Applied Physics Letters

The fundamental downscaling limit of field effect transistors

Cite as: Appl. Phys. Lett. **106**, 193503 (2015); https://doi.org/10.1063/1.4919871
Submitted: 11 February 2015 • Accepted: 25 April 2015 • Published Online: 12 May 2015

Denis Mamaluy and Xujiao Gao

openaccess

View Online · Export Citation · CrossMark

ARTICLES YOU MAY BE INTERESTED IN

MOSFET Scaling Trends, Challenges, and Key Associated Metrology Issues Through the End of the Roadmap
AIP Conference Proceedings **788**, 203 (2005); https://doi.org/10.1063/1.2062964

A critical review of recent progress on negative capacitance field-effect transistors
Applied Physics Letters **114**, 090401 (2019); https://doi.org/10.1063/1.5092684

High-κ gate dielectrics: Current status and materials properties considerations
Journal of Applied Physics **89**, 5243 (2001); https://doi.org/10.1063/1.1361065

AIP Publishing

Appl. Phys. Lett. **106**, 193503 (2015); https://doi.org/10.1063/1.4919871 **106**, 193503

© 2015 Author(s).

The fundamental downscaling limit of field effect transistors

Denis Mamaluy[a)] and Xujiao Gao
Sandia National Laboratories, Albuquerque, New Mexico 87185-1322, USA

(Received 11 February 2015; accepted 25 April 2015; published online 12 May 2015)

We predict that within next 15 years a fundamental down-scaling limit for CMOS technology and other Field-Effect Transistors (FETs) will be reached. Specifically, we show 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. These findings were confirmed by performing quantum mechanical transport simulations for a variety of 6-, 5-, and 4-nm gate length Si devices, optimized to satisfy high-performance logic specifications by ITRS. Different channel materials and wafer/channel orientations have also been studied; it is found that altering channel-source-drain materials achieves only insignificant increase in switching energy, which overall cannot sufficiently delay the approaching downscaling limit. Alternative possibilities are discussed to continue the increase of logic element densities for room temperature operation below the said limit. *© 2015 Author(s). All article content, except where otherwise noted, is licensed under a Creative Commons Attribution 3.0 Unported License*. [http://dx.doi.org/10.1063/1.4919871]

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, 1980s, and 1990s, as summarized in Ref. 2. Since then, there have been many studies[2–4] 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 field effect transistor (FET) based. Arguments are frequently made that III-V-, carbon-nanotube-, 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 CMOS for high performance computing applications. In fact, ITRS now projects[5] that the emerging trend of Si Multi-Gate FET (MuGFET) technology should allow Moore's law to continue for at least another decade until 6-nm gate length is reached (see Fig. 1, black diamonds).

The question of how small the CMOS/FET devices can become remains an active subject of research and debate. In 2003, Zhirnov *et al.*[6] estimated that the minimal feature size of a "binary logic switch" is given by $x_{min} = \hbar/\sqrt{2m_e E_s} = \hbar/\sqrt{2m_e k_B T \ln 2} \approx 1.5$ nm at T = 300 K. This estimation is based on the Heisenberg uncertainty ($\Delta x \Delta p \geq \hbar$) and the Landauer principle[7] which states that the switching energy, E_s, of a binary switch must be higher than $k_B T \ln 2$ for irreversible computing. It is obvious, however, that the estimate is only applicable to the Landauer switching energy limit. On the other hand, it has been shown[8] that modern CMOS architectures cannot operate at such low switching energies due to prohibitively high expenses associated with the necessity to compensate for thermally induced errors. In fact, the minimal switching energy of a realistic FET transistor that guarantees error-free lifetime circuit operation is on the order of $E_s = 100 k_B T$.[8–10] Thus, for realistic transistors that operate sufficiently far from the Landauer limit, Zhirnov's estimate is not relevant, since $x_{min} = \hbar/\sqrt{2m_e k_B T 100} \approx 0.12$ nm at 300 K, which is on the order of atomic size.

We have initiated this study by utilizing the recent ITRS projections[5] for CMOS technology downscaling and characteristics to compute the device switching energy, $E_s = C_g V_g^2$, where C_g is the gate capacitance and V_g is the gate voltage needed to turn on a FET device. We note that this concept of switching energy applies to all FETs, including MOSFETs,

