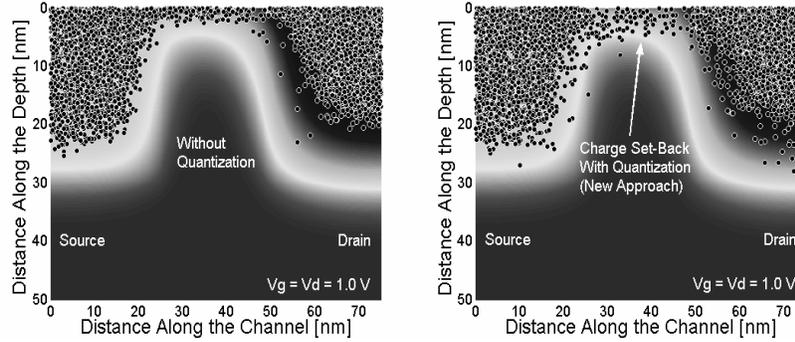
Fig. 5. Electron distribution in the device without (top panel) and with (bottom panel) the incorporation of quantum-mechanical size-quantization effects.
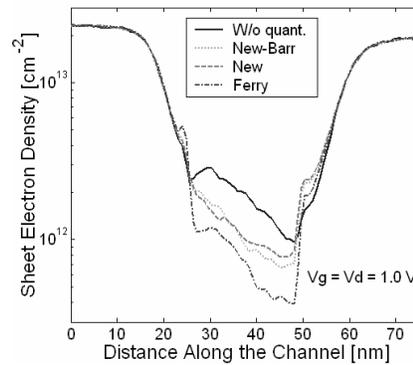

Fig. 6. Variation of the sheet electron density along the channel. *New-bar* corresponds to the case when we only include the influence of the barrier field. *New* represents the case when we include both the barrier and the Hartree contributions to the total electric field.

Also note that carriers behave more like bulk carriers at the drain end of the channel and are displaced in the same manner when using both the classical and the quantum-mechanical model.

The channel length variation of the sheet electron density is shown in Fig. 6 for classical, fully-quantum ($V_H^Q + V_B^Q$) and quantum-barrier field (V_B^Q) models [62]. Also compared are the simulation results for the sheet electron density from the new method with those utilizing the approach due to Ferry. There are several noteworthy features to be observed in these figures. First, the pinch-off of the sheet electron density near the drain end of the channel is evident in all models used. Second, the barrier and the full-effective potential scheme give almost the same value for the sheet electron density, which suggests that the repulsive barrier field dominates over the attractive field due to the Hartree potential. Third, the method due to Ferry leads to significantly lower value for the sheet electron density which can be improved by choosing lower values of the Gaussian smoothing parameter.

The average electron velocity and the average electron energy are shown in the left and the right panels of Fig. 7, respectively. Comparing the results for the average carrier energy on the right panel, one can see that the data for the case when one has not included the effective potential and the case when one has used the new model for the effective potential agree very well with each other. The slight increase in the carrier energy in the channel region, which is non-physical, when one uses the new effective potential approach is because of the very high value of the quantum field being present in the vicinity of the Si/SiO_2 interface proper. The situation can be improved by using a sufficiently small time-step (for example 0.01 fs) during Monte Carlo simulation. The approach due to Ferry gives significantly lower value for the carrier energy near the source end of the channel which has been explained to be due to the bandgap widening effect. Also, here we do not observe the non-physical carrier heating because of the fact that Ferry's effective potential is calculated from the very mesh potential which depends on both the meshing and the Gaussian parameter used in the model. The quantum field, calculated from direct differentiation of the effective mesh-potential, which has every possibility of being underestimated due to the finite size of the meshing used in simulations, also is independent on carrier energy (according to the current implementation of the model). When one confronts these data with the results for the average electron velocity, one observes that in the low-energy region near the source end of the channel the velocity is almost the same for all cases considered. At the drain end, one finds degradation of the velocity due to the smearing introduced by the quantum potential. Again, the inclusion of the barrier field and of the quantum-corrected Hartree term give similar values, which suggests that for the device being considered in this study only the barrier field has significant impact [63].
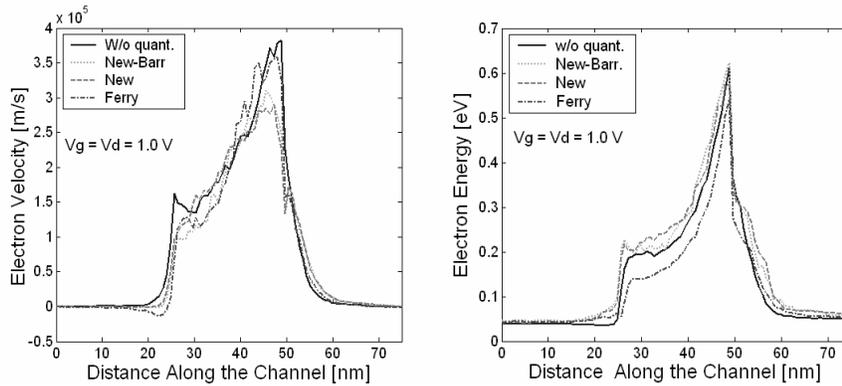

Fig. 7. Average electron velocity (left panel) and average electron energy (rightpanel) variation along the channel.

The device transfer characteristics are shown in the left panel of Fig. 8. Again, it becomes clear that the proposed full quantum potential and the barrier potential give similar values for the current. Looking more in detail the device transfer characteristics one finds that the quantization effects lead to threshold voltage increase of about 220 mV. When

properly adjusted for the oxide thickness difference, this result is consistent with previously published data [39]. Evidently, as deduced from the output characteristics shown in the right panel of Fig. 8, the shift in the threshold voltage leads to a decrease in the on-state current by 30%. The later observation confirms earlier findings that one must include quantum effects into the theoretical model to be able to properly predict the device threshold voltage and its on-state current.
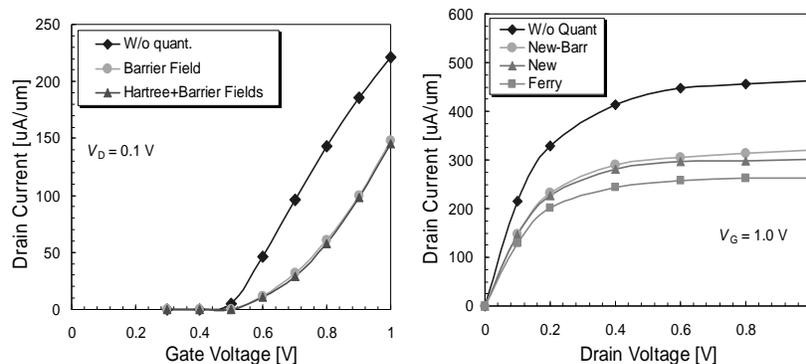

Fig. 8. Device transfer characteristic for $V_D = 0.1$ V (left panel). Device output characteristics VG = 1.0 V (right panel).

Next, the simulation results of a 15 nm conventional *n*-channel MOSFET device are discussed. Similar devices have been fabricated by Intel Corporation [7]. The physical gate length of the device used is 15 nm. The source/drain length equals 15 nm and the junction depth is also 15 nm. The bulk substrate thickness used for simulations is 45 nm. The height of the fabricated polysilicon gate electrode for this device is 25 nm. The gate oxide used was SiO_2 with physical thickness of only 0.8 nm. The source/drain doping density is 2×10^{19} cm^{-3} and the channel doping is 1.5×10^{19} cm^{-3}. The substrate doping used is 1×10^{18} cm^{-3}. The simulated device output characteristics are shown in Fig. 9.
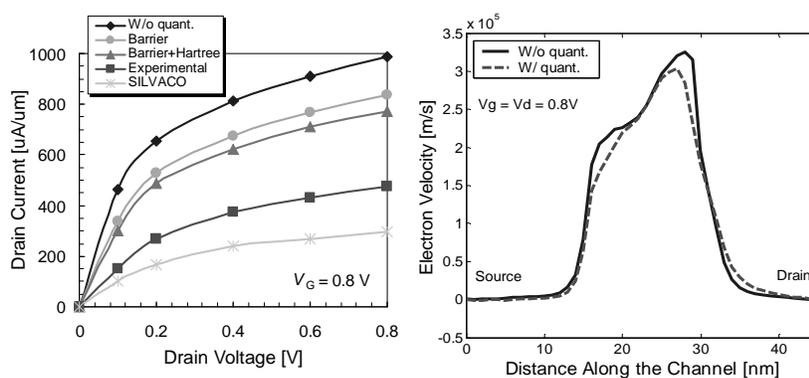

Fig. 9. Left panel: Conventional 15 nm MOSFET device output characteristics. Right panel: Average electron velocity along the channel.

There are again several noteworthy features in these results: (1) Quantum-mechanical size quantization increases the threshold voltage as observed from the decrease in the slope in the linear region and hence degrades the device transconductance. (2) Drain current degradation due to the quantum effects is not uniform rather decreases with the increase in drain bias. The reason may be attributed again to the fact that the electrons tend to behave as classical particles as average carrier energy increases with the increase in drain bias, (3) there is a considerable difference between the barrier-correction and the barrier-Hartree (full) correction which is mainly due to the use of higher doping density (1.5×10^{19} cm^{-3}) in the channel region than was used in the 25 nm MOSFET (1×10^{19} cm^{-3}) case. The higher doping density has a direct impact on the Hartree potential making the triangular channel potential steeper and hence introducing a pronounced quantum effects. But the overall degradation of the drain current as compared to the 25 nm MOSFET device structure has reduced in the 15 nm device because of the ballistic nature of the carrier motion in the latter case. This fact becomes clear if one observes the velocity profile of the device as depicted in the right panel of Fig. 9. What is important in this figure is that the carriers attain a velocity which is comparable to that in the 25 nm device structure even with a lesser biases applied i.e. $V_{\mathrm{G}} = V_{\mathrm{D}} = 0.8$V. Also, the gate oxide thickness is lesser in the 10 nm device which means that the gate oxide capacitance constitutes the major portion of the total effective gate capacitance thereby reducing the impact of the quantum capacitance. (4) The discrepancy between the experimental and the simulated results is attributed mainly to two reasons: (a) the series resistance coming from the finite width of the actual device structure and the contact resistances, and (b) the gate polysilicon depletion effects which as previously mentioned can introduce further degradation of the drain current on the order of 10-30% depending on the doping density and the height of the polysilicon gate used. The limited data as supplied by the Intel Corporation shows that the polysilicon gate is of 25 nm height which can indeed contribute to a significant degradation of the drain current. (4) The use of a commercial simulator like the drift-diffusion based SILVACO Atlas fails considerably to predict the device behavior mainly because of the ballistic and quantized nature of the carriers in these nanoscale device structures.

2.2.2. Size-quantization in Nanoscale SOI Devices

Because of using lightly/nearly undoped channel region, size-quantization effects in nanoscale fully-depleted SOI devices find a major source in the very physical nature of the confined region which remains sandwiched between the two oxide layers. In order to verify the applicability of the quantum potential approach developed in this work, a single gated SOI device structure will be studied first. Simulations will be carried out to calculate the threshold voltage as a function of the silicon film thickness and the results will be compared to other available methods. The SOI device used here has the following specifications: gate length is 40 nm, the source/drain length is 50 nm each, the gate oxide thickness is 7 nm with a 2 nm source/drain overlap, the box oxide thickness is 200 nm, the channel doping is uniform at 1×10^{17} cm^{-3}, the doping of the source/drain regions equals 2×10^{19} cm^{-3}, and the gate is assumed to be a metal gate with workfunction equal to the semiconductor affinity. There is a 10 nm spacer region between the gate and the source/drain contacts. The silicon (SOI) film thickness is varied over a range of 1–10 nm for the different simulations that were performed to capture the trend in the variations of

the device threshold voltage. Similar experiments were performed in Refs. [64,65] using the Schrödinger-Poisson solver and Ferry's effective potential approaches, respectively. For comparison purposes, threshold voltage is extracted from the channel inversion density versus gate bias profile and extrapolating the linear region of the characteristics to a zero value. This method also corresponds well to the linear extrapolation technique using the drain current-gate voltage characteristics.
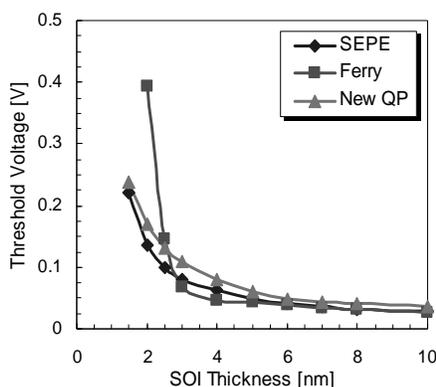

Fig. 10. Threshold voltage variation with SOI film thickness.

