SANDIA REPORT

SAND200X-XXXX
Unlimited Release
Printed October 2014

The ultimate downscaling limits of FETs

Denis Mamaluy, Xujiao Gao, Brian Tierney

Prepared by
Sandia National Laboratories
Albuquerque, New Mexico 87185 and Livermore, California 94550

Sandia National Laboratories is a multi-program laboratory managed and operated by Sandia Corporation,
a wholly owned subsidiary of Lockheed Martin Corporation, for the U.S. Department of Energy's
National Nuclear Security Administration under contract DE-AC04-94AL85000.

Issued by Sandia National Laboratories, operated for the United States Department of Energy by Sandia Corporation.

NOTICE: This report was prepared as an account of work sponsored by an agency of the United States Government. Neither the United States Government, nor any agency thereof, nor any of their employees, nor any of their contractors, subcontractors, or their employees, make any warranty, express or implied, or assume any legal liability or responsibility for the accuracy, completeness, or usefulness of any information, apparatus, product, or process disclosed, or represent that its use would not infringe privately owned rights. Reference herein to any specific commercial product, process, or service by trade name, trademark, manufacturer, or otherwise, does not necessarily constitute or imply its endorsement, recommendation, or favoring by the United States Government, any agency thereof, or any of their contractors or subcontractors. The views and opinions expressed herein do not necessarily state or reflect those of the United States Government, any agency thereof, or any of their contractors.

Printed in the United States of America. This report has been reproduced directly from the best available copy.

Available to DOE and DOE contractors from
U.S. Department of Energy
Office of Scientific and Technical Information
P.O. Box 62
Oak Ridge, TN 37831

Telephone: (865) 576-8401
Facsimile: (865) 576-5728
E-Mail: reports@adonis.osti.gov
Online ordering: http://www.osti.gov/bridge

Available to the public from
U.S. Department of Commerce
National Technical Information Service
5285 Port Royal Rd.
Springfield, VA 22161

Telephone: (800) 553-6847
Facsimile: (703) 605-6900
E-Mail: orders@ntis.fedworld.gov
Online order: http://www.ntis.gov/help/ordermethods.asp?loc=7-4-0#online

SAND200X-XXXX
Unlimited Release
Printed October 2014

The ultimate downscaling limit of FETs

Denis Mamaluy (1425), Xujiao Gao (1355), Brian Tierney (1425)
Sandia National Laboratories
P.O. Box 5800
Albuquerque, New Mexico 87185-MS1322

Abstract

We created a highly efficient, universal 3D quantum transport simulator. We demonstrated that the simulator *scales linearly* – both with the problem size (N) and number of CPUs, which presents an important break-through in the field of computational nanoelectronics. It allowed us, for the first time, to accurately simulate and optimize a large number of realistic nanodevices in a much shorter time, when compared to other methods/codes such as RGF[$\sim N^{2.333}$]/KNIT, KWANT, and QTBM[$\sim N^3$]/NEMO5. In order to determine the best-in-class for different beyond-CMOS paradigms, we performed rigorous device optimization for high-performance logic devices at 6-, 5- and 4-nm gate lengths. We have discovered that there exists a fundamental down-scaling limit for CMOS technology and other Field-Effect Transistors (FETs). We have found that, at room temperatures, all FETs, irrespective of their channel material, will start experiencing unacceptable level of thermally induced errors around 5-nm gate lengths.

ACKNOWLEDGMENTS

Authors would like to thank our excellent manager John Aidun (1425) for the support, Erik DeBenedictis (1425) and Wolfgang Porod (University of Notre Dame) for very useful discussions.

This work was funded under LDRD Project #165618 and Title "A Universal Quantum Transport Computational Capability for Cross-Technology Comparisons of Beyond-CMOS Nano-electronic Devices".

CONTENTS

FIGURES

NOMENCLATURE

FET	field-effect transistor
CMOS	complementary metal–oxide–semiconductor
TFET	tunneling field-effect transistor
SET	single electron transistor
CNT	carbon nano-tube
MuGFET	multiple gate field-effect transistor
FinFET	MuGFET utilizing a thin silicon "fin" which forms the device body
ITRS	International Technology Roadmap for Semiconductors

DOS	density of states
LDOS	local density of states
NEGF	non-equilibrium Green’s function
CBR	Contact Block Reduction
RGF	recursive Green’s function
QTBM	quantum-transmitting boundary method

EXECUTIVE SUMMARY

How small can transistors, the basic building blocks of modern electronics that currently drive $250-billion semiconductor industry, get? Are there any fundamental limits for Moore's law – besides the obvious atomic size limit? If so, are there any ways to overcome such limits? – These are the questions that we asked ourselves when we began EC LDRD project of creating a universal quantum transport device simulator. The simulator uses a very efficient algorithm developed by the PI that allows us to dramatically reduce computational costs of quantum transport calculations and hence vastly increases our ability to simulate a large variety of different ultra-scaled CMOS and beyond-CMOS structures at the end of ITRS roadmap for semiconductor devices. Surprisingly, we have discovered that at room temperatures, all field-effect-transistors (FETs), *irrespective of their channel material*, will start experiencing unacceptable levels of thermally induced errors when gate lengths approach 5nm. This prediction effectively means that the end is in sight for downscaling and Moore's law for all FETs, including the most crucial FET technology – CMOS. According to the current ITRS projections, 5nm gate lengths will occur in about 15 years from now. What are the industry possibilities after the thermal fluctuation limit is reached? We see at least three:

- Accept the end of Moore's law and concentrate efforts on power dissipation reduction with the reversible/adiabatic computing
- Develop non-FET alternatives: memristors, super-conductive logic, etc.
- Continue Moore's law, using single-electron transistors!

The latter option seems to be the most exciting to us: indeed there will be still a lot of "room" below 5nm-gate lengths (atomic size is about 0.2 nm) and any possibility to extend Moore's law, which has been so beneficial to world's economy, should be closely investigated.

INTRODUCTION

The ultimate end of CMOS scaling was predicted almost immediately after the now ubiquitous technology was invented by Frank Wanlass [1] in 1963. Indeed, many possible limitations to downscaling were discussed in the 1970s, 80s, and 90s [2]. In 2003 Zhirnov *et al.*, estimated [3] the minimal feature size of a "binary logic switch" to be around 1.5nm, based on the Heisenberg uncertainty and Landauer principles. Since then, there have been many papers [2,4,5] discussing the likely end of CMOS scaling due to lithographical, power-thermal, material, and other technological, as opposed to fundamental physical, limitations.

Despite the aforementioned predictions, however, CMOS has famously survived, albeit with adaptations (high-k gate dielectrics, revival of metal gates, etc.). Furthermore, the immense increase in the understanding of semiconductor physics since the 1960s has resulted in a plethora of "alternative CMOS" technologies that are generally FET-based. In fact, arguments are frequently made that III-V-, CNT-, or 2D-material-based FETs have the potential to someday replace present Si FETs due to their superior mobilities and ultra-scale manufacturing capability. However, proponents of these devices have also made it evident that such devices are not yet ready to compete with state-of-the-art Si FinFETs for high performance computing applications. In fact, ITRS now projects [6] that the emerging trend of Si FinFET technology should allow Moore's law to continue for at least another decade until 6-nm gate lengths are reached.

In this work we present our findings from EC LDRD #165618 "A Universal Quantum Transport Computational Capability for Cross-Technology Comparisons of Beyond-CMOS Nanodevices" by D. Mamaluy (PI) and X. Gao. Within the scope of this LDRD we have created a universal quantum transport simulator that allows us to assess and compare, within a reasonable simulation time, the performance of the different types of beyond-CMOS transistor technologies. The simulation tool is based on a novel numerical method called Contact Block Reduction (CBR) [7,8] that was developed by the PI. The CBR method provides an efficient and accurate implementation of the Keldysh Non-Equilibrium Green's Function (NEGF) formalism [9] for quantum transport simulation that turns out to be significantly faster than other existing methods, as identified by independent reviews [10,11].

This report is organized as follows: a brief description of the EC LDRD project is given, followed by a presentation of the details of the of the universal quantum transport simulator, CBR3D. We then list and analyze the most significant findings obtained with CBR3D. This report is concluded with a discussion of further ideas for research and the associated directions industry may consider after the thermal fluctuation limit is reached.

1. UNIVERSAL QUANTUM TRANSPORT SIMULATOR: CBR3D

1.1. Motivation

As industry and academia work to increase the speed of transistors by shrinking their size to nanometer dimensions with gate lengths of 10 nm or less, conventional theories used to predict device behavior are becoming obsolete because they fail to account for quantum effects. Existing semi-classical Technology Computer Aided Design (TCAD) tools work only with conventional CMOS devices. While there has been a significant effort in the computational nanoelectronics community to simulate a *particular* promising novel beyond-CMOS device, until now there has been no common tool that allows performance comparisons of *different types* of beyond-CMOS devices. Examples of these devices include ultra-scaled Si and III-V FinFETs, carbon nanotube FETs, graphene-based and other two-dimensional material-based transistors, tunneling FETs.