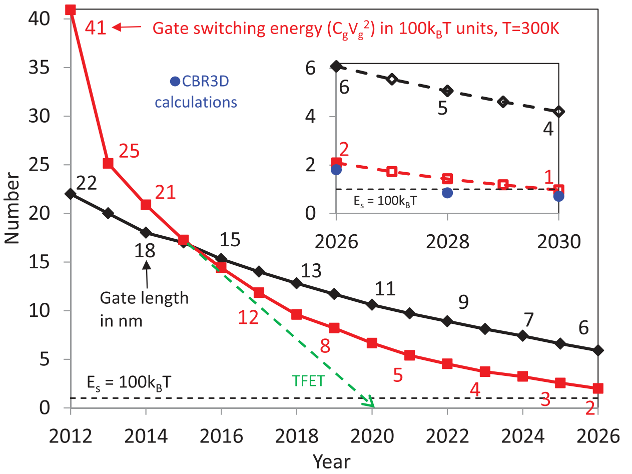

FIG. 1. ITRS gate length projection (black diamonds) for high performance Si MuGFET devices and the calculated switching energy per device projection (red squares). Solid black and red curves are guides for the eyes. The black dashed line indicates the switching energy of $100k_BT$. The green dashed line represents the approximate projection of switching energy for tunneling FETs. Inset: open symbols represent our extrapolation of ITRS data, while filled blue circles are our CBR3D simulated switching energies at 6-, 5-, and 4-nm nodes.

[a)]Electronic mail: mamaluy@sandia.gov

© Author(s) 2015

MuGFETs, TFETs, SpinFETs, and SETs, but is not applicable to non-FET devices, such as memristors. 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 have been obtained,[11] shown as red curve in Fig. 1. The downscaling projection of the switching energy leads to an interesting observation that, as the gate length scales down to sub-10 nm values, the switching energy rapidly approaches the $100k_BT$ value, making the device susceptible to 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 could be addressed in one way or another. However, by the year 2030, downscaling will reach a fundamental limit, when the switching energy becomes less than $100k_BT$, below which reliable FET-based logic operations would not be possible due to thermal fluctuations and the consequent logic errors. In this analysis, we do not consider possibilities for hardware or software error correction that could somewhat soften this limit; we note, however, that due to the exponential increase of the thermally induced error rates with reducing switching energy,[7] such error correction would become impractical for sub-5 nm gate lengths. Though the projection data are estimated for Si FETs, in the following we will demonstrate that this fundamental thermal fluctuation limit also holds true for FETs with alternative (e.g., Ge, III-V) channel (and/or source-drain) materials. Moreover, the same analysis remains valid for other FET-based technologies, such as TFETs. Indeed, from basic geometry considerations, the gate capacitance of a TFET is the same or smaller than that of a CMOS transistor of the corresponding size; while the operating gate voltage of the TFET could be much lower than that for CMOS, due to the much steeper turn-on characteristics of TFETs. It is therefore easy to see that this power-saving advantage of TFETs could become detrimental at smaller nodes, as illustrated by the dashed green line in Fig. 1.

To investigate the validity of the projected switching energy in relation to the thermal fluctuation limit, we employed our fully 3D charge-self-consistent quantum transport simulator, CBR3D, to simulate and optimize the electrical performances of MuGFETs at gate lengths of 6-, 5-, and 4-nm. The CBR3D simulator is based on a numerical method called Contact Block Reduction (CBR),[12,13] which provides an efficient implementation of the Keldysh Non-Equilibrium Greens Function (NEGF) formalism[14] for open-system quantum transport. The CBR quantum transport is self-consistently coupled with the Poisson equation in the CBR3D simulator to satisfy the charge self-consistency. The self-consistent convergence is achieved by adopting the predictor-corrector algorithm[15] to open systems.[13,16] Surface and interface roughness are included with the real-space treatment,[17] inelastic scattering processes are emulated with an analog of relaxation time approximation or "Buttiker probes."[18] We note, however, that in this study the emulation of inelastic scattering only affected the on-current values (about 10% reduction compared with the case of elastic scattering only) and practically did not affect the capacitance and the switching energy values. The CBR3D simulator shows a linear scaling with the number of grid points (i.e., problem size) and a nearly linear speed-up with the number of CPUs as shown in Fig. 2. This linear scaling allowed us to simulate a large number of MuGFET devices with different geometry parameters and doping profiles to perform device optimization at different gate lengths.