The results showing the trend in the threshold voltage variation with respect to the SOI film thickness are depicted in Fig. 10. One can see that Ferry's effective potential approaches overestimates the threshold voltage for a SOI thickness of 3 nm due to the use of rather approximate value for the standard deviation of the Gaussian wave packet which results in a reduced sheet electron density. As the silicon film thickness decreases, the resulting confining potential becomes more like rectangular from a combined effects of both the inversion layer quantization and the SOI film (physical) quantization, which also emphasizes the need for using a more realistic quantum-mechanical wavepacket description for the confined electrons. Of most importance in this figure is the very fact that the new quantum potential approach is free from this large discrepancy and can capture the trend in the threshold voltage as obtained from the more accurate Schrödinger-Poisson solver. These results indicate that the new quantum potential method can be applied to the simulations of SOI devices with a greater accuracy and predictive capability as it will be seen from the results presented in the next section.

2.2.3. *Size-quantization in Nanoscale DG SOI Devices*

Fig. 11 shows the simulated DG SOI device structure used in this section, which is similar to the devices reported in Ref. [66]. For quantum simulation purposes only the dotted portion of the device which has been termed as the *intrinsic* device is taken into considerations. The device was originally designed in order to achieve the ITRS performance specifications for the year 2016.
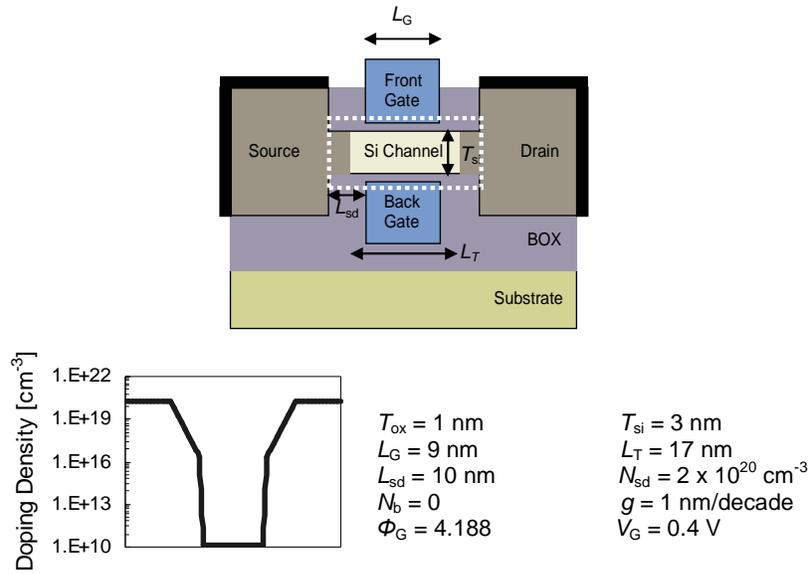
$T_{ox} = 1$ nm $\qquad$ $T_{si} = 3$ nm
$L_{G} = 9$ nm $\qquad$ $L_{T} = 17$ nm
$L_{sd} = 10$ nm $\qquad$ $N_{sd} = 2 \times 10^{20}$ cm^{-3}
$N_{b} = 0$ $\qquad$ $g = 1$ nm/decade
$\Phi_{G} = 4.188$ $\qquad$ $V_{G} = 0.4$ V

Fig. 11. DG device structure.

The effective intrinsic device consists of two gate stacks (gate contact and SiO$_2$ gate di-electric) above and below a thin silicon film. For the intrinsic device, the thickness of the silicon film is 3 nm. Use of a thicker body reduces the series resistance and the effect of process variation but it also degrades the short channel effects (SCE). From SCE point of view, a thinner body is preferable but it is harder to fabricate very thin films of uniform thickness, and the same amount of process variation ($\pm 10\%$) may give intolerable fluctuations in the device characteristics. A thickness of 3 nm seems to be a reasonable compromise, but other body thicknesses are also examined. The top and bottom gate insulator thickness is 1 nm, which is expected to be near the scaling limit for SiO$_2$. As for the gate contact, a metal gate with tunable workfunction, Φ_{G}, is assumed, where Φ_{G} is adjusted to 4.188 to provide a specified off-current value of 4 μA/μm. The background doping of the silicon film is taken to be intrinsic, however, due to diffusion of the dopant ions, the doping profile from the heavily doped S/D extensions to the intrinsic channel is graded with a coefficient of g which equals to 1 nm/dec. For convenience, the doping scheme is also shown in Fig. 11. According to the roadmap, the high performance (HP) device should have a gate length of $L_{G} = 9$ nm at the year 2016. At this scale, two-dimensional (2D) electrostatics and quantum mechanical effects both play an important role and traditional device simulators may not provide reliable projections. The length, L_{T}, is an important design parameter in determining the on-current, while gate metal workfunction, Φ_{G}, directly controls the off-current. The doping gradient, g, affects both on-current and off-current. Values of all the structural parameters of the device are shown in Fig. 11.

The intrinsic device is simulated using the new quantum potential approach in order to gauge the impact of size-quantization effects on the DG SOI performance. The results are then compared to that from a full quantum approach based on the non-equilibrium

Green's function (NEGF) formalism (NanoMOS–2.5) developed at Purdue University [67]. In this method, scattering inside the intrinsic device is treated by a simple Buttiker probe model, which gives a phenomenological description of scattering and is easy to implement under the Greens' function formalism. The simulated output characteristics are shown in Fig. 12. Devices with both 3 nm and 1 nm channel thickness are used with applied gate bias of 0.4 V. The salient features of this figure are discussed as follows: (1) Even with an undoped channel region, the devices achieve a significant improvement with respect to the short channel effects (SCEs) as depicted in flatness of the saturation region. This is due to the use of the two gate electrodes and an ultrathin SOI film which makes the gates gain more control on the channel charge. (2) Reducing the channel SOI film thickness to 1 nm further reduces the SCEs and improves the device performance. However, the reduction in the drive current at higher drain biases is due to series resistance effect pronounced naturally when the drain current increases. (3) Regarding the quantum effects, one can see that quantum-mechanical size quantization plays not a very dominant role in degrading the device drive current mainly because of using an undoped channel region. Also, looking at the 3 nm (or 1 nm) case alone one can see that the impact of quantization effects reduces as the drain voltage increases because of the growing bulk nature of the channel electrons. (4) Percentage reduction in the drain current is more pronounced in 1nm case throughout the range of applied drain bias because of the stronger physical confinement arising from the two SiO_2 layers sandwiching the silicon film. (5) Finally, the comparison between the quantum potential formalism and the NEGF approach for the device with 3 nm SOI film thickness shows reasonable agreement which further establishes the applicability of this method in the simulations of different technologically viable nanoscale classical and nonclassical MOSFET device structures.
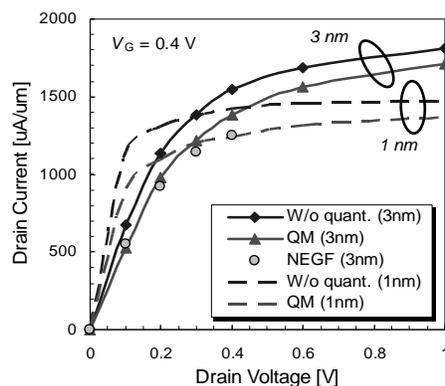

Fig. 12. Generic DG SOI device output characteristics.

3. Discrete Impurity Effects

The pioneering experimental studies by Mizuno and co-workers [68] in the mid 1990's clearly demonstrated that threshold voltage fluctuations due to the discrete nature of the

impurity atoms, are going to be a significant problem in future ultra-small devices. They showed that the threshold voltage standard deviation is inversely proportional to the square root of the gate area, to the oxide thickness, and to the fourth root of the average doping in the device channel region. They also observed that the statistical variation of the channel dopant number accounts for about 60% of the experimentally derived threshold voltage fluctuation. In a later study, Mizuno [69] also found that the lateral and vertical arrangement of ions produces variations in the threshold voltage that depend upon the drain and substrate biases. Horstmann and co-workers [70] investigated global and local matching of sub-100 nm *n*-channel metal-oxide-semiconductor (NMOS) and *p*-channel metal-oxide-semiconductor (PMOS)-transistors and confirmed the area law proposed in [68]. The empirical analytical expression by Mizuno was generalized by Stolk *et al.* [71] by taking into account the finite thickness of the inversion layer, the depth-distribution of the charge in the depletion layer and the influence of the source and drain impurity distributions.

Numerical drift-diffusion and hydrodynamic simulations [72,73,74,75] have also confirmed the existence of the fluctuations in the threshold voltage in ultra-small devices. Two-dimensional (2D) [76] and three-dimensional (3D) [77,78,79,80] ensemble Monte Carlo (EMC) particle-based simulations have also been carried out. An important observation was made in Ref. [29], where it was shown that there is a significant correlation between the threshold voltage shift and the actual position of the impurity atoms. A rather systematic analysis of the random dopant induced threshold voltage fluctuations in ultra-small metal-oxide-semiconductor field-effect transistors (MOSFETs) was carried out by Asenov [81] using 3D drift-diffusion device simulations and confirming previous results. Recent simulation experiments by Asenov and Saini [82] have shown that discrete impurity effects are significantly suppressed in MOSFETs with a δ-doped channel.

However, the majority of the above-mentioned simulation experiments, except [29,79], utilized 2D or 3D device simulators, in which the "discreteness" of the ions was only accounted for through the charge assignment to the mesh nodes. There, the long-range portion of the electron-ion forces is inherent in the mesh force and is found from the solution of the Poisson equation. The short-range portion of these interactions is either completely ignored or treated in the **k**-space portion of the EMC transport kernel (in particle based simulations) or via the doping dependence of the mobility (in drift-diffusion simulations). Because of the complexity and obscurity of the treatment of the Coulomb interaction in the MC simulations, a more direct approach has been introduced [29], in which the MC method is supplemented by a *molecular dynamics* (MD) routine. In this approach, the mutual Coulomb interaction among electrons and impurities is treated in the drift part of the MC transport kernel. Indeed, the various aspects associated with the Coulomb interaction, such as dynamical screening and multiple scatterings, are automatically taken into account. Very recently, the MC/MD method has been extended for spatially inhomogeneous systems. Since a part of the Coulomb interaction is already taken into account by the solution of the Poisson equation, the MD treatment of the Coulomb interaction is restricted only to the limited area near the charged particles. It is claimed that the full incorporation of the Coulomb interaction is indispensable to reproduce the correct electron mobility in highly doped silicon samples.