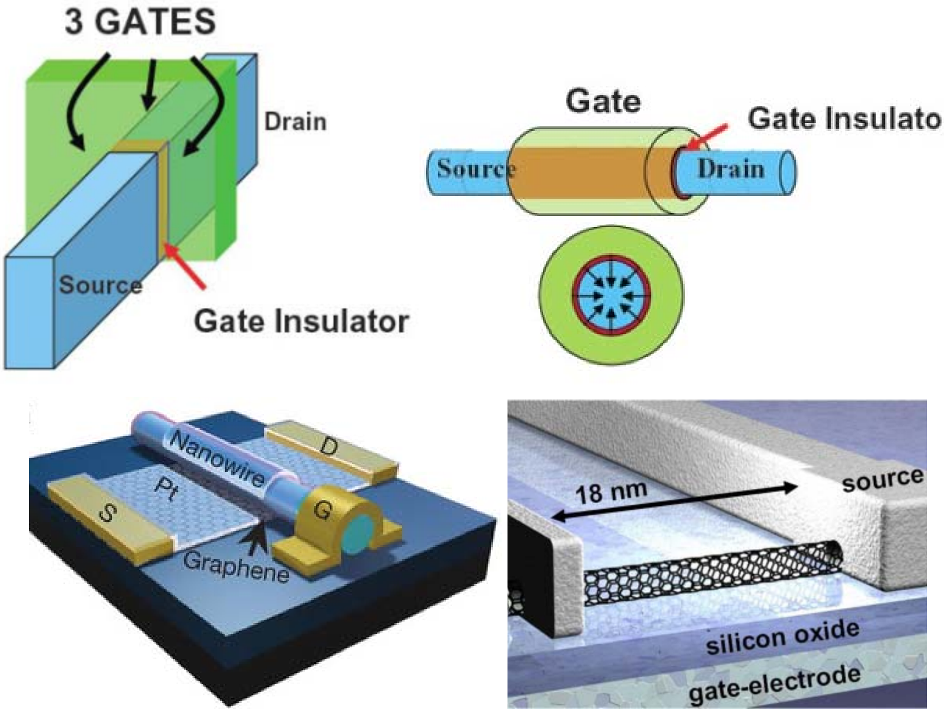


Figure 1. Some examples of the most promising emerging alternatives to conventional CMOS: FinFET (Tri-gate) (top-left), nanowire (top right), graphene-based and carbon nanotube transistors (lower left and right, respectively).

The CBR3D code enables such cross-technology comparisons and can be invaluable for guided development of advanced nanoelectronics, including the corresponding experimental and fabrication efforts.

1.2. Overview of the existing methods and simulators

A number of powerful methods have been developed through time to calculate the quantum transport through nano-structures. A well-known approach is the *transfer matrix method* [12,13]. While this approach is unstable [14] for larger devices in its original form, this drawback was overcome by a series of generalizations developed by Frensley [15], Lent *et al.* [16], and Ting *et al.* [17]. These approaches use the *quantum transmitting boundary method* (QTBM) [16] to account for the coupling to the leads, and can handle structures of arbitrary geometry. There are QTBM implementations applied to one-dimensional tight-binding Hamiltonian [17,18] and k-dot-p based [19] multi-band calculations, and two-dimensional (2D) single-band calculations [16,20]. A three-dimensional self-consistent scheme based on QTBM has been presented [21] in a case when device potential is separable. The *boundary element method* [22] is more computationally efficient, but so far the published applications are limited to wave-guide structures, i.e. structures possessing a flat potential [23] or consisting of piecewise homogeneous materials with constant potentials [24]. Another efficient and widely used algorithm is the *recursive Green's function* (RGF) method [25,26] that has been successfully implemented for two-dimensional devices [27,28] and for small three-dimensional structures such as nano-wires [29]. It is well suited for 2-terminal devices that can be discretized into cross-sectional slices with nearest neighbor interactions, but has serious difficulties dealing with additional (*i.e.* more than two) contacts [7]. A closely related *modular recursive Green's function* method [30] is applicable to devices that can be divided into regions of sufficiently high symmetry, where the Schrödinger equation is separable, and has been adapted to include magnetic fields [31]. Another method that is applicable for the case of spatially separable device potentials and when current can flow through two leads (*e.g.* in the situation of quasi-1D transport) has been termed the *mode-space* approach [28]. This method has been implemented in the effective-mass simulator NanoMOS that has been extended to arbitrary crystallographic directions [32,33]. A 3D simulator within the validity of the effective-mass approximation using the mode-space approach for silicon nanotransistor has been reported where scattering is taken into account by Buttiker probes [34]. An application of the mode-space approach to ultra-small FinFET simulations has been presented [35] in which phonon scattering and surface roughness have been taken into account. Another simulator for 3D silicon nanowires based on the effective-mass approximation and the mode-space approach has been presented in [36]. In that work the simulation was performed assuming that the device potential is separable in the confinement direction (mode representation) and the transport direction, along which the potential is assumed to be nearly uniform. The resulting quasi-1D transport problem is solved using a simplified NEGF formalism self-consistently. Finally, a modified version of the QTBM has been developed that expands the scattering solutions in terms of two different closed system wave functions in an efficient way for 2D systems with arbitrary number of contacts [37].

Thus, among reviewed simulators only a few are able to take into account more than two Ohmic contacts in order to treat gate leakage effects properly, and all of these simulators are 2D. Additionally, many "fully 3D quantum simulators" treat the transport as a quasi-1D problem, wherein it is automatically assumed that there are only two Ohmic contacts in the system and thus the effects of gate leakage are neglected.

In contrast, in our research we utilized *the CBR method* [7,8], within which all open system quantities of interests such as currents and charge density can be obtained from I) eigenstates of a specially defined closed system $H^0|\alpha\rangle = \varepsilon_\alpha |\alpha\rangle$ with generalized Neumann boundary conditions [7] and II) solution of a very small linear algebraic system of the size of the "contacts" (boundary regions between the active device and external leads) for every energy step. The quantum transport simulator based on the CBR method has been developed by the PI and allows calculating transport properties of 2D and 3D nano-devices that may have arbitrary shape, potential profile, and, notably, any number of external contacts, which differentiates it from simulators based on the abovementioned recursive Green's functions method (e.g. NEMO5 [38]). The CBR simulator takes into account surface and interface roughness, scattering on impurities, discrete dopants, phonons and electron-electron interaction. The charge self-consistency is achieved by adopting the predictor-corrector approach, which has been shown to reach the convergence within only 5-7 iterations [8,39].

In the following sections, we will present some details of the underlying CBR method of the newly created universal quantum transport simulator CBR3D and will discuss the code scalability with problem size and number of CPUs.

1.3. The Contact Block Reduction (CBR) method

In the CBR method, quantities such as the transmission function and charge density of the open system can be obtained from the eigenstates of a corresponding closed system $H^0 |\alpha\rangle = E_\alpha |\alpha\rangle$, and a solution of a very small linear algebraic system for every energy step. Remarkably, we have shown that the calculation of relatively few eigenstates of the closed system is sufficient to obtain accurate results [7,8].

The first step of the CBR method consists of formally dividing the device space into two regions: the boundary region corresponding to the contact with the leads, C, and the region corresponding to the rest of the device, D. The self-energy matrix, representing the coupling of the device to the leads, is non-zero only in the region C and therefore has the following structure $\mathbf{\Sigma} = \begin{bmatrix} \mathbf{\Sigma}_C & \mathbf{0}_{CD} \\ \mathbf{0}_{DC} & \mathbf{0}_D \end{bmatrix}$, where $\mathbf{0}$'s denote the zero-matrices for the corresponding regions. In a similar manner, we can subdivide all the Green's function matrices. In particular, the Green's function of the decoupled device (i.e. the corresponding closed system) can be written as $\mathbf{G}^0 = \begin{bmatrix} \mathbf{G}_C^0 & \mathbf{G}_{CD}^0 \\ \mathbf{G}_{DC}^0 & \mathbf{G}_D^0 \end{bmatrix}$. It has been shown [7,8] that after some algebraic manipulations, this leads to the following structure of the retarded Green's function matrix of the *open* system:

$$\mathbf{G}^{\mathbf{R}} = \begin{bmatrix} \mathbf{A}_C^{-1}\mathbf{G}_C^0 & \mathbf{A}_C^{-1}\mathbf{G}_{CD}^0 \\ \mathbf{G}_{DC}^0 \mathbf{\Sigma}_C \mathbf{A}_C^{-1}\mathbf{G}_C^0 + \mathbf{G}_{DC}^0 & \mathbf{G}_{DC}^0 \mathbf{\Sigma}_C \mathbf{A}_C^{-1}\mathbf{G}_{CD}^0 + \mathbf{G}_D^0 \end{bmatrix}, \text{ where } \mathbf{A}_C = \mathbf{1}_C - \mathbf{G}_C^0 \mathbf{\Sigma}_C.$$

Of note, this equation shows that all elements of the retarded Green's function can be calculated by inverting only a small matrix, $\mathbf{A}_C$, of the same size as the interface (termed ***contacts***) between the device and external leads. In Ref. [43] it was shown that the transmission function is determined only by the elements of the following submatrix, of the same small size as the contacts: $\mathbf{G}_C^R = \mathbf{A}_C^{-1}\mathbf{G}_C^0$. In Refs. [7,8,39] it has been shown that any other quantity of interest, such as the density matrix, particle (charge) density, *etc.*, can be found with similar computational effort. Indeed, the density matrix $\xi_{\alpha\beta}$ can be calculated as:

$$\begin{aligned}
&\boldsymbol{\xi}_{\alpha\beta} = \sum_{\lambda=1}^{L} \int \Xi_{\alpha\beta}^{(\lambda)}(E) f_{\lambda}(E)\, dE, \\
&\boldsymbol{\Xi}_{\alpha\beta}^{(\lambda)}(E) = \frac{1}{2\pi} \frac{\mathrm{Tr}\left(\left[|\beta\rangle\langle\alpha|\right]_C \mathbf{B}_C^{-1} \boldsymbol{\Gamma}_C^{(\lambda)} \mathbf{B}_C^{-1\dagger}\right)}{\left(E-\varepsilon_{\alpha}+i\eta\right)\left(E-\varepsilon_{\beta}-i\eta\right)}\Bigg|_{\eta\to 0+},
\end{aligned}$$

where the equilibrium distribution functions within each lead are denoted by $f_{\lambda}(E)$; the matrix $\mathbf{B}_C = \mathbf{1}_C - \boldsymbol{\Sigma}_C \mathbf{G}_C^0$ and all the other matrices are of the size of the *contacts* only; $|\alpha\rangle$ are the eigenfunctions of the closed (decoupled) system. Then the charge density can be expressed as

$$n(\mathbf{r}) = \sum_{\alpha,\beta} \psi_{\alpha}(\mathbf{r}) \psi_{\beta}(\mathbf{r}) \xi_{\alpha\beta} .$$

The current $J_{\lambda\lambda'}$ from lead λ to lead λ' can be expressed in terms of the transmission function $T_{\lambda\lambda'}(E)$ and the distribution functions of the leads (the most general form has been given in [44])

$$J_{\lambda\lambda'} = \frac{2e}{h} \int T_{\lambda\lambda'}(E)\left[f_{\lambda}(E) - f_{\lambda'}(E)\right] dE,$$

where e is the electron charge, h is Planck's constant, and E is the energy. Finally, the transmission function $T_{\lambda\lambda'}(E)$ and the reflection coefficient $R_{\lambda}(E)$ can be obtained from the *contact part* of the retarded Green's function G^R of the open device

$$\begin{aligned}
&T_{\lambda\lambda'}(E) = \mathrm{Tr}\left(\boldsymbol{\Gamma}_C^{\lambda} \mathbf{G}_C^{R} \boldsymbol{\Gamma}_C^{\lambda'} \mathbf{G}_C^{R\dagger}\right),\ \lambda \neq \lambda \\
&R_{\lambda}(E) = T_{\lambda\lambda'}(E) = \mathrm{Tr}\left(\left[\mathbf{1}_C - i\boldsymbol{\Gamma}_C^{\lambda} \mathbf{G}_C^{R}\right]\left[\mathbf{1}_C - \boldsymbol{\Gamma}_C^{\lambda'} \mathbf{G}_C^{R\dagger}\right]\right).
\end{aligned}$$

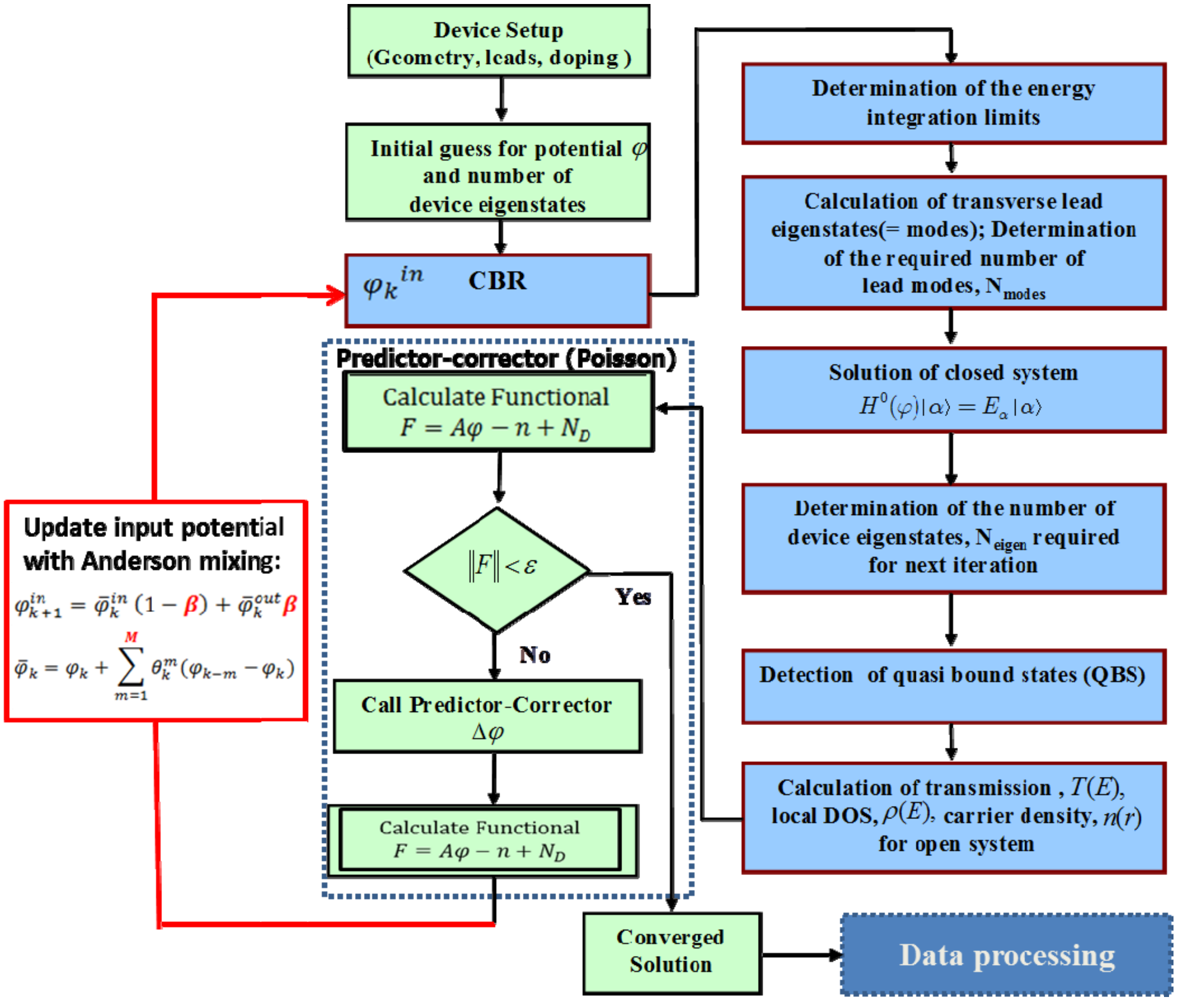
The crucial property of the CBR expressions shown above is that all quantities of interest can be found from a small part of the retarded Green's function matrix corresponding to the contact region.

Finally, the calculation of the Green's function of the decoupled device $\mathbf{G^0}$ is performed through its spectral representation $\mathbf{G}_C^0(E) = \sum_{\alpha} \frac{|\alpha_C\rangle\langle\alpha_C|}{E - E_{\alpha}}$, and is only needed in the contact region. The eigenfunctions $|\alpha\rangle$ are obtained by solving the Schrödinger equation of the decoupled device employing *generalized Neumann boundary conditions* [7] at the contacts. The use of these boundary conditions drastically reduces the required range of the eigenvalue spectrum in the calculation of the transmission function. As such, only a few percent of the eigenvalues are usually sufficient. The remaining task of solving the Dyson equation for each energy value is also reduced, typically by an order of magnitude, because only propagating modes need to be taken into account, providing that the device region is defined properly [7,8].