Although real space treatments eliminate the problem of double counting of the force, a drawback is that the 3D Poisson equation must be solved repeatedly to properly describe the self-consistent fields which consumes over 80% of the total simulation time. To further speed up simulations, in this work a new idea has been proposed: to use a 3D

Fast Multi-Pole Method (FMM) [83,84,85,86], instead. The FMM allows calculation of the field and the potential in a system of n particles connected by a central force within $O(n)$ operations given certain prescribed accuracy. The FMM is based on the idea of condensing the information of the potential generated by point sources in truncated series expansions. After calculating suitable expansions, the long range part of the potential is obtained by evaluating the truncated series at the point in question and the short range part is calculated by direct summation. The field due to the applied boundary biases is obtained at the beginning of the simulation by solving the Poisson equation. Hence the total field acting on each electron is the sum of this constant field and the contribution from the electron-electron and electron-impurity interactions handled by the FMM calculations. *The image charges, which arrive because of the dielectric discontinuity, are handled by the method of images.*

Quite recently, several groups, including ours [87], have shown that the Coulomb effects become even more prominent when the device size scales into the nm range. Even in undoped samples, a single unintentional dopant atom can cause significant fluctuations in the threshold voltage and, therefore, the device on-state current due to the randomness of its position within the device active area. Therefore, *proper inclusion of the short – range Coulomb interactions is a MUST when considering state of the art SOI FD-MOSFETs and alternate device structures, such as dual gate and FinFET devices.*

3.1. *The P³M Method*

The particle-particle-particle-mesh (P³M) algorithms are a class of hybrid algorithms developed by Hockney and Eastwood [88]. These algorithms enable correlated systems with long-range forces to be simulated for a large ensemble of particles. The essence of P³M algorithms is to express the inter-particle force as a sum of a short-range part calculated by a direct particle-particle force summation and a long-range part approximated by the particle-mesh (PM) force calculation. Using the notation of Hockney, the total force on a particle i may be written as

$$F_i = \sum_{j \neq i} F_{ij}^{coul} + F_i^{ext}. \tag{19}$$

F_i^{ext} represents the external field or boundary effects of the global Poisson solution. F_{ij}^{coul}, is the force of particle j on particle i given by Coulomb's law as

$$F_{ij}^{coul} = \frac{q_i q_j}{4\pi\varepsilon} \frac{\left(r_i - r_j\right)}{\left|r_i - r_j\right|^3}, \tag{20}$$

where q_i and q_j are particle charges and r_i and r_j are particle positions. In a P³M algorithm, the total force on particle i is split into two sums

$$F_i = \sum_{\substack{j \neq i \\ SRD}} F_{ij}^{sr} + \sum_{\substack{j \neq i \\ GD}} F_{ij}^{m}. \tag{21}$$

The first sum represents the direct forces of particles j on particle i within the short-range domain (SRD), while the second sum represents the mesh forces of particles j on particle i over the global problem domain (GD) that includes the effect of material boundaries and

the boundary conditions on particle i. F_{ij}^{sr} is the short-range particle force of particle j on particle i, and F_{ij}^{m} is the long-range mesh force of particle j on particle i. The short-range Coulomb force can be further defined as

$$F_{ij}^{sr} = F_{ij}^{coul} - R_{ij},$$ (22)

where F_{ij}^{coul} is given by Eq. (20) and R_{ij} is called the reference force. The reference force in Eq. (22) is needed to avoid double counting of the short-range force due to the overlapping domains in Eq. (21). The reference force should correspond to the mesh force inside the short-range domain (SRD) and equal to the Coulomb force outside the short-range domain. In other words, a suitable form of the reference force for a Coulombic long-range force is one which follows the point particle force law beyond the cutoff radius r_{sr}, and goes smoothly to zero within that radius. Such smoothing procedure is equivalent to ascribing a finite size to the charged particle. As a result, a straightforward method of including smoothing is to ascribe some simple density profile $S(r)$ to the reference inter-particle force. Examples of shapes, which are used in practice, and give comparable total force accuracy, are the uniformly charged sphere, the sphere with uniformly decreasing density

$$S(\mathrm{r}) = \begin{cases} \dfrac{48}{\pi r_{sr}^4}\left(\dfrac{r_{sr}}{2} - r\right), & r \le r_{sr}/2 \\ 0, & otherwise, \end{cases}$$ (23)

and the Gaussian distribution of density. The second scheme gives marginally better accuracies in 3D simulations. For this case the reference force can be obtained [89] as

$$\begin{cases} R_{ij}(r) = \dfrac{q_i q_j}{4\pi\varepsilon} \times \dfrac{1}{35 r_{sr}^2}(224\xi - 224\xi^3 + 70\xi^4 + 48\xi^5 - 21\xi^6) & \xi = \dfrac{2r}{r_{sr}} \text{ and } 0 \le r \le r_{sr}/2 \\ R_{ij}(r) = \dfrac{q_i q_j}{4\pi\varepsilon} \times \dfrac{1}{35 r_{sr}^2}(\dfrac{12}{\xi^2} - 224 + 896\xi - 840\xi^2 + 224\xi^3 + 70\xi^4 - 48\xi^5 + 7\xi^6) & r_{sr}/2 \le r \le r_{sr} \\ R_{ij}(r) = \dfrac{q_i q_j}{4\pi\varepsilon} \times \dfrac{1}{r^2} & r > r_{sr} \end{cases}$$ (24)

Hockney advocates pre-calculating the short-range force, $F_{ij}^{sr}(r)$ in Eq. (22) including the reference force above for a fixed mesh. It is important to extend the P³M algorithm to nonuniform meshes for the purpose of semiconductor device simulation since practical device applications involve rapidly varying doping profiles and narrow conducting channels which need to be adequately resolved. Since the mesh force from the solution to the Poisson equation is a good approximation within about two mesh spaces, r_{sr} is locally chosen as the shortest distance which spans two mesh cells in each direction of every dimension of the mesh at charge i.

In order to incorporate the effects of material boundaries and boundary conditions, the reference force would be found most precisely in the short-range domain by associating particle j with the particle-mesh and calculating the resulting force on particle i with $F_i^{ext} = 0$. Since such a procedure would be required for each particle, it is obviously too costly for reasonable ensemble sizes and defeats the purpose of the P³M algorithm [89]. Instead, it is desirable to use an approximation for this force, which minimizes the effects of the transition error in going from the long-range domain to the short-range domain.

One approach developed in [90] is to choose a particular orientation of approaching particles relative to the mesh and find a radial approximation to the reference force. This method is straightforward and computationally efficient per particle for a fixed uniform mesh, but it is not easily adaptable to nonuniform meshes where the mesh force is not isotropic.

3.2. *The Fast Multipole Method (FMM)*

FMM was first introduced by Rokhlin [83] and was later refined by Greengard [84] for the application of two and three-dimensional *N*-body problems whose interactions are Coulombic or gravitational in nature. In a system of *N* particles, the decay of the Coulombic or gravitational potential is sufficiently slow so that all interactions must be accounted for, resulting in CPU time requirements on the order of $O(N^2)$. On the other hand, the FMM requires an amount of work proportional to *N* to evaluate all interactions to within a round off error, making it practical for large-scale problems encountered in plasma physics, fluid dynamics, molecular dynamics, and celestial mechanics.

There have been a number of previous efforts aimed at reducing the computational complexity of the *N*-body problem. Assuming the potential satisfies Poisson's equation, a regular mesh is laid out over the computational domain and the method proceeds by: (1) interpolating the source density at mesh points; (2) using a fast Poisson solver to obtain potential values on the mesh; (3) computing the force from the potential and interpolating to the particle positions. The complexity of these methods is of the order of $O(N + M \log M)$, where *M* is the number of mesh points. The number of mesh points is usually chosen to be proportional to the number of particles, but with a small constant of proportionality so that $M \langle\langle N$. Therefore, although the asymptotic complexity for the method is $O(N \log N)$ the computational cost in practical calculations is usually observed to be proportional to *N*. Unfortunately, the mesh provides limited resolution, and highly non-uniform source distributions cause a significant degradation of performance. Further errors are introduced in step (3) by the necessity for numerical differentiation to obtain the force. To improve the accuracy of particle-in-cell calculations, short-range interactions can be handled by direct computation, while far-field interactions are obtained from the mesh, giving rise to the so-called particle-particle-particle-mesh (P^{3}M) method described previously. While these algorithms still depend for their efficient performance on a reasonably uniform distribution of particles, in theory they do permit arbitrarily high accuracy to be obtained. As a rule, when the required precision is relatively low, and the particles are distributed more or less uniformly in a rectangular region, P^{3}M methods perform satisfactorily. However, when the required precision is high (as, for example, in the modeling of highly correlated systems), the CPU time requirements of such algorithms tend to become excessive.

3.2.1. *Multipole Moment*

A multipole expansion is a series expansion which describes the effect produced by a given system in terms of an expansion parameter [83] that becomes smaller as the distance of the observation point from the source point increases. Therefore the leading order terms in a multipole expansion are generally the dominant. The first order behavior of

the system at large distances can therefore be predicted from the first terms of the series, which is much easier to compute than the general solution.

Let r be the vector from the fixed reference point to a point in the system and r_1 be the vector from reference point to the observation point, and $d \equiv r_1 - r$ be the vector from a point in the system to the observation point. From the laws of cosines, d can be expressed as

$$d^2 = r_1^2 + r^2 - 2r_1 r \cos\varphi = r_1^2 \left(1 + \frac{r^2}{r_1^2} - 2\frac{r}{r_1}\cos\varphi\right) \qquad (25)$$

where $\cos\varphi \equiv \hat{r}.\hat{r_1}$. Therefore

$$d = r_1 \sqrt{1 + \frac{r^2}{r_1^2} - 2\frac{r}{r_1}\cos\varphi} \qquad (26)$$

Let $\xi \equiv \dfrac{r}{r_1}$ and $y = \cos\varphi$. Then

$$\frac{1}{d} = \frac{1}{r_1}\left(1 - 2\xi y + \xi^2\right)^{-1/2} \qquad (27)$$

But $\left(1 - 2\xi y + \xi^2\right)^{-1/2}$ is the generating function for Legendre Polynomials, i.e.

$$\left(1 - 2\xi y + \xi^2\right)^{-1/2} = \sum_{i=0}^{\infty} \xi^i P_i(y) \qquad (28)$$

so

$$\frac{1}{d} = \frac{1}{r_1}\sum_{i=0}^{\infty}\left(\frac{r}{r_1}\right)^i P_i(\cos\varphi) = \sum_{i=0}^{\infty}\frac{1}{r_1^{i+1}}r^i P_i(\cos\varphi) . \qquad (29)$$

Any physical potential that obeys a $1/d$ law can therefore be expressed as a multipole expansion

$$V = \sum_{i=0}^{\infty}\frac{1}{r_1^{i+1}}\int r^i P_i(\cos\varphi)\rho(r)d^3r \ . \qquad (30)$$

In MKS unit

$$V = \frac{1}{4\pi\varepsilon_0\varepsilon_r}\sum_{i=0}^{\infty}\frac{1}{r_1^{i+1}}\int r^i P_i(\cos\varphi)\rho(r)d^3r , \qquad (31)$$

where ε_0 is the permittivity of the free space , ε_r is the dielectric constant of the medium and $\rho(r)$ is the charge density.

3.2.2 *How FMM speeds up the computation*?

In FMM *multipole moments* are used to represent distant particle groups and a *local expansion* is used to evaluate the contribution from distant particles in the form of a series. The multipole moment associated with a distant group can be *translated* into the coefficient of the local expansion associated with a local group. In FMM the computational domain is decomposed in a hierarchical manner with a quad-tree in two dimensions and an oct-tree in three dimensions to carry out efficient and systematic grouping of particles

with tree structures. The hierarchical decomposition is used to cluster particles at various spatial lengths and compute interactions with other clusters that are sufficiently far away by means of the series expansions.