The CBR provides an extremely efficient method to solve the quantum transport problem in an open system. The CBR computational cost is primarily determined by the partial solution of the Hermitian eigenvalue problem of a closed system. The CBR method is capable of taking into account several connected leads precisely (i.e. without neglecting any off-diagonal elements of the self-energy, unlike in the recursive Green's function method, see e.g. [7,8]). This makes it possible to apply the method to complicated 3D structures with an **arbitrary number of leads**. The CBR method has been generalized to multi-band Hamiltonians [7], and fully charge **self-consistent** calculations [8]. To achieve the charge self-consistency in our simulator we are using the predictor-corrector approach [42], modified for open systems [8]. As discussed in the following section, we can obtain a solution converged up to the third significant digit of the current within a few Poisson-(open) Schrödinger cycles.

1.4. Self-consistent CBR method with Anderson acceleration

For a detailed description of the self-consistent procedure please refer to Ref. [8] and [39]. Here, we summarize the essence of the self-consistent CBR algorithm with the flowchart shown in Figure 2.



Device Setup
(Gcometry, lcads, doping)
Initial guess for potential φ
and number of
device eigenstates
$\varphi_k{}^{in}$ CBR
Predictor-corrector (Poisson)
Calculate Functional
$F = A\varphi - n + N_D$
$\|F\| < \varepsilon$
Yes
No
Call Predictor-Corrector
$\Delta\varphi$
Calculate Functional
$F = A\varphi - n + N_D$
Update input potential
with Anderson mixing:
$\varphi_{k+1}^{in} = \bar{\varphi}_k^{in}(1-\beta) + \bar{\varphi}_k^{out}\beta$
$\bar{\varphi}_k = \varphi_k + \sum_{m=1}^{M} \theta_k^m(\varphi_{k-m} - \varphi_k)$
Determination of the energy
integration limits
Calculation of transverse lead
eigenstates(= modes); Determination
of the required number of
lead modes, N_{modes}
Solution of closed system
$H^0(\varphi)|\alpha\rangle = E_\alpha|\alpha\rangle$
Determination of the number of
device eigenstates, N_{eigen} required
for next iteration
Detection of quasi bound states (QBS)
Calculation of transmission , $T(E)$,
local DOS, $\rho(E)$, carrier density, $n(r)$
for open system
Converged
Solution
Data processing

Figure 2. Flow-chart of CBR3D algorithm with Anderson acceleration

First, we initialize the system's geometry, material parameters (Si, Ge, III-V, SiO_2, high-k dielectrics, metal work-functions), crystallographic orientation, doping profiles and external lead configuration (number, position and geometry of leads). Then an initial guess is made for the electrostatic potential φ ,either a flat potential (zero, if references to the equilibrium Fermi level) or the potential taken from the previously converged bias point. The energy integration range based on the applied voltages and temperature is then determined, followed by the calculation of transverse lead modes (2D Eigen states) that determine the self-energy. Afterwards comes the first major computational step: solution of the closed-system eigenproblem $H^0\left|\alpha\right\rangle = \varepsilon_\alpha\left|\alpha\right\rangle$ with generalized Neumann boundary conditions [7]. In the CBR3D code we utilize two eigensolvers, ARPACK [40] and FEAST [41]. A comparison of each eigensolver will be given in the next section. After the closed system eigenvectors and eigenvalues are computed, weakly coupled "quasi-bound" states are detected and extracted to facilitate better convergence [8]. The last significant computational step is the energy integration of LDOS, transmission function, and the other open-system quantities of interest that are efficiently computed in the basis of the closed-system (generalized Neumann) states. Once the open-system properties (electron density, LDOS) are determined, we invoke the predictor-corrector algorithm [41,8] to solve the non-linear Poisson equation and satisfy the charge self-consistency. However, to further aid the convergence in highly-resonant systems, we follow up the predictor-corrector method with the Anderson acceleration scheme that, as we found [39], provides faster convergence in many cases, as illustrated in Figure 3.

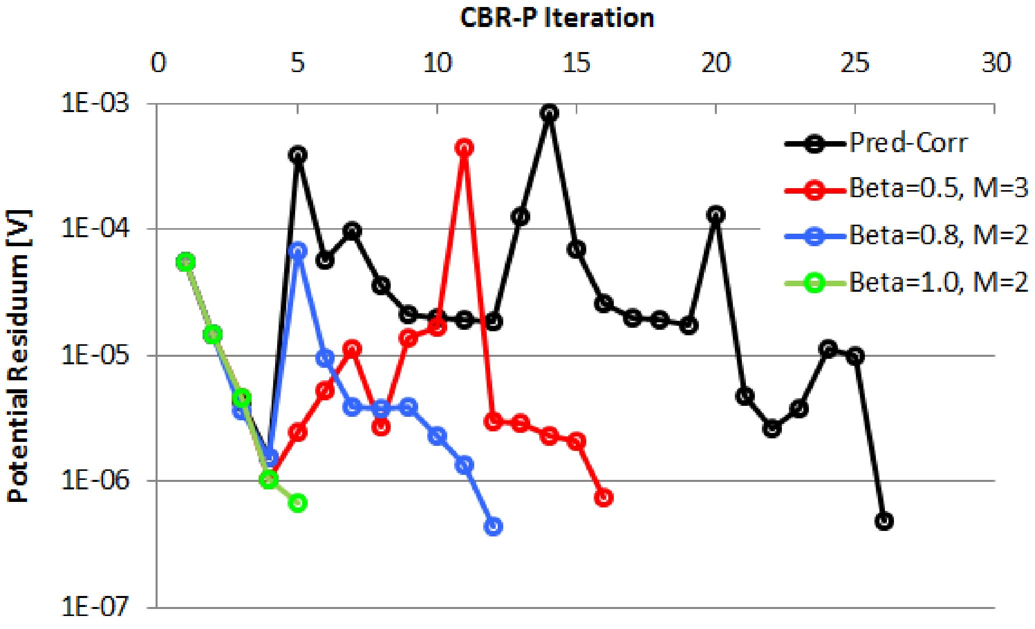


Figure 3. CBR3D: self-consistent loop convergence of the original predictor-corrector method (Pred-Corr) and with the same method with Anderson acceleration technique (colored plots).

1.5. Scaling of CBR3D code with problem size and number of CPUs

The theoretical analysis of the CBR method predicts linear scaling (e.g. [8]) with the problem size N, which can be compared to other methods/codes: Recursive Green Function (RGF) method [~$N^{2.333}$]/KNIT, KWANT, and Quantum transmitting boundary method (QTBM) [~N^3]/NEMO5. However, the CBR is a rather complicated method, so it is important to demonstrate that the linear scaling is not just theory, but the achievable practice. Figure 4 shows total CBR3D iteration time (normalized with size) vs problem size (number of grid points). Perfect linear (~N) scaling is clearly demonstrated by the flattening of the curve with an increasing number of grid points.

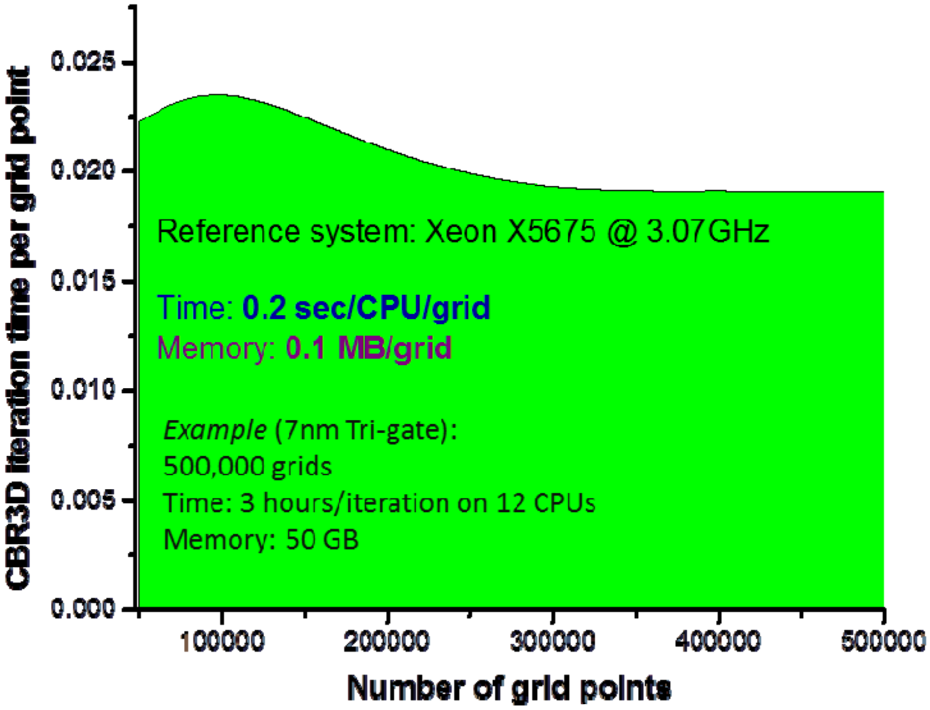


Figure 4. CBR3D scaling with the problem size (number of grid points)

Our results shown in Figure 4 are for a simulated 7-nm gate length Tri-gate device discretized with 10^5 to 10^6 grid points. We find that a typical 500,000 grid simulation requires 3 hours per iteration using a 12-core workstation (CPU: Xeon X5675 @ 3.07GHz, 32 nm Westmere-EP) with 50GB of shared memory. The average number of iterations per bias point is 7. We note that previously simulations of such scale could only be performed on expensive clusters.