For a given input configuration of particles, the sequential FMM first decomposes the data-space in a hierarchy of blocks and computes local neighborhoods and *interaction-lists* involved in subsequent computations. Then, it performs two passes on the decomposition tree. The first pass starts at the leaves of the tree, computing *multipole expansion coefficients* for the Columbic field. It proceeds towards the root accumulating the multipole coefficients at intermediate tree-nodes. When the root is reached, the second pass starts. It moves towards the leaves of the tree, *exchanging* data between blocks belonging to the neighborhoods and interaction-lists calculated at tree-construction. At the end of the downward pass all long-range interactions have been computed. Subsequently, nearest-neighbor computations are performed directly to take into consideration interactions from nearby bodies. Finally, short- and long-range interactions are accumulated and the total forces exerted upon particles are computed. The algorithm repeats the above steps and simulates the evolution of the particle system for each successive time-step.

3.3   The Role of Discrete Impurities as Observed by Simulations and With Comparisons to Experiments

3.3.1. Previous Knowledge on Threshold Voltage and On-State Current Fluctuations in Sub-Micrometer MOSFET Devices

As already discussed in great details in the Introduction, continued scaling of devices has led to a number of undesirable effects, including fluctuations in the threshold voltage that arise because of the discrete, or atomistic nature, of the impurity atoms in the device active region. For better insight of the importance of this issue, we have considered a prototypical MOSFET with 0.07 µm channel length, 0.07 µm channel width and channel doping of 10^{18} cm^{-3}. The number of dopant atoms in the depletion region of this device is on the order of several hundreds, and well below 100 in the active region. In addition, there are regions where the impurity atoms cluster and other regions in which the impurity density is well below the average value expected from the doping level. With such a small number of the impurity atoms in the device active region, the local variations in the "doping concentration" across the channel become a significant factor in determining the threshold voltage, mobility and drain current characteristics. This, in turn, causes considerable problems for circuit design, especially for circuits in which the devices must be well matched, such as operational amplifiers [91] and static random access memories [92]. The *SIA* roadmap technology requirements state that the variation in gate length should be less than 10% and the variation in threshold voltage should be less than 40 mV for devices in the 150 nm generation and beyond [3].

It is interesting to note that the existence of these surface potential fluctuations in MOS devices was postulated by Nicollian and Goetzberger [93] to explain the departures from the theoretical predictions in conductance versus frequency measurements in MOS structures. In addition to their effect on the *ac*-conductance results, surface potential fluctuations were also found to have significant influence on a variety of other device characteristics, such as threshold voltage, transconductance, substrate current and off-state leakage currents. Experimental studies by Mizuno, Okamura, and Toriumi [23] have shown

that the threshold voltage standard deviation is related to the average number of ionized impurities beneath the channel according to

$$\sigma_{vt} = \left(\frac{\sqrt[4]{q^3 \varepsilon_s \phi_b}}{\sqrt{2}\varepsilon_{ox}} \right) \frac{T_{ox} \sqrt[4]{N}}{\sqrt{L_{eff} W_{eff}}} \ , \tag{32}$$

where N is the average channel doping density, ϕ_b is the built-in potential, T_{ox} is the oxide thickness, L_{eff} and W_{eff} are the effective channel length and width, and ε_s and ε_{ox} are the semiconductor and oxide permittivity, respectively. They found that the statistical variation of the channel dopant number accounts for about 60% of the experimentally derived threshold voltage fluctuations. In a later study, Mizuno [69] also found that the lateral and vertical arrangement of ions produces variations in the threshold voltage dependence upon the drain and substrate bias. Quite recently, Horstmann, Hilleringmann and Goser [94], who investigated the global and local matching of sub-100 nm NMOS- and PMOS-transistors, confirmed the law of area given in (1). Also, Stolk, Widdershoven and Klaassen [95] generalized the analytical result by Mizuno and his co-workers, by taking into account the finite thickness of the inversion layer, depth-distribution of charges in the depletion layer and the influence of the source and drain dopant distributions and depletion regions. For a uniform channel dopant distribution, the analytical expression for the threshold voltage standard deviation given in [95] simplifies to

$$\sigma_{vt} = \left(\frac{\sqrt[4]{q^3 \varepsilon_s \phi_b}}{\sqrt{3}} \right) \left[\frac{k_b T}{q} \cdot \frac{1}{\sqrt{4\varepsilon_s \phi_b N_a}} + \frac{T_{ox}}{\varepsilon_{ox}} \right] \frac{\sqrt[4]{N}}{\sqrt{L_{eff} W_{eff}}} \ . \tag{33}$$

In Eq. (33), the first term in the square brackets represents the surface potential fluctuations whereas the second term represents the fluctuations in the electric field.

The purpose of this section is twofold. First, we will clarify some issues related to the origin of the threshold voltage fluctuations in ultra-small devices. The second, and more important issue discussed here is how discrete impurities affect device high-field characteristics, such as carrier drift velocity and the on-state currents in conventional MOSFETs.

A. *The role of the short-range e-e and e-i interactions*

To be able to study the effect of the proper inclusion of the short-range Coulomb force to the mesh force, the energy and position of several electrons were monitored during a simulation run. The simulated device has channel length L_G= 80 nm, channel width W_G=80 nm and oxide thickness T_{ox}=3 nm. The lateral extension of the source and drain regions is 50 nm. The channel doping equals 3×10^{18} cm^{-3}. The applied bias is $V_G = V_D =$ 1 V. Only those electrons that entered the channel region from the source side were "tagged" and their energy and position was monitored and used in the average energy calculation. The average velocity and the average energy of the electrons that reach the drain end of the device is shown in

Fig. 13. From the average velocity simulation results, it follows that the short-range *e-e* and *e-i* interaction terms damp the velocity overshoot effect, thus increasing the transit time of the carriers through the device, thus reducing its cut-off frequency (

Fig. 13(a)). It is also quite clear that when we use the mesh force only, i.e. we skip the MD loop that allows us to correct for the short-range *e-e* and *e-i* interactions, those electrons that enter the drain end of the device from the channel never reach equilibrium (Fig. 13(b)). Their average energy is more than 60 meV far into the drain region. Also, the average energy peaks past the drain junction. The addition of the short-range Coulomb forces to the mesh force via the MD loop, leads to rapid thermalization of the carriers once they enter the drain region. The characteristic distance over which carriers thermalize is on the order of a few nm.
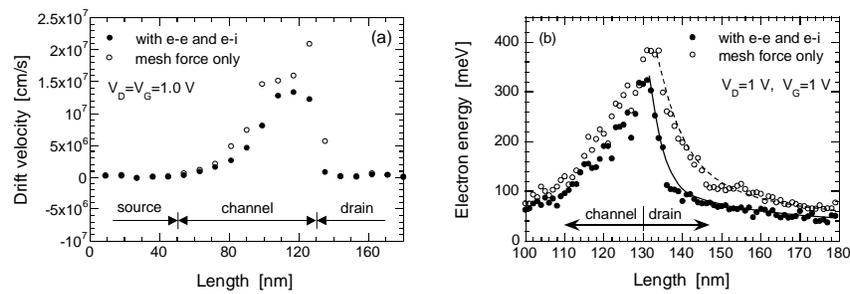
Fig. 13. (a) Average velocity of the electrons along the channel, with and without the inclusion of the *e-e* and *e-i* interactions. (b) Average energy of the electrons coming to the drain from the channel. The applied bias equals $V_D = V_G = 1$ V. Filled (open) circles correspond to the case when the short-range *e-e* and *e-i* interactions are included (omitted) in the simulations.

In Fig. 14 we show the phase-space trajectory of 10 randomly selected electrons that reach the drain region. We use $V_G = 0.5$ V, $V_D = 0.8$ V, $T_{ox} = 3$ nm, and $N_A = 3 \times 10^{17}$ cm^{-3} in these simulations. Notice that some of the electrons reach the end of the device and are reflected back without loosing much energy when we use the mesh force only (Fig. 14(a)). The addition of the short-range Coulomb force leads to very fast thermalization of the carrier energy once they enter the drain end (Fig. 14(b)). None of the randomly selected electrons reach the device boundary, as opposed to 3 out of 10 electrons reaching the boundary when the short-range Coulomb force is turned off.
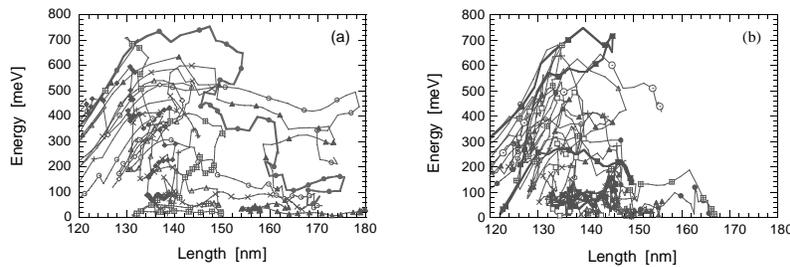

Fig. 14. (a) Phase-space trajectories of 10 randomly chosen electrons for the case when the mesh force is only considered in the free-flight portion of the simulator. (b) Phase-space trajecto-ries of 10 randomly chosen electrons for the case when the short-range e-e and e-i interac-tions are included via our MD routine.

B. *Threshold Voltage Fluctuations*

The threshold voltage fluctuations versus device gate width, channel doping and oxide thickness, are shown in Fig. 15. Also shown in this figure are the analytical model predictions given by Eqs. (32) and (33). The decrease of the threshold voltage fluctuations with increasing the width of the gate is due to the averaging effects, in agreement with the experimental findings by Horstmann *et al.* [70]. We want to point out that we still observed significant spread of the device transfer characteristics along the gate voltage axis even for devices with $W_G = 100$ nm. This is due to the nonuniformity of the potential barrier, which allows for early turn-on of some parts of the channel. As expected, the increase in the channel doping leads to larger threshold voltage standard deviation $\sigma_{V_{TH}}$.

These results also imply that the fluctuations in the threshold voltage can be even larger in devices in which counter ion implantation is used for threshold voltage adjustments. Similarly, the increase in the oxide thickness leads to linear increase in the threshold voltage standard deviation. The results shown in Fig. 15(a-c) also suggest that reconstruction of the established scaling laws is needed to reduce the fluctuations in the threshold voltage. In other words, within some new scaling methodology, T_{ox} should become much thinner, or N_A much lower that what the conventional scaling laws give.
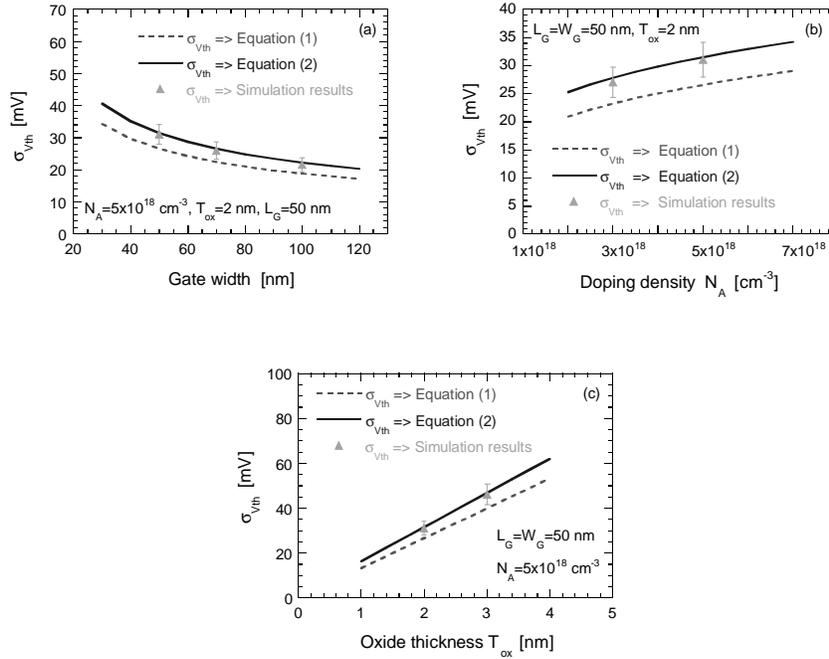

Fig. 15. Variation of the threshold voltage with (a) gate width, (b) channel doping, and (c) oxide thickness.