We also analyzed the scaling properties of CBR3D code with number of CPUs available on shared memory systems. The latter choice is determined by the nature of the computational problem in the CBR method: we need to solve for the lowest states of a large eigen problem and then perform the energy integration of LDOS. LDOS is a massive array of the size of NxN_E, where N is the size of the problem in real space (i.e. number of grids that discretize the volume of simulated device), typically $N\sim10^5$-10^6; and N_E is the number of energy grid points, typically $N_E\sim1000$-10000. For realistic 6-4nm gate length FET simulations it is sufficient to use $N\sim10^6$ and $N_E\sim2000$; hence, distributed memory systems (clusters) would not be a very efficient choice for these type of problems, since relatively inexpensive shared memory systems ($5k workstations) have sufficient memory.

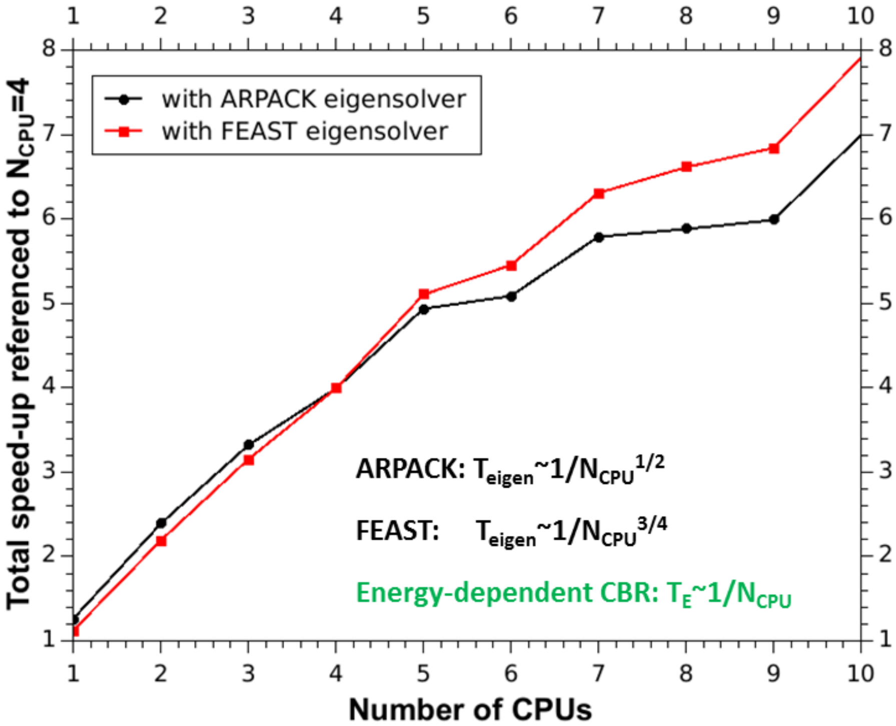


Figure 5. CBR3D scaling with the number of CPUs/cores

Indeed, the required memory estimate for CBR3D is $L*N*N_E*8$ bytes = L*16 GB, where L is the number of independent external leads that may have different applied voltages. For FETs L is equal to 3, and therefore a 48GB shared memory system would be the most efficient choice for these types of calculations. We note that both the solution of eigenproblem and energy integration of LDOS are also much more *efficient* on shared memory systems as illustrated in Figure 5, where we observe an 8x speed-up on a 10-CPU system is observed. This is a rather remarkable achievement, given the relative complexity of the CBR algorithm.

2. PERFOMANCE OF OPTIMIZED NANO-DEVICES AT THE END OF THE ITRS ROADMAP

2.1 The ultimate limit of FET downscaling (overview)

When discussing fundamental limits of FET downscaling, the following three considerations are usually considered pertinent: 1) physical atomic size ~ 0.2 nm, 2) the Landauer principle [45], which states that the switching energy (of a binary switch) must be higher than $kT\ln 2$, and 3) the minimal feature size of a binary switch, estimated by Zhirnov et al. [46] to be 1.5 nm. The latter estimate is based on the Landauer principle and the Heisenberg uncertainty principle: the authors argue that since

$$\begin{aligned} & E_{switch} \geq kT\ln 2 \\ & \begin{cases} \Delta x \Delta p \geq \hbar \\ \Delta E \Delta t \geq \hbar \end{cases} \quad , \\ & \Rightarrow x_{\min} = \frac{\hbar}{\Delta p} = \frac{\hbar}{\sqrt{2m_e E_{switch}}} \end{aligned}$$

the binary switch feature size (e.g. gate length) must be larger than [46]

$$x_{\min} = \frac{\hbar}{\sqrt{2m_e kT\ln 2}} \approx 1.5\,nm\ \left(T = 300K\right).$$

However, it is easy to see that this estimate only applies to the minimal possible (for irreversible process) switching energy: $E_{switch} = kT\ln 2$. It has been shown, however, that modern CMOS architectures cannot operate at such low switching energies due to prohibitively high expenses associated with necessity to compensate for thermally induced errors [47]. In fact, the minimal switching energy of a realistic (FET) transistor that guarantees its error-free lifetime operation is of the order of $E_{switch} = 100kT$ [47], [48]. Thus, for realistic transistors (that operate sufficiently far from the thermal fluctuations/Landauer limit) Zhirnov's estimate is not relevant. Indeed, we have

$$x_{\min} = \frac{\hbar}{\sqrt{2m_e kT100}} \approx 0.12\,nm\ \left(T = 300K\right)$$

which is smaller than the atomic size and therefore does not affect the fundamental downscaling limit of FETs.

2.2 Gate switching energy

Before we proceed further, let us discuss the downscaling limits imposed by the existence of the **minimal switching energy**. The gate switching energy is the amount of energy necessary to switch a FET on/off. It's easy to see that the switching energy is related to the electrostatic energy of the corresponding MOS capacitor:

$$E_{switch} = C_g V^2 = Q_g V$$

We note that the concept of switching energy applies to *all* FETs, including MOSFETs, MuGFETs, TFETs, SpinFETs, SETs, etc., but is not applicable to non-FET devices, such as memristors. We also point out that *any scalable technology* results in the reduction of the gate capacitance, C_g. For CMOS technology, the voltage is roughly constant and therefore the switching energy simply scales down with the gate capacitance.

2.3 Switching energy estimates based on ITRS data

Recent projections for CMOS technology downscaling and characteristics for the next decade [6], published periodically by ITRS, are as illustrated in Figure 6:

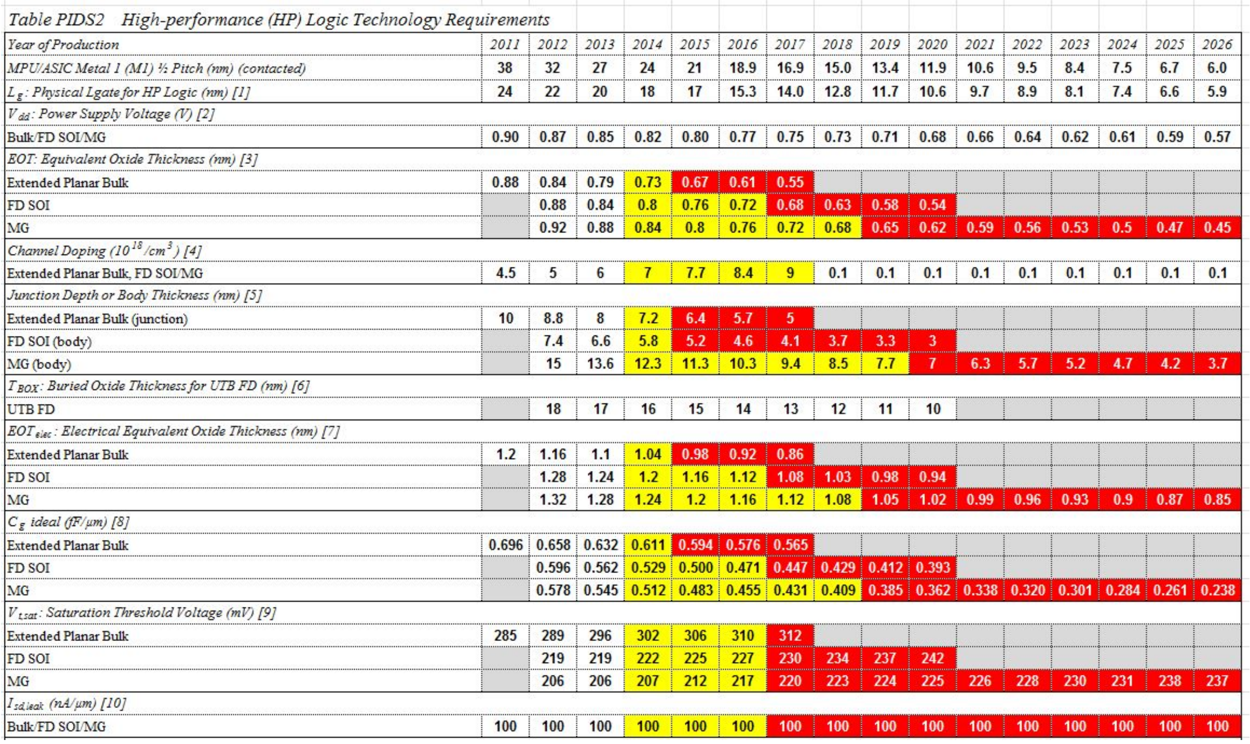


Table PIDS2 High-performance (HP) Logic Technology Requirements

Year of Production	*2011*	*2012*	*2013*	*2014*	*2015*	*2016*	*2017*	*2018*	*2019*	*2020*	*2021*	*2022*	*2023*	*2024*	*2025*	*2026*
MPU/ASIC Metal 1 (M1) ½ Pitch (nm) (contacted)	38	32	27	24	21	18.9	16.9	15.0	13.4	11.9	10.6	9.5	8.4	7.5	6.7	6.0
L_g: Physical Lgate for HP Logic (nm) [1]	24	22	20	18	17	15.3	14.0	12.8	11.7	10.6	9.7	8.9	8.1	7.4	6.6	5.9
V_{dd}: Power Supply Voltage (V) [2]																
Bulk/FD SOI/MG	0.90	0.87	0.85	0.82	0.80	0.77	0.75	0.73	0.71	0.68	0.66	0.64	0.62	0.61	0.59	0.57
EOT: Equivalent Oxide Thickness (nm) [3]																
Extended Planar Bulk	0.88	0.84	0.79	0.73	0.67	0.61	0.55									
FD SOI		0.88	0.84	0.8	0.76	0.72	0.68	0.63	0.58	0.54						
MG		0.92	0.88	0.84	0.8	0.76	0.72	0.68	0.65	0.62	0.59	0.56	0.53	0.5	0.47	0.45
Channel Doping ($10^{18}/cm^3$) [4]																
Extended Planar Bulk, FD SOI/MG	4.5	5	6	7	7.7	8.4	9	0.1	0.1	0.1	0.1	0.1	0.1	0.1	0.1	0.1
Junction Depth or Body Thickness (nm) [5]																
Extended Planar Bulk (junction)	10	8.8	8	7.2	6.4	5.7	5									
FD SOI (body)		7.4	6.6	5.8	5.2	4.6	4.1	3.7	3.3	3						
MG (body)		15	13.6	12.3	11.3	10.3	9.4	8.5	7.7	7	6.3	5.7	5.2	4.7	4.2	3.7
T_{BOX}: Buried Oxide Thickness for UTB FD (nm) [6]																
UTB FD		18	17	16	15	14	13	12	11	10						
EOT_{elec}: Electrical Equivalent Oxide Thickness (nm) [7]																
Extended Planar Bulk	1.2	1.16	1.1	1.04	0.98	0.92	0.86									
FD SOI		1.28	1.24	1.2	1.16	1.12	1.08	1.03	0.98	0.94						
MG		1.32	1.28	1.24	1.2	1.16	1.12	1.08	1.05	1.02	0.99	0.96	0.93	0.9	0.87	0.85
C_g ideal (fF/μm) [8]																
Extended Planar Bulk	0.696	0.658	0.632	0.611	0.594	0.576	0.565									
FD SOI		0.596	0.562	0.529	0.500	0.471	0.447	0.429	0.412	0.393						
MG		0.578	0.545	0.512	0.483	0.455	0.431	0.409	0.385	0.362	0.338	0.320	0.301	0.284	0.261	0.238
V_{tsat}: Saturation Threshold Voltage (mV) [9]																
Extended Planar Bulk	285	289	296	302	306	310	312									
FD SOI		219	219	222	225	227	230	234	237	242						
MG		206	206	207	212	217	220	223	224	225	226	228	230	231	238	237
$I_{sd,leak}$ (nA/μm) [10]																
Bulk/FD SOI/MG	100	100	100	100	100	100	100	100	100	100	100	100	100	100	100	100

Figure 6. ITRS projections for future high-performance logic devices (table PIDS2).

While the specific numerical figures presented in ITRS reports tend to be revised in each new edition, representative data of the continuing downscaling trend for switching energy has been obtained [49] , as shown in Figure 7:

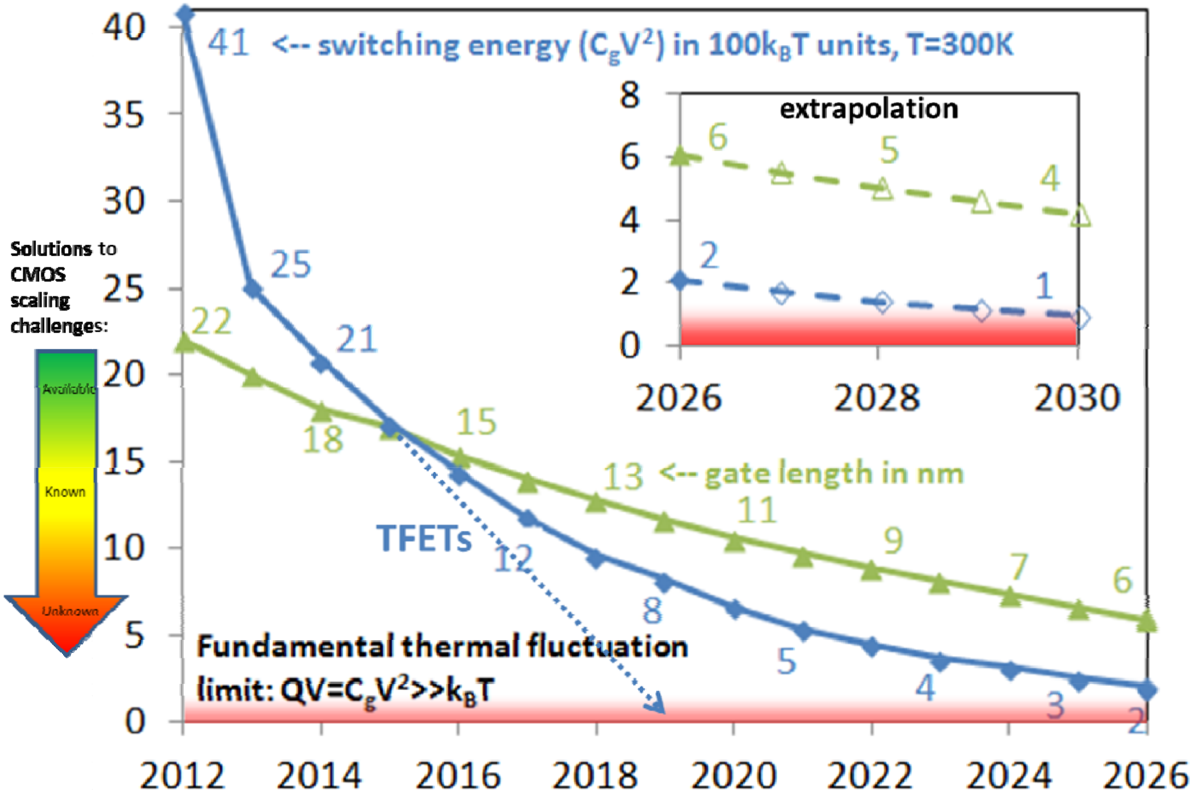
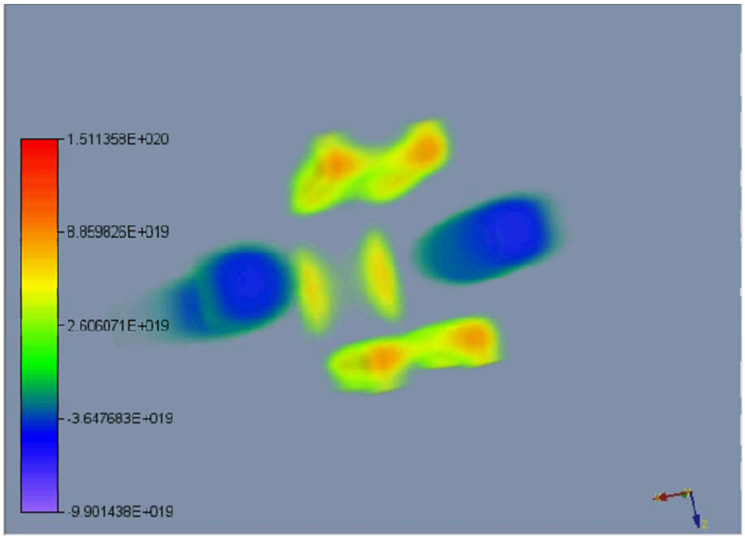


Figure 7. ITRS gate length projection (green) for high performance Si FinFET devices and the calculated switching energy per device projection (blue). Inset shows the extrapolated data.