C.*Fluctuations in the on-state currents*

Besides investigating the threshold voltage fluctuations, our 3D EMC particle-based simulator also allsows us to investigate the fluctuations in the high-field characteristics, such as the saturation drain current. The variation of the drain current versus the number of channel dopant atoms for the 15 devices from Ref. [96] described in terms of the number of dopants in Fig. 16(a), is shown in Fig. 16(c). Each device was simulated for a total of 4 ps. The gate voltage was set to 1.5 V and the drain voltage to 1.0 V. The drain current was measured by averaging the velocity of electrons in the channel over the last 2.4 ps of the simulation. It is important to note that at these bias conditions, the devices were in the saturation region of the I_D-V_G curve, but were not velocity saturated.
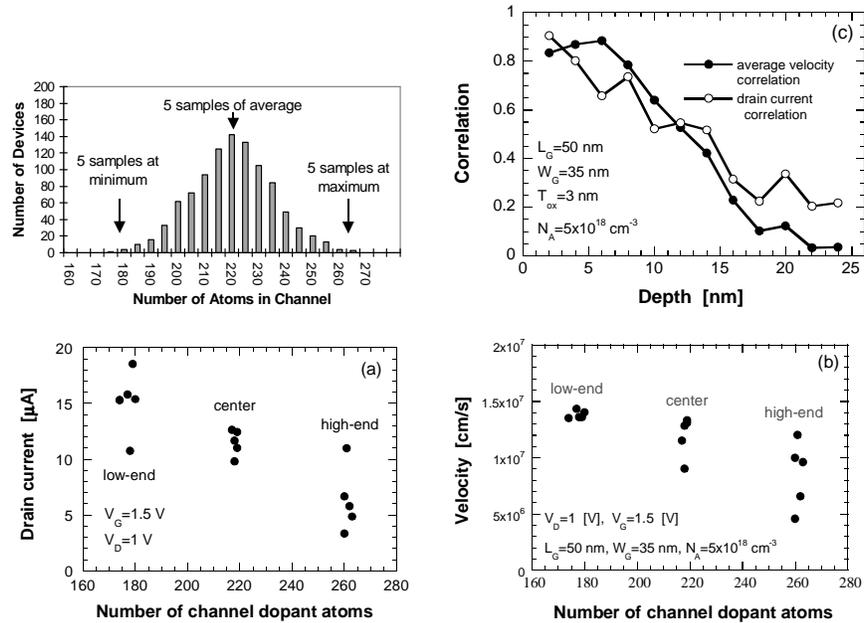
Fig. 16. (a) Histogram of the number of dopant atoms in the channel for a population of 1000 type #5 devices. (b) Correlation of the drain current and average electron velocity to the number of dopant atoms within a 10 nm range at various depths beneath the channel. (c) Drain current versus the number of channel dopant atoms. (d) Average velocity of channel electrons versus the number of channel dopant atoms.

As expected, as the number of channel dopant atoms increases, the drain current decreases due to the increase in the V_T. More importantly, for the five devices from the high-end of the distribution, due to the larger probability that some of the impurity atoms will be located near the semiconductor/oxide interface, there is larger fluctuation in the saturation current. This is also reflected in the average velocity of channel electrons versus the number of dopant atoms in the channel, shown in Fig. 16(d). Again, the velocity decreases as the number of dopant atoms increases due to increased ionized impurity scattering. At the low end of the dopant number distribution, the average electron velocity is roughly the same for each dopant configuration. However, the fluctuation in the

electron velocity increases with the number of dopant atoms, with a 3× spread in the velocity seen for the devices at the high dopant number extreme.

The average electron velocity and device drain current characteristics were correlated to the number of dopant atoms in a 10 nm range at various depths. Fig. 16(b) shows a plot of the square of the correlation coefficient versus depth (beneath the semiconductor/oxide interface). The correlation to the electron velocity is very high for the first 6 nm, and steadily decreases up to 18 nm depth, beyond which the correlation is nearly zero. It appears that only the dopant atoms in the first 6-10 nm from the semiconductor/oxide interface have significant effect on the velocity. This is reinforced by the fact that the correlation nearly goes to zero at a depth of 18 nm, as opposed to the threshold voltage correlation, which remains fairly high at a larger depth. The correlation of the drain current to the number of dopant atoms is also high near the surface, but the drop-off is not as steep as the velocity correlation. Beyond 18 nm depth, the correlation of the drain current is non-zero due to the correlation of the threshold voltage to the number of dopant atoms (see previous discussion).

3.3.2. Threshold Voltage Fluctuations due to Unintentional Doping in Narrow-Width SOI Device Structures

The SOI device structure that has been simulated in this work to study comprehensively the effects of quantum mechanical size-quantization and discrete/unintentional dopant effects on the performance of nanoscale devices is shown in Fig. 17. It consists of a thick (600 nm) silicon substrate, on top of which is grown 400 nm of buried oxide. The thickness of the silicon on insulator layer is 7 nm, *with p⁻ region width of 10 nm* (if not stated otherwise) making it a fully-depleted device under normal operating conditions. The channel length is 50 nm and the doping of the p^- active layer is 10^{16} cm^{-3} which corresponds to a nearly undoped channel region. The source/drain length is 15 nm, width being three times the channel width i.e. 30 nm. On top of the SOI layer sits gate-oxide layer, the thickness of which is 34 nm. This is rather thick gate oxide, but it is used to compare the simulation results with the experimental data of Majima *et al*. [97]. The doping of the source/drain junctions equals 10^{19} cm^{-3} (if not stated otherwise), and the gate is assumed to be a metal gate with workfunction equal to the semiconductor affinity. The use of the low source-drain doping is justified by the fact that most of the carriers that are being simulated are residing in the source/drain regions and the reduction of the source/drain doping leads to a smaller ensemble of carriers. It has been found via Silvaco ATLAS Drift-Diffusion simulations of similar device structures that a reduction in the source/drain doping by one order of magnitude leads to approximately 20-30% decrease in the on-state current due to the additional source/drain series resistances.

In a 50 nm by 10 nm by 7 nm SOI device structure in Fig. 17, with a channel doping of 10^{16} cm^{-3}, one has merely a single dopant atom in the channel region. Even if the channel is undoped, the unavoidable background doping gives rise to at least 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 active silicon body, it will introduce a fixed charge in the channel region. These potential sources of localized single charge will introduce a highly *localized barrier* to the carrier/current flow. Such a localized barrier is shown in Fig. 18. The device operation is affected by this localized barrier from both electrostatics (effective increase in doping) and dynamics (transport) points of view. The transport is affected through modulation of carrier velocity and energy charac-

teristics as shown in Fig. 19(left panel) where the dip is due to the presence of a single impurity in the center of the channel region. In Fig. 19(right panel), the device transfer characteristics are shown for a device with continuum doping and with an unintentional dopant present in the center of the channel. The channel width is 10 nm. One observes increase in the device threshold voltage, V_{th} and degradation of the drain current due to the presence of a single charge.
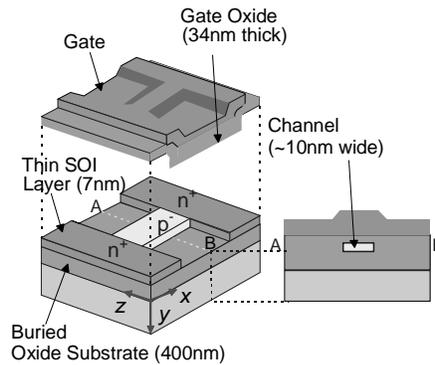

Fig. 17. Device structure of ultra-narrow channel FD-SOI device.
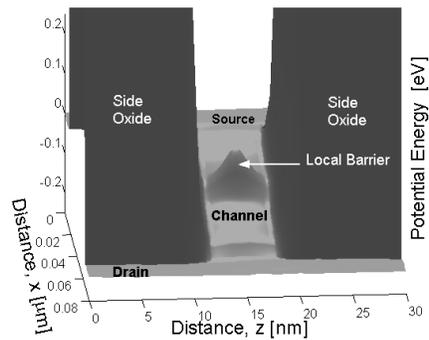
Fig. 18. Shape of the conduction band profile when a single impurity is localized in the center of the channel.

In Fig. 20 shown are the fluctuations in the drain current as a function of the position of a single dopant ion in the channel region of the device. Simulations have been performed using $V_G = 1.0$ V and $V_D = 0.1$ V. Results for devices with channel width of both 10 nm and 5 nm are shown. Due to the size-quantization effect, which, as a consequence of the charge set-back, results in the majority of current flowing through the middle portion of the channel, a dopant ion trapped in the center region of the channel produces maximum fluctuations in the on-state current. The drain-end is less affected due to two

reasons: (a) the presence of a weaker quantization effect therein due to the least vertical field experienced by the electrons and (b) the presence of the largest in-plane (x-component) electric field along the length of channel region which obviously minimizes the effect of the single dopant.
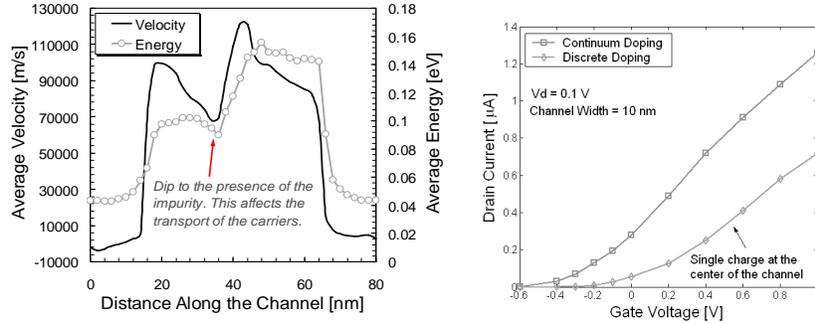

Fig. 19. Left panel: Velocity and energy plots for $V_G = 1.0$ and $V_D = 0.2$ V when a single impurity is present at the center of the channel. Right panel: Device transfer characteristics for the case of a continuum and discrete impurity model with a single charge at the center of the channel.

To investigate the impact of screening effect for the impurity positioned along the center of the channel region on the drain current a detailed simulations were performed. The results are shown in Fig. 21. One can see that the impurity positioned in the very vicinity of the source-end has lower effect than when positioned little away from the source-end. This is attributed to the fact that the very presence of a large number of electrons in the source region try to screen further the impurity and thereby its effect on the drain current.

The impurity position dependence of the drain current is shown in Fig. 22(left panel) in the device output characteristics. There are several noteworthy conclusions that can be drawn from these simulations:

- Single impurity at the source-end of the channel affects the drain current most.
- Impurities at the drain-end of the channel reduce the DIBL (drain-induced-barrier-lowering) in the output characteristics.
- Dopant atoms trapped in the center region of the channel produce the maximum fluctuations than the dopant atoms near the interface.

The observed impurity position dependence of the drain current may be attributed to both the inhomogeneities in the electrostatics and the non-uniform carrier quantization in the channel region. Another potential source arises from the modulation of the transport characteristics which is reflected in the carrier velocity behavior as shown in the right panel of Fig. 22. Here, the velocity profiles for impurities at three different positions are shown. One can see that the impurity near the source end affects (reduces) the electron velocity most throughout the channel region. Simulations have been performed using $V_G = 1.0$ V and $V_D = 0.2$ V.
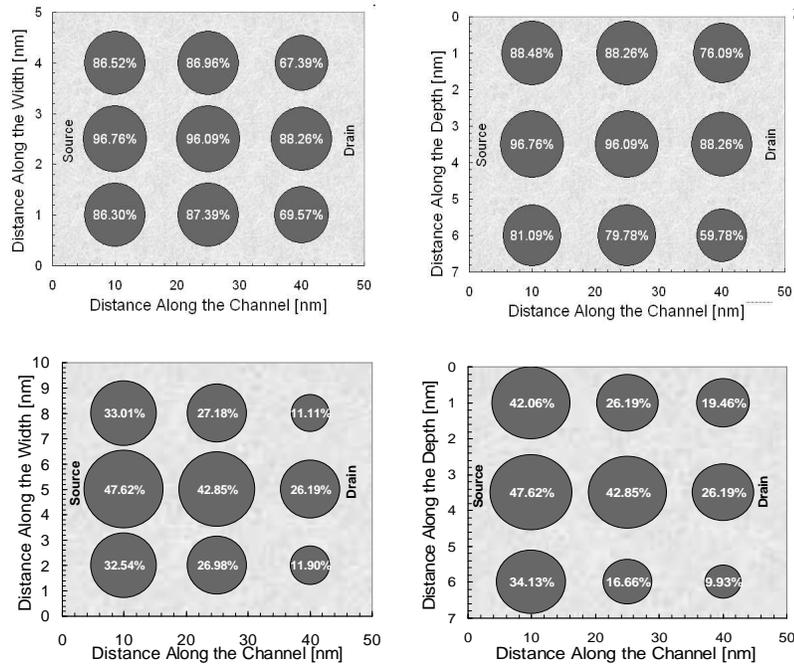
Fig. 20. Fluctuation in the drain current with the different positioning of the single impurity along the width (*X-Z* plane) and along the depth (*X-Y* plane) of the device. (a) The device width is 5 nm, and (b) The device width is 10 nm.
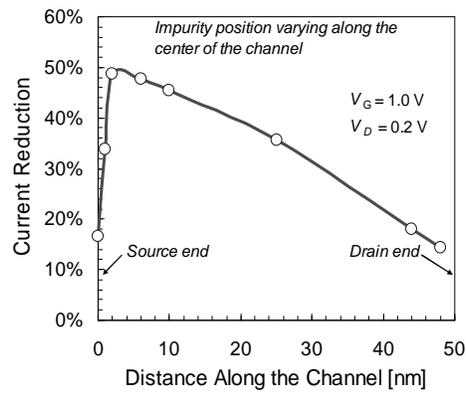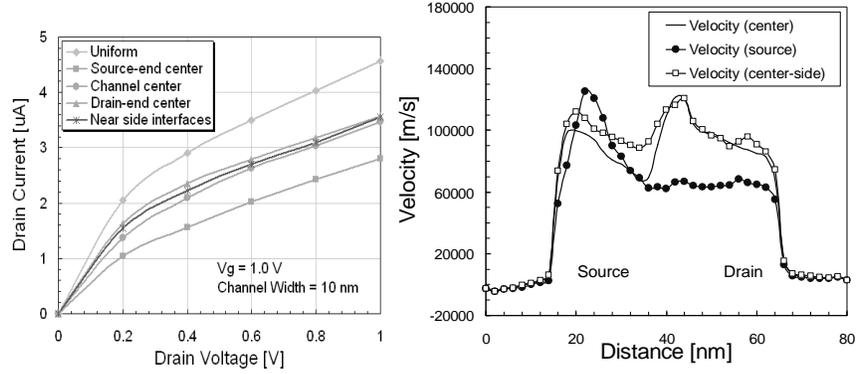

Fig. 21. Impact of screening on the drain current.

Fig. 22. Left panel: Variations of the device drain current as a function of the placement of a single impurity at various positions in the channel. We have used $V_G = 1.0$ V in these simulations. Right panel: Variations of the electron velocity as a function of the placement of a single impurity at various positions in the channel. We have used $V_G = 1.0$ V and $V_D = 0.2$ V in these simulations.

The results presented in Fig. 22 also suggest that there might be fluctuations in the device threshold voltage for devices fabricated on the same chip due to unintentional doping and random positioning of the impurity atoms. This can also be deduced from the scatter of the experimental data from Ref. [97]. The simulation results of the transfer characteristics with a single impurity present in different regions in the channel of the device, shown in the left panel of Fig. 23, clearly demonstrate the origin of the threshold voltage shifts for devices with 10 nm and 5 nm channel width. The width dependence of the threshold voltage for the case of a uniform (undoped) and a discrete impurity model is shown in the right panel of Fig. 23. This figure suggests that *both size-quantization effects and unintentional doping must be concurrently considered to explain threshold voltage variation in small devises*.
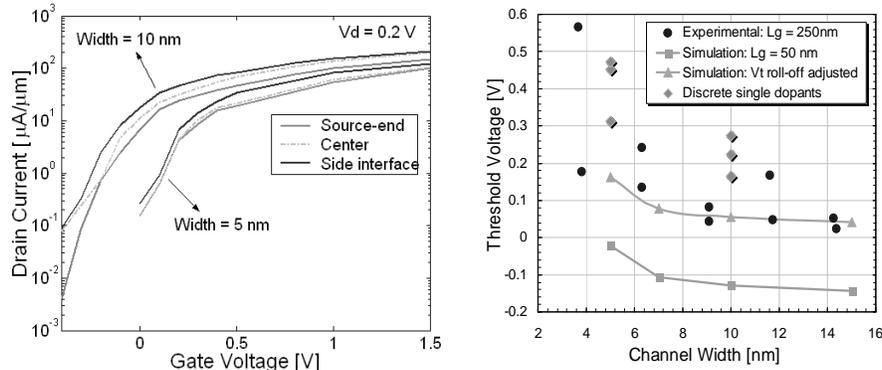

Fig. 23. Left Panel: Transfer characteristics of the device with 10 nm and 5 nm channel widths and different location of the impurity atoms. We have used $V_G = 1.0$ V in these simulations. Right Panel: Width dependence of the threshold voltage for the case of a uniform and a discrete impurity model. Clearly seen in this Fig. are two trends: (a) Threshold voltage increase with decreasing channel width due to quantum-mechanical size quantization effects, and (b) Scatter in the threshold voltage data due to unintentional doping.

3.3.3   The role of Unintentional Doping on FinFET Device Design Parameters

 The FinFET device structure that has been simulated in this work is shown in Fig. 24 [98]. It consists of a thick (100nm) buried oxide on top of which source/drain regions and a vertical fin are formed. The channel length is 40 nm with a gate length of 20 nm and a fin extension length of 10 nm on each side of the gate. The fin height and width are 30 nm and 10 nm, respectively. The source/drain length is 20 nm, the width being three times the channel width, i.e. 30 nm. The doping of the source/drain junctions equals 2×10^{19} cm^{-3}. The fin is assumed intrinsic. The gate is assumed to be n$^+$ polysilicon with work function equal to the semiconductor affinity. Gate oxide of 2.5 nm has been used for both side and top gates. To simulate this device structure, a convenient meshing scheme has been adopted. Meshing is uniform along the *x* (channel length) and *z* (width) directions and is non-uniform along the *y* (depth) direction, with the exception of the semiconductor region, where uniformity in meshing has been kept in order to facilitate the Monte Carlo transport simulations.

Significant velocity overshoot is observed in small geometry devices due to the presence of very high electric fields. Fig. 25(left panel) depicts the average velocity profile along the channel length of a FinFET device. Equal amount of velocity overshoot is observed near the source and the drain end of the channel when fin extension length on each side of the gate is equal. Note that the magnitude of the velocity overshoot also depends on the fin extension length on each side of the gate and this observation is discussed later in the text. Fig. 25(right panel) depicts the average energy profile along the device channel length. Near the source end the average carrier energy equals the thermal energy. Along the channel the average energy increases progressively reaching its peak value near the drain end. Note that carriers are not thermalized near the drain end of the channel due to the omission of the short-range electron-electron and electron-ion interactions in these simulations. Fin extension of 10 nm has been used on each side of the gate. The applied bias equals $V_D = V_G = 0.8$ V.
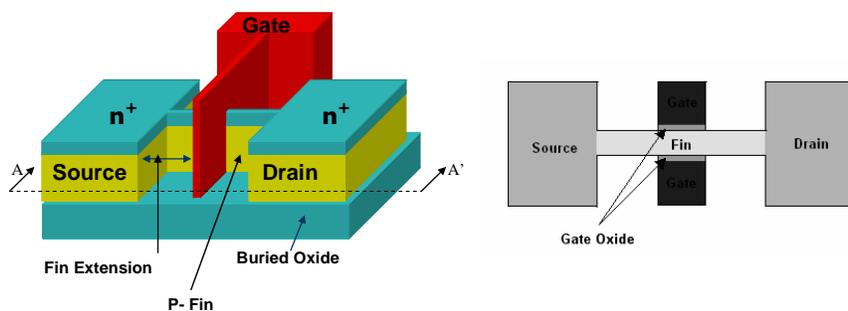

Fig. 24. Left panel: 3D schematic view of FinFET. Right panel: Top view of the FinFET shown in top panel along the cross section A-A'.

The amount of velocity overshoot the carriers experience within the FinFET devices shown previously heavily depends on the fin extension length on each side of the gate. Keeping D/G gap fixed, gradual increase in S/G gap causes the source end to experience

more overshoot and the drain side overshoot to gradually diminish as shown in Fig. 26(left panel). This is due to the fact that with an increase in extension length, source and drain lateral fields along the channel redistribute which changes the velocity profiles which can be seen from the 1-D conduction band profile along the x-direction as shown in Fig. 26(right panel). Near the drain end and in the channel, the slope of conduction band decreases with increase in S/G gap, resulting in lower electric field. Also note that near the source end the slope of conduction band increases giving higher electric field at that region. D/G gap is fixed at 10nm and $V_D = V_G = 0.8$ V is used in the simulation. The same phenomena happen for varying the D/G gap while keeping S/G gap constant at 10nm.
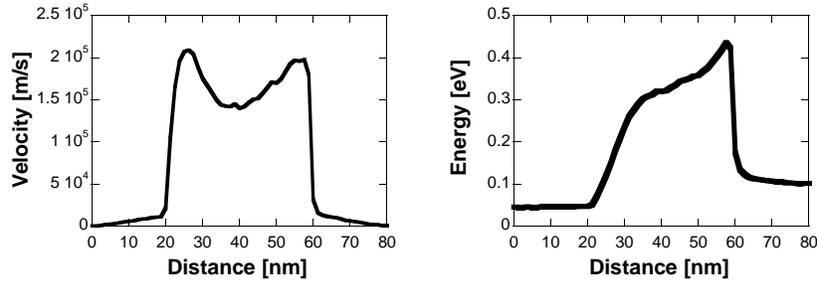

Fig. 25. Left panel: Average velocity (x-component) profile of carriers along the channel. Right panel: Average energy of carriers along the length of the device. $V_G = V_D = 0.8$V and S/G=D/G=10nm.
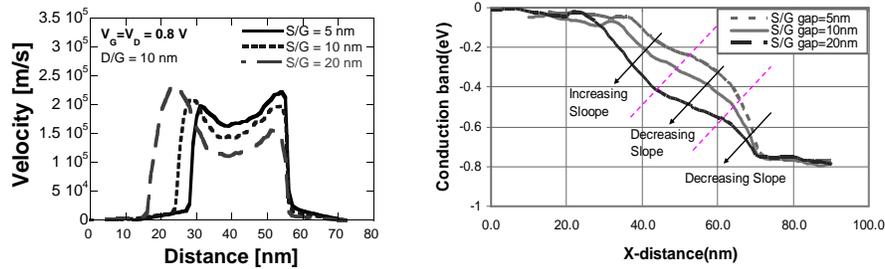

Fig. 26. Left panel: Average velocity (x-component) profile of carriers along the channel as a function of S/G gap. The applied bias equals $V_G = V_D = 0.8$ V. Right panel: Conduction band profile along x-direction.