The dashed curves and open symbols in the inset represent our extrapolation of ITRS data. This new way of plotting the switching energy leads to the interesting observation that, as gate length scales down, the switching energy becomes ever closer to the energy of thermal fluctuations. Hence, according to ITRS projections, scaling of the FET technology is likely capable of continuing for another 15 years (provided that the UV lithography, gate dielectric/work function engineering and other significant technological challenges can be addressed in one way or another). However, by the year 2030, scaling will reach a fundamental limit, when the switching energy becomes less than $100k_BT$, below which reliable FET-based logic operations may not be possible due to the thermal noise induced logic errors.

2.4 Switching energy obtained from CBR3D calculations

To investigate the reality of the projected data and the thermal fluctuation limit, we employed our fully 3D charge-self-consistent quantum transport simulator CBR3D to simulate the electrical performances of MuGFETs at gate lengths of 6-, 5- and 4-nm. All MuGFETs structures considered have their geometry and doping profiles been optimized in order to satisfy ITRS specifications (Fig. 6) for each node. We note that the CBR method allows us to calculate the source-drain and gate leakage currents rigorously (i.e., in the same full 3D quantum-mechanical way). The effective gate capacitance C_g is extracted using the quasi-static approximation: the corresponding capacitive (i.e. induced) charge distribution is given by $c(r) = q\Delta n(r)/\Delta V_g$, with $\Delta n = n(V_g=V_{dd}) - n(V_g=0)$. An example induced (capacitive) charge distribution is shown in Figure 8 for an optimized 6-nm Si FinFET.



$$\Delta Q(\mathbf{r}) = q\left[n_{on}(\mathbf{r}) - n_{off}(\mathbf{r})\right]$$

$\Delta Q(\mathbf{r})$ – induced charge distribution

$\Rightarrow$

$$C_g = \int_{\Delta Q(\mathbf{r})>0} \frac{\Delta Q(\mathbf{r})}{V} d\mathbf{r}, \quad \left[\int \Delta Q(\mathbf{r}) d\mathbf{r} = 0\right]$$

$\Rightarrow$

$$E_{switch} = C_g V^2$$

E_{switch} – (electrostatic) energy required to switch a transistor on/off

Figure 8. Induced (capacitive) charge calculation and distribution in a 6-nm Si FinFET.

We simulated and analyzed a number of MuGFET structures consisting of Si/Ge/GaAs channels, state-of-the-art HfSiON/SiO_2 gate dielectrics, and TaN metal gates, with a variety of doping profiles, and at several gate lengths including 6-, 5-, and 4-nanometers.

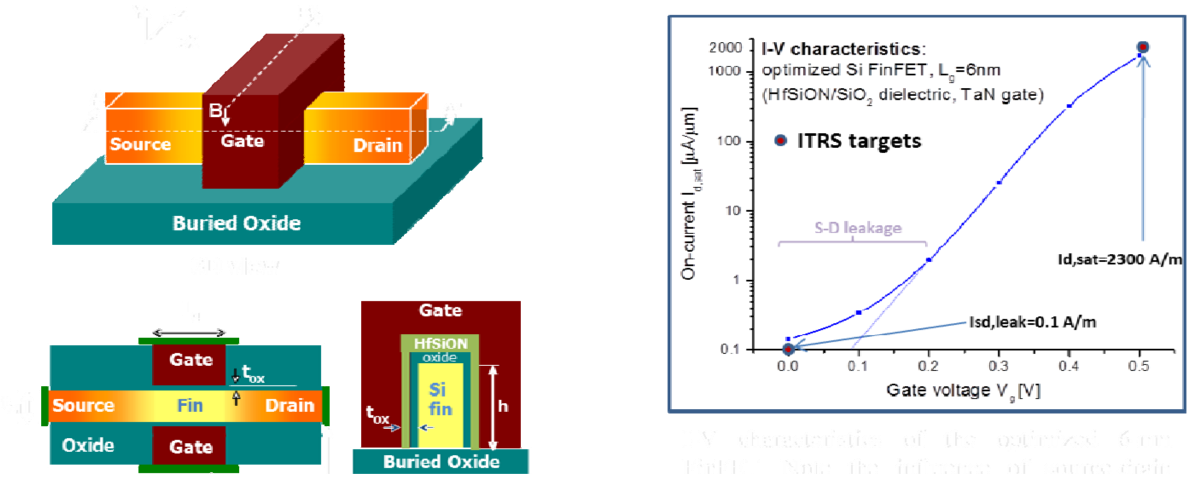


Figure 9. Geometry/doping profile optimization procedure and I-V characteristics of 6-nm FinFET (HfSiON/SiO2 dielectric, TaN metal gate)

We obtained optimized devices for each gate length (see Figure 9 for I-V characteristics of an optimized 6-nm FinFET device), and extracted the switching energies for these candidates to compare with (and confirm) the ITRS projected data, as shown in Figure 10.

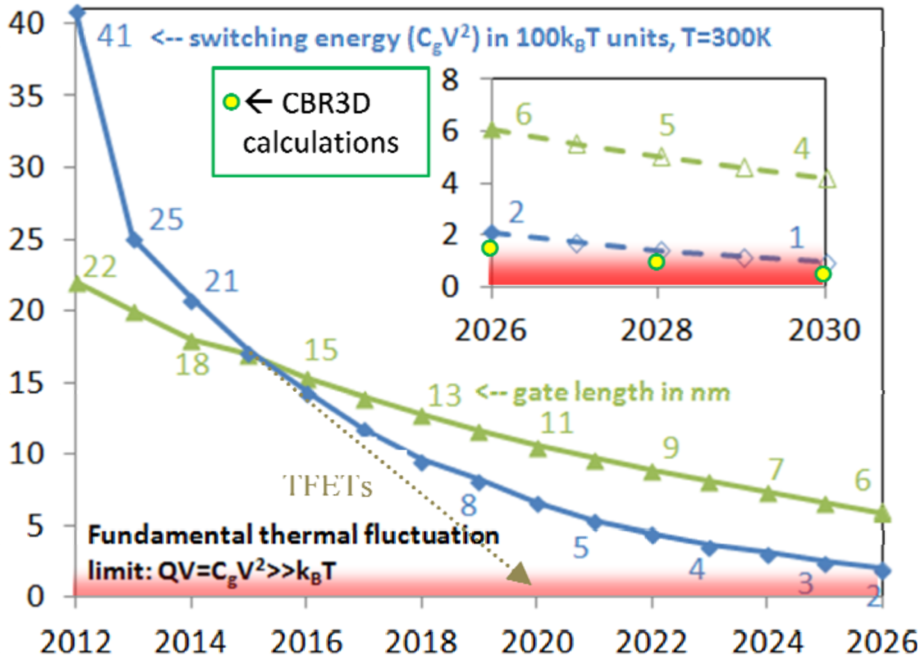


Figure 10. Switching energy vs gate length per CBR3D calculations (yellow circles).

As is clear from Fig. 10, MuGFETs will hit the thermal fluctuation limit at gate lengths of approximately 5nm. We have also determined that an alternative choice of channel materials (Ge, III-V, etc.) does not alter our finding that thermal fluctuations set the fundamental downscaling limit for FETs, since the gate capacitance is mainly determined by the node geometry and dielectric material. Specifically, we have considered the following materials with different crystallographic orientations (wafer)/[channel direction]:

- Si in 3 different orientations: (100)/[001], (110)/[001], (110)/[1-10];
- Ge in 2 different orientations: (100)/[001], (111)/[-211];
- GaAs (conduction band only, spherical effective mass: 0.067*m_e)

and found [39,52] that while alternative channel materials/orientations may significantly affect the electron mobility and current density, the gate capacitance varies by less than 10% comparing with the standard Si wafer/channel (100)/[001].