From the transfer characteristics of the device as shown in Fig. 27(left panel), it is evident that the threshold voltage is negative and is around -0.1 V. Negative threshold voltage results due to the use of n^+-polysilicon as a gate electrode. The metal work function equal to the electron affinity of Si is assumed in the simulation. Polysilicon gates also suffer from depletion and high gate resistance. A nominal threshold voltage of 0.2-0.4 V for n-channel FinFET can be achieved using metal gates with work function close to the mid band-gap of silicon (~4.6 eV). Achieving symmetric threshold voltages for both n-channel and p-channel FinFETs requires metals with different work functions [99]. The output characteristics of the device from Fig. 24 are presented in Fig. 27(right panel). Equal fin extension of 10nm is assumed on both sides of the gate. Gate voltage $V_G = 0.4$

V is used. The inclusion of the electron–electron and electron-ion interaction results in lower drain current. Also the Fast Multipole method (FMM) gives output characteristic, which is in good agreement with that using the P^{3}M approach.
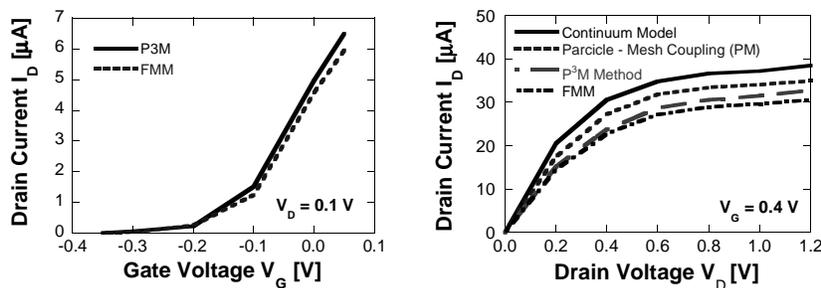

Fig. 27. Left panel: Transfer characteristics. Right panel: Output characteristics.

It is important to note that the CPU time requirement when using the FMM is much smaller compared to the traditional P^{3}M approach. Table 4 gives a comparison of the CPU time requirements for simulating FinFET device with 3D mesh of $64 \times 24 \times 24$ node points. The number of particles simulated is around 1500. The speedup due to using FMM depends on the number of particles, mesh size and computational resources. As the number of particles increases, FMM becomes slower but still much faster when compared to the P^{3}M approach. Also for very small number of particles, it is better to calculate e-e and e-ion interaction directly than using FMM [100]. Correction for image charges is incorporated in our simulator to get the precise results.

Table 4. P^{3}M vs. FMM speed-up.

Approach	CPU time per iteration
P^{3}M	~24 sec
FMM	<1 sec

FinFET devices use undoped or lightly doped fin. In a 40 nm by 10 nm by 30 nm channel region, with a channel doping of 10^{16} cm^{-3}, one has merely 0.12 dopant atoms in the channel region. Even if the channel is undoped, the unavoidable background doping gives rise to at least 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. These potential sources of localized single charge will introduce a localized barrier to current flow. The position of a single dopant at the center of the channel along with the localized barrier it creates is shown in Fig. 28(left and right panel). The device operation is affected by this localized barrier from both electrostatics (effective increase in doping) and dynamics

(transport) points of view. The effective increase in doping in the channel region results in increase in the threshold voltage and consequently, the drain current reduces. The transport is affected through modulation of the carrier velocity and energy characteristics.
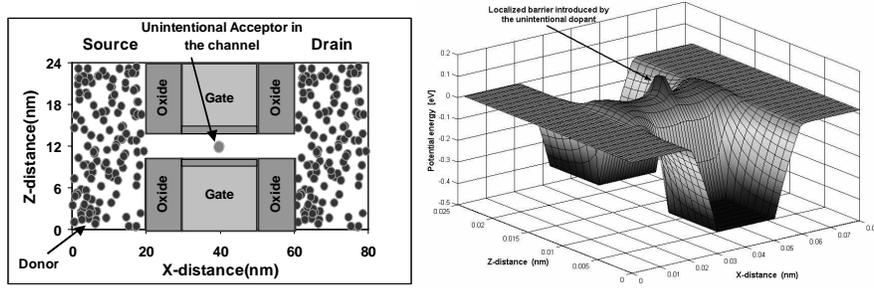

Fig. 28. Left panel: Top view of the FinFET device showing dopant position at the center region of the channel. Right panel: Potential profile showing the localized barrier introduced by the unintentional dopant.

Due to the presence of multiple channels in the FinFET device, the effect of unintentional doping is not that much pronounced. The reduction in drain current heavily depends on the fin width. With decrease in fin width, the localized barrier has more pronounced effect on carrier motion through the channel, and the reduction in drain current is significant. This trend is schematically shown on the left panel of Fig. 29. Fin extension length of 10 nm is used on each side of the gate. $V_D = 0.1$ V, $V_G = 0.4$ V is used in the simulation. The unintentional dopant is placed near the source end close to the top interface. Fin extension length on each side also influences the reduction in drain current due to unintentional dopant as it is shown in the right panel of Fig. 29. Longer fin extension results in more reduction in drain current than that due to smaller fin extension for any dopant position. With longer fin extension, lateral field from source and drain has less influence on the barrier produced by the unintentional dopant thereby reducing the drain current more when compared to the case with smaller fin extension. Fin extension length can, therefore, be optimized for suppressing unintentional doping effects while keeping the drive current within required range. $V_G = 0.4$ V and $V_D = 0.1$ V is used. The dopant atom is placed near the source end close to the top interface. Fin width of 4nm is used. As noted in earlier device structures, the reduction in drain current due to unintentional dopant significantly depends on the position of the dopant atom in the channel. It is found that dopant placed near the source end has greater effect on the drain current. Near the drain end, the effect is less pronounced. Since in FinFET devices channels are formed symmetrically in vertical plane on each side of the fin, placing the unintentional dopant near the center along the width will reduce drain current more than that caused by dopant for any other position.

The effect of unintentional doping on device operation is relatively strong near sub threshold regime/weak inversion when few carriers are present in the channel. Thus the presence of unintentional dopant in the channel is expected to affects the switching behavior of the device. Increasing either the gate voltage or the drain bias will reduce the effects. As the gate voltage is increased, the number of carriers in the channel region increases and screens the localized potential produced by the unintentional dopant as shown

in the left panel of Fig. 30. Drain bias of 0.1 V is applied in the simulation. Unintentional dopant is placed at the center of the channel near the top interface. Similarly with increase in drain voltage carriers are accelerated more along the channel and can easily overcome the localized barrier. Therefore the reduction in drain current gradually decreases with increasing drain bias as shown in the right panel of Fig. 30. Gate bias of 0.4 V is applied in the simulation. Dopant is placed near the source end of the fin close to the top interface.
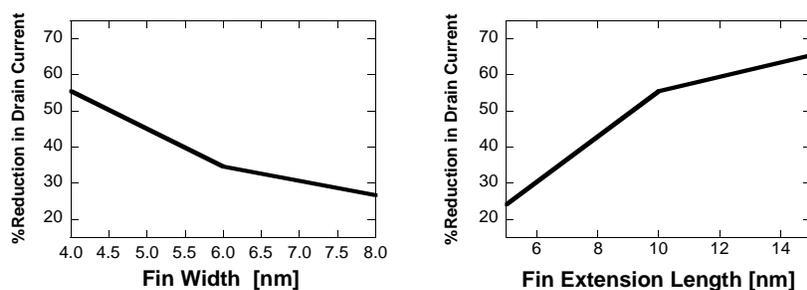

Fig. 29. Left panel: Reduction in drain current due to unintentional dopant as a function of fin width. $V_G = 0.4$ V, $V_D = 0.1$ V. Right panel: Reduction in drain current due to unintentional dopant as a function of fin extension length. $V_G = 0.4$ V, $V_D = 0.1$ V.
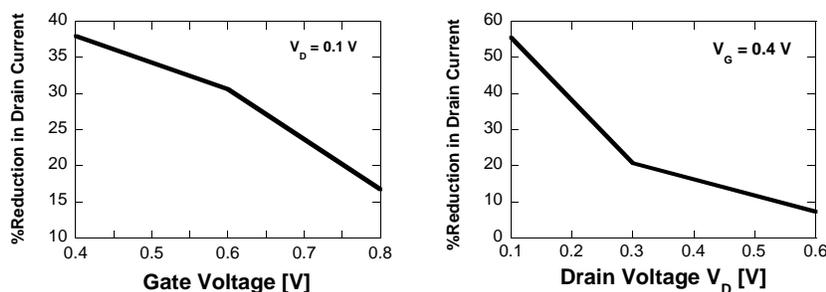

Fig. 30. Left panel: Screening behavior of the carriers on reduction of drain current due to unintentional dopant. Right panel: Reduction in drain current due to unintentional dopant as a function of drain voltage.

3. Conclusions

A recently proposed *effective potential* approach has been utilized to successfully simulate two-dimensional space-quantization effects in a model of a narrow-channel SOI device structure. The incorporation of the *effective potential* approach into a full 3D Monte Carlo particle-based simulator allows one to investigate the device transfer and output characteristics with proper treatment of the size-quantization effects, velocity overshoot and carrier heating on an equal footing. The *effective potential* provides a set-back of the charge from the interface proper and a quantization energy within the channel. Both of these effects

lead to an increase in the threshold voltage. A threshold voltage increase of about 180 mV has been observed when the effective potential is included in the SOI device with 10 nm channel width. Also, observed is a pronounced channel width dependency of the threshold voltage which is termed as the *quantum mechanical narrow channel effect*. The width dependence of the threshold voltage is in close agreement with the experimental results. The increase in the threshold voltage is found to give rise to a significant on-state current reduction (20-30%), which depends upon the gate bias. Larger degradation is observed for larger gate voltages. The energy characteristics along the channel do not change with the inclusion of quantum mechanical size-quantization effects. The average drift velocity shows a small decrease due to the smearing of the potential.

A novel effective potential approach has been proposed and tested in simulations of quantization effects in 25 nm nano-MOSFET device. The approach is parameter free as the size of the electron depends upon its energy. We have justified the correctness of the approach with simulations of the gate voltage dependence of the sheet electron density. The excellent agreement between the simulations and SCHRED results suggests that one is able to correctly predict the effective oxide thickness increase due to quantum-mechanical size-quantization effects that leads to a reduction of the sheet electron density. The nano-MOSFET simulation results also confirm this charge displacement effect near the source end of the channel where quantization effects play significant role. Due to the larger smearing of the potential for high energy electrons, we see a decrease in the carrier velocity when quantization effects are included in the model. This leads to smaller drain current in both the device transfer and output characteristics. The charge displacement from the interface, and the effective increase of the oxide thickness, gives rise to a threshold voltage shift of ~220 mV which is consistent with earlier observations. The shift in the threshold voltage leads, in turn, to a drain current degradation of about 30 %. Hence, the observations presented here, that utilize the new effective potential approach, confirm that quantum-mechanical space-quantization effects must be included in the theoretical model to correctly predict the device behavior. In some cases, this can be achieved with the incorporation of the barrier field that is pre-computed in the initial stages of the simulation and does not require additional CPU time during the simulation sequence. We believe that this new effective potential approach is more reliable in simulation of quantization effects in nano-scale devices with barriers that have different size and shape.

To treat the short-range Coulomb (electron-ion and electron-electron) interactions properly, *three* different but consistent real-space *molecular dynamics* (MD) schemes have been implemented in the simulator: the particle-particle-particle-mesh (P^3M) method, the corrected Coulomb approach and the Fast Multipole Method (FMM). It is believed that the FMM algorithm has been used for the first time in the simulations of semiconductor devices. The correctness of the approaches is verified via the simulations of the doping dependence of the low-field electron mobility in a 3D resistor and through its comparison with available experimental data. These approaches are then applied in the investigations of the role of unintentional doping on the operation of narrow-width SOI devices. We find significant correlation between the location of the impurity atom and the magnitude of the drain current. Namely, impurities near the source end of the channel have maximum influence on the drain current. This observation suggests that one has to take into account transistor mismatches due to unintentional doping when performing circuit designs. We have also investigated in depth the fluctuations in the threshold voltage due to discrete distribution of the impurity atoms in narrow width SOI devices with

10 nm and 5 nm channel width. The simulated data for the threshold voltage are in perfect agreement with the experimental values and do explain the fluctuations in the experimentally derived threshold voltage data.