CONCLUSIONS

By developing a novel CBR3D simulator, have discovered that there is a fundamental, as opposed to a technological, downscaling limit that will prevent the room-temperature operation of all sub-5nm gate length FETs. This limit effectively indicates an end to downscaling and Moore's law for all FETs, including the most crucial technology – CMOS. According to the current ITRS projections [51], this will occur no later than in 15 years from now when gate lengths approach 5nm. Even more alarming, this downscaling limit could be reached even earlier, if CMOS technology is replaced by more power-efficient structures such as tunneling FETs. This is because TFETs have much steeper subthreshold slopes, and consequently lower operating voltages [48] - for example 0.1V could be used in a TFET instead of 0.5V for the corresponding CMOS node. In this situation, the gate capacitance of such TFETs would be the same or smaller than capacitance of the corresponding CMOS transistors. That means an additional factor of 25 reduction for the switching energy, which would bring TFET technology against the thermal fluctuation limit much sooner than CMOS.

We argue that this fundamental thermal fluctuation limit holds true for all charge-based FET technologies since they all operate on similar principles. Thus, it is necessary to rethink the question, *what is the actual "beyond-Moore" challenge?* The tremendous deflationary influence of Moore's law on the global economy is yet to be fully appreciated [50]. It is clear, however, that when the *density of transistors* stops increasing, the exponential decline of the price per function in computing is also going to stop. With these considerations in mind, we argue that **the actual beyond-CMOS challenge lies in *extending Moore's law beyond the 5-nm feature size down to sub-nm (few atoms) size***. In doing so, the essential problem is that downscaling below 4-5nm feature size cannot be realized using FETs (or TFETs) – irrespective of any high device channel mobility, steep threshold slopes that can be achieved, or whatever non-Si, alternative, 2D, *etc*. materials may be generally used.

As such, we foresee at least three industry possibilities after the thermal fluctuation limit is reached

(1) Accept the end of Moore's law and concentrate efforts on reducing power dissipation with the adiabatic or reversible computing;

(2) Use non-FET alternatives: memristors, super-conducting logic, etc.

(3) Continue Moore’s law using single-electron transistors!

The third option seems to be the most promising in our opinion. Indeed, there will be still a lot of “room” below 5nm-gate lengths (atomic size is about 0.2 nm) and any possibility to extend Moore’s law, which has been so enormously beneficial to world’s economy, should be closely investigated. Since $E_{switch} = q^2 / C_g$ for SETs, their switching energy trend vs. gate/island capacitance is opposite to that of all other FETs, which may allow their downscaling to sub-5nm gate lengths.

REFERENCES

[1] F. M. Wanlass and C. T. Sah, Digest of Technical Papers, ISSCC, pp. 32-33 (1963).

[2] H. Iwai, Proc. of the 17th Intl. Conference on VLSI Design (VLSID04), pp. 30-35 (2004).

[3] V.V. Zhirnov *et al.*, Proc. IEEE **91**, 1934 (2003).

[4] T. Skotnicki, J. A. Hutchby, T. J. King, H. S. Philip Wong, and F. Boeuf, IEEE Circ. Dev. Mag., pp. 16-26 (2005).

[5] N. Z. Haron and S. Hamdioui, Proc. of the 3rd Intl. Design and Test Workshop, pp. 98-103 (2008).

[6] ITRS Reports: 2011, 2012-editions for HP logic devices: http://www.itrs.net/Links/2011ITRS/Home2011.htm, Fig. PIDS2.

[7] D. Mamaluy et al., Phys. Rev. B 71, 245321 (2005).

[8] H. R. Khan, D. Mamaluy and D. Vasileska, IEEE T-ED 54, pp. 784-796 (2007).

[9] L. V. Keldysh, Sov. Phys. JETP 20, 1018 (1965).

[10] Y. M. Sabry et al., Int. J. Numerical Modeling 24, 322 (2011).

[11] Hao Wang et al., IEEE T-ED 56, 3106 (2009).

[12] E. O. Kane, Tunneling Phenomena in Solids, edited by E. Burstein and S. Lundqvist (Plenum, New York, 1969).

[13] J. N. Schulman and Y. C. Chang, Phys. Rev. B 27, 2346-2354 (1983).

[14] C. Mailhiot and D. L. Smith, Phys. Rev. B 33, 8360-8372 (1986).

[15] W. Frensley, Rev. Mod. Phys. 62, 745-791 (1990).

[16] C. Lent, D. Kirkner, J. Appl. Phys. 67, 6353-6359 (1990).

[17] Z.-Y. Ting et al., Phys. Rev. B 45, 3583-3592 (1992).

[18] C. Strahberger and P.Vogl, Phys. Rev. B 62, 7289-7297 (2000).

[19] Y. X. Liu et al., Phys. Rev. B 54, 5675-5683 (1996).

[20] E. Polizzi et al., J. Appl. Phys. 87, 8700-8706 (2000).

[21] E. Polizzi, N. Ben Abdallah, Phys. Rev. B 66, 245301-1- 245301-9 (2002).

[22] P. A. Ramachandran, Boundary Element Methods in Transport Phenomena, p. 1 (WIT Press, 1993).

[23] H. Frohne, M. McLennan, and S. Datta, J. Appl. Phys. 66, 2699-2706 (1989).

[24] P. A. Knipp and T. L. Reinecke, Phys. Rev. B 54, 1880-1891 (1996).

[25] D. K. Ferry and S. M. Goodnick, Transport in Nanostructures (Cambridge University Press, Cambridge, 1997).

[26] R. Lake et al., J. Appl. Phys. 81, 7845-7869 (1997).

[27] A. Svizhenko et al., J. Appl. Phys 91, 2343-2354 (2002).

[28] R. Venugopal et al., J. Appl. Phys. 92, 3730-3739 (2002).

[29] C. Rivas and R. Lake, Phys. Stat. Sol. (b) 239, 94-102 (2003).

[30] S. Rotter, Phys. Rev. B 62, 1950-1960 (2000).

[31] S. Rotter, Phys. Rev. B 68, 165302-1-165302-14 (2003).

[32] A. Rahman, A. Ghosh, and M. Lundstrom, Tech. Dig. - Int. Electron Devices Meet., 471-474 (2003).

[33] A. Rahman, M. S. Lundstrom, and A. W. Ghosh, J. Appl. Phys. 97, 053702-1-053702-12 (2005).

[34] J. Wang, E. Polizzi, and M. Lundstrom, J. Appl. Phys. 96, 2192-2203 (2004).

[35] H. Takeda and N. Mori, J. Comp. El. 4, 31-34 (2005).

[36] S. Jin, Y. J. Park, and H. S. Min, J. Appl. Phys. 99, 123719-1-12719-10 (2006).
[37] S. E. Laux, A. Kumar, and M. V. Fischetti, J. Appl. Phys. 95, 5545-5582 (2004).
[38] https://engineering.purdue.edu/gekcogrp/software-projects/nemo5/nemo5_manual.pdf
[39] X. Gao, D. Mamaluy, E. Nielsen, R. W. Young, A. Shirkhorshidian, M. P. Lilly, N. C. Bishop, M. S. Carroll and R. P. Muller, J. Appl. Phys. 115, 133707 (2014).
[40] http://www.caam.rice.edu/software/ARPACK/
[41] http://www.ecs.umass.edu/~polizzi/feast/
[42] A. Trellakis, A. T. Galick , A. Pacelli and U. Ravaioli, J. Appl. Phys. 81, 7880 (1997).
[43] D. Mamaluy, M. Sabathil, P. Vogl, J. Appl. Phys. 93, 4628 (2003).
[44] A. Di Carlo, P. Vogl, and W. Pötz, Phys. Rev. B 50, 8358 (1994).
[45] Rolf Landauer, IBM Journal of Research and Development, vol. 5, pp. 183–191 (1961).
[46] V.V. Zhirnov et al., Proc. IEEE 91, 1934 (2003).
[47] M.P. Frank, Computing Science & Engineering 4, 16-26 (2002).
[48] Erik DeBenedictis, et al. "What's Beyond Moore's Law", SAND2009-2325.
[49] D. Mamaluy, X. Gao, B. Tierney, 10.1109/IWCE.2014.6865875
[50] http://www.itrs.net/Links/2010ITRS/IRC-ITRS-MtM-v2%203.pdf
[51] http://www.itrs.net/Links/2013ITRS/Summary2013.htm
[52] D. Mamaluy, X. Gao, B. Tierney, "How much time does FET downscaling have left?", Appl. Phys. Letters (in prep.)

DISTRIBUTION

4 Lawrence Livermore National Laboratory
Attn: N. Dunipace (1)
P.O. Box 808, MS L-795
Livermore, CA 94551-0808

1	MS0899	Technical Library	9536 (electronic copy)
1	MS0359	D. Chavez, LDRD Office	1911

Sandia National Laboratories