Another device structure that has been investigated regarding the influence of the discrete impurities is the FinFET. Among different double gate structures FinFET attracts the researchers due to its inherent immunity to short channel effects and ease of fabrication using the existing planar fabrication process flow. Single fin FinFET can easily be extended to multiple fin structure for higher drive current. Again, in this structure as well, we find significant correlation between the magnitude of the drain current and the position of the discrete dopant for the case when screening effects do not play considerable role.

Acknowledgements

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.* **46**, 317 (2002).

3   *International Technology Roadmap for Semiconductors*, 2002 Edition, Semiconductor Industry Association (SIA), Austin, Texas: SEMATECH, USA, 2706 Montopolis Drive, Austin, Texas 78741; http://www.itrs.net/ntrs/publntrs.nsf

4   G. Moore, *IEDM Tech. Digest*, 11 (1975)

5   R. Dennard, F. H. Gaensslen, H. N. Yu, L. Rideout, E. Bassous, and A. R. LeBlanc, *IEEE J. Solid State Circuits* **9**, 256 (1974).

6   B. Yu *et al*., *IEDM Tech. Dig*., 937 (2001).

7   Robert Chau, B. Boyanov, B. Doyle, M. Doczy, S. Datta, S. Hareland, B. Jin, J. Kavalieros, and M. Metz, *4th Int. Symp. on Nanostructures and Mesoscopic Systems*, 17 (2003).

8   H. S. Wong, *IBM J., Res. & Dev*. **46**, 133 (2002).

9   W. Zhu, J. P. Han, T. P. Ma, *IEEE Trans. Electron Dev*. 51, 98 (2004).

10 Welser, J. L. Hoyt and J. F. Gibbons, *IEDM Tech. Dig*., 1000 (1992).

11 G. Formicone, D. Vasileska, D.K. Ferry, *VLSI Design* **6**, 167 (1998).

12 D. Vasileska, G. Formicone and D.K. Ferry, *Nanotechnology* **10**, 147 (1999).

13 P. M. Garone, V. Venkataraman, and J. C. Sturm, *IEEE Electron Device Lett*. **13**, 56 (1992).

14 World-wide-web: http://eetimes.com/semi/news/OEG20021210S0049

15 M. Leong, H.-S. Wong, E. Nowak, J. Kedzierski, E. Jones, *ISQED*, 492 (2002).

16 Geppert, *IEEE Spectrum*, April 9, 2004.

17 Vasileska, Goodnick

18 Ferry, Goodnick, Transport in Nanostructures

19 Fischetti, Ren

20 N. Sano, A. Hiroki, K. Matsuzawa, *IEEE Trans. Nanotechnology* **1**, 63 (2002).

21 Irena Knezevic, *Ph. D. Dissertation*, Arizona State University, August 2004.

22 R. W. Keyes, *Appl. Phys*. **8**, 251 (1975).

23 T. Mizuno, J. Okumtura, A, Toriumi, *IEEE Trans. Electron Devices* **41**, 2216 (1994).

24 M. V. Fischetti, *Journal of Computational Electronics* **2**, 73 (2000).

25 Coulomb drag experiments

26 H. S. Wong and Y. Taur, in *Proc. IEDM*, 29.2.1 (1993).

27 Vasileska, Discrete impurities, VLSI design paper

28 A. Asenov, *IEEE Trans. Electron Devices* **45**, 2505 (1998).

29 William Gross PhD thesis

30 N. Sano, K. Matsuzawa, M. Mukai, and N. Nakayama, *in Proc. IEDM Tech. Digest*, 275 (2000).

31 D. K. Ferry, A. M. Kriman, M. J. Kann, and R. P. Joshi, *Comp. Phys. Comm*. **67**, 119 (1991).

32 L. R. Logan and J. L. Egley, *Phys. Rev. B* **47**, 12532 (1993).

33 C. Jacoboni and P. Lugli, *The Monte Carlo Method for Semiconductor Device Simulation* (New York: Springer-Verlag, 1989).

34 R. H. Dennard, F. H. Gaensslen, H.-N. Yu, V. L. Rideout, E. Bassous and A. R. leBlanc, *IEEE J. Solid-State Circuits* **9**, 256 (1974).

35 J. R. Brews, W. Fichtner, E. H. Nicollian and S. M. Sze, *IEEE Electron Dev. Lett.* **1,** 2 (1980).

36 G. Bacarani and M. R. Worderman, in *Proceedings of the IEDM*, 278 (1982).

37 M.-S. Liang, J. Y. Choi, P.-K. Ko and C. Hu, *IEEE Trans. Electron Devices* **33**, 409 (1986).

38 A. Hartstein and N. F. Albert, *Phys. Rev. B* **38**, 1235 (1988).

39 M. J. van Dort, P. H. Woerlee, A. J. Walker, C. A. H. Juffermans and H. Lifka, *IEEE Trans. Electron Dev*. **39**, 932 (1992).

40 M. J. van Dort, P. H. Woerlee and A. J. Walker, *Solid-State Electronics* **37**, 411 (1994).

41 D. Vasileska and D. K. Ferry, in the *Technical Proceedings of the First International Conference on Modeling and Simulation of Microsystems, Semiconductors, Sensors and Actuators*, 408 (1998).

42 S. Takagi and A. Toriumi, *IEEE Trans. Electron Devices* **42**, 2125 (1995).

43 S. A. Hareland, S. Krishnamurthy, S. Jallepali, C.-F. Yeap, K. Hasnat, A. F. Tasch, Jr. and C. M. Maziar*, IEEE Trans. Electron Devices* **43**, 90 (1996).

44 D. Vasileska, D. K. Schroder and D. K. Ferry, *IEEE Trans. Electron Devices* **44**, 584 (1997).

45 K. S. Krisch, J. D. Bude and L. Manchanda, *IEEE Electron Dev. Lett*. **17**, 521 (1996).

46 L. de Broglie, *C. R. Acad. Sci. Paris* **183**, 447 (1926).

47 L. de Broglie, *C. R. Acad. Sci. Paris* **184**, 273 (1927).

48 E. Madelung, *Z. Phys.* **40**, 322 (1926).

49 D. Bohm, *Phys. Rev*. **85**, 166 (1952).

50 D. Bohm, *Phys. Rev*. **85**, 180 (1952).

51 C. Dewdney and B. J. Hiley, *Found. Phys*. **12**, 27 (1982).

52 G. J. Iafrate, H. L. Grubin, and D. K. Ferry, *Journal de Physique* **42** (Colloq. 7), 307 (1981).

53 E. Wigner, *Phys. Rev*. **40**, 749 (1932).

54 D. K. Ferry and J.-R. Zhou, *Phys. Rev. B* **48**, 7944 (1993).

55 P. Feynman and H. Kleinert, *Phys. Rev. A* **34**, 5080 (1986).

56 C. L. Gardner and C. Ringhofer, *Phys. Rev. E* **53**, 157 (1996).

57 C. Ringhofer and C. L. Gardner, *VLSI Design* **8**, 143 (1998).

58 D. Vasileska, S. Ahmed, IEEE TED

59 Ferry, Superlattices and Microstructures

60   C. Ringhofer, S. Ahmed and D. Vasileska, *Journal of Computational Electronics* **2**, 113 (2003).

61   C. Ringhofer, C. Gardner and D. Vasileska, *Inter. J. on High Speed Electronics and Systems* **13**, 771 (2003).

62   S.S. Ahmed, PhD Disertation

63   AVR Superlattices and Microstructices

64   Y. Omura, S. Horiguchi, M. Tabe, and K. Kishi, *IEEE Elec. Device Lett.* **14**, 569 (1993).

65   S. M. Ramey and D. K. Ferry, *IEEE Transactions on Nanotechnology* **2**, (2003).

66   S. Hasan, J. Wang, and M. Lundstrom, *Solid–State Elect.* **48**, 867 (2004).

67   S. Datta book

68   T. Mizuno, J. Okamura and A. Toriumi, *IEEE Trans. Electron Devices* **41**, 2216 (1994).

69   T. Mizuno, *Jpn. J. Appl. Phys.* **35**, 842 (1996).

70   J. T. Horstmann, U. Hilleringmann and K. F. Goser, *IEEE Trans. Electron Devices* **45**, 299 (1998).

71   P. A. Stolk, F. P. Widdershoven and D. B. M. Klaassen, *IEEE Trans. Electron Devices* **45**, 1960 (1998).

72   K. Nishinohara, N. Shigyo and T. Wada, *IEEE Trans. Electron Devices* **39**, 634 (1992).

73   J.-R. Zhou and D. K. Ferry, *IEEE Comput. Science and Eng.* **2**, 30 (1995).

74   D. Vasileska, W. J. Gross, V. Kafedziski and D. K. Ferry, *VLSI Design* **8**, 301 (1998).

75   D. Vasileska, W. J. Gross and D. K. Ferry, *Extended Abstracts IWCE-6*, Osaka 1998, IEEE Cat. No. 98EX116, 259.

76   X. Tang, V. K. De and J. D. Meindl, *IEEE Trans. on VLSI Systems* **5**, 369 (1997).

77   P. Lugli and D. K. Ferry, *IEEE Trans. Electron Dev.* **32**, 2431 (1986).

78   A. M. Kriman, M. J. Kann, D. K. Ferry and R. Joshi, *Phys. Rev. Lett.* **65**, 1619 (1990).

79   W. J. Gross, D. Vasileska, and D. K. Ferry, *VLSI Design* **10**, 437 (2000).

80   D. Vasileska, W. J. Gross, and D. K. Ferry, *Superlattices and Microstructures* **27**, 147 (2000).

81   A. Asenov, *IEEE Trans. Electron Dev.* **45**, 2505 (1998).

82   A. Asenov and S. Saini, *IEEE Trans. Electron Dev.* **46**, 1718 (1999).

83   L. Greengard and V. Rokhlin, *J. Comput. Phys.* **135**, 280 (1997).

84   R. Beatson and L. Greengard, "A short course on fast multipole methods," in Wavelets, Multi-level Methods and Elliptic PDEs (Leicester, 1996), ser. *Numer. Math. Sci. Comput.* New York: Oxford Univ. Press, pp. 1–37, 1997

85   H. Cheng, L. Greengard, and V. Rokhlin, *J. Comput. Phys.* **155**, 468 (1999).

86   FMMPART3D user's guide, version 1.0 ed., *MadMax Optics*, Hamden, CT, USA.

87   Our IEEE TED paper on unintentional doping

88   R. W. Hockney and J. W. Eastwood, *Computer Simulation Using Particles* (New York, McGraw-Hill, 1981).

89   C. J. Wordelman and U. Ravaioli, *IEEE Tran. Electron Devices* **47**, 410 (2000).

90   W. J. Gross, D. Vasileska, and D. K. Ferry, *IEEE Electron Devices* **47**, 1831 (2000).

91   Allen, D. Holberg, *CMOS Analog Circuit Design* (Saunders College Publishing, New York, 1987).

92   Bohr, Y. A. El-Mansy, *IEEE Trans. Electron Dev.* **45**, 620 (1998).

93   E. H. Nicollian and A. Goetzberger, *Bell Syst. Techn. J.* **46**, 1055 (1967).

94   J. T. Horstmann, U. Hilleringmann and K. F. Goser, *IEEE Trans. Electron Devices* **45**, 299 (1998).

95   P. A. Stolk, F. P. Widdershoven and D. B. M. Klaassen, *IEEE Trans. Electron Devices* **45**, 1960 (1998).

96   GVF IEEE TED

97   H. Majima, H. Ishikuro, and T. Hiramoto, *IEEE Electron Dev. Lett.* 21, 396 (2000).

98   Khan, IWCE-10

99   L. Chang et al., *IEDM* Tech. Dig., 719 (2002)

100 Clemens IWCE-10