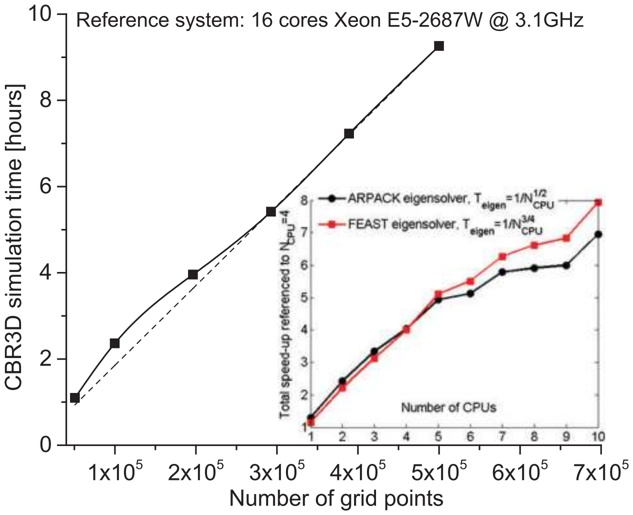

FIG. 2. CBR3D total simulation time (per bias point) scaling with the number of grid points. Inset: CBR3D relative speed-up with the number of CPUs used; red curve corresponds to the use of FEAST[19] eigensolver, black curve—ARPACK[20] eigensolver.

We first simulated and analyzed a number of MuGFET structures consisting of Si(100)/[001] channels, state-of-the-art HfSiON/SiO_2 gate dielectrics, and TaN metal gates, at gate lengths of 6-, 5-, and 4-nm. Figure 3 shows the schematics of a representative MuGFET structure that was simulated. We obtained an optimized device at each gate length by varying the geometry dimensions (e.g., fin width and fin height) and doping profiles (e.g., step versus Gaussian doping shape at the source/channel junction). Figure 4 shows the

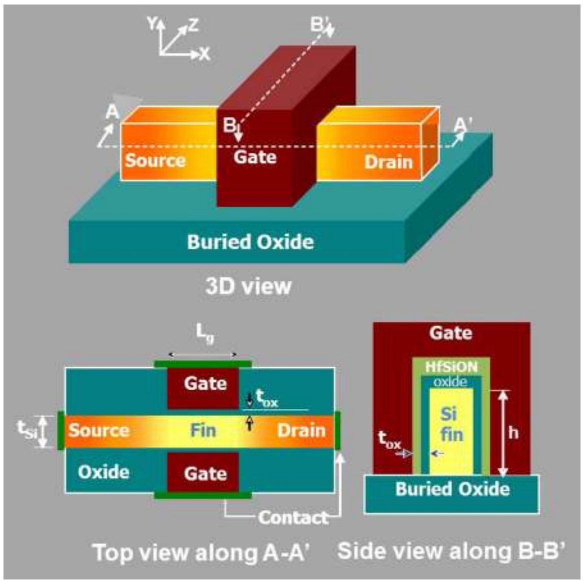

FIG. 3. Schematics of a MuGFET/FinFET structure. Top panel: 3D view, bottom-left panel: top view along the A-A′ cross section, bottom-right panel: side view along the B-B′ cross section.

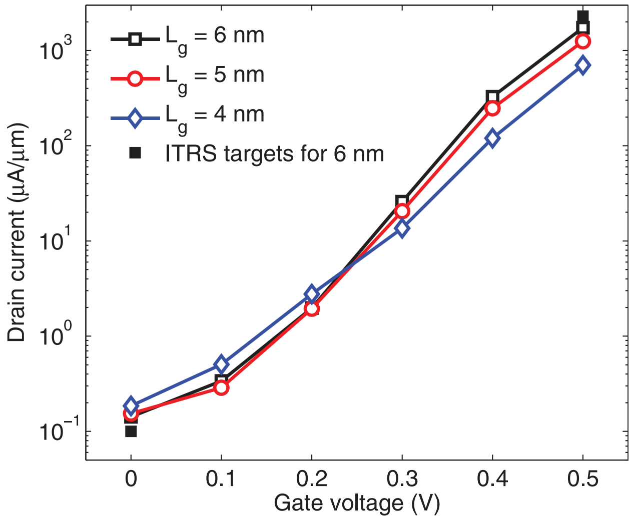

FIG. 4. Drain current versus gate voltage characteristics obtained from our CBR3D simulator for optimized 6-, 5-, and 4-nm Si MuGFETs. Filled black squares indicate ITRS projection targets for the 6-nm MuGFET node.

drain current versus gate voltage characteristics obtained from our CBR3D simulator for optimized 6-, 5-, and 4-nm Si MuGFETs. For the 6-nm node, the device was optimized to closely match the ITRS specifications, including off- and on-current values.[5] For the 5- and 4-nm nodes, we optimized the devices such that their drain off- and on-currents are close to the values obtained by extrapolating the ITRS specifications[5] to smaller gate lengths. The appreciable leakage current for $V_g < 0.2$ V is due to source-drain band-to-band tunneling, which becomes more dominant at the 4-nm node due to a shorter gate length.

Once an optimized device geometry and doping profile were determined for a given gate length, we extracted the effective gate capacitance C_g using the quasi-static approximation: the induced charge distribution $\Delta Q(\mathbf{r})$ has been calculated as $\Delta Q(\mathbf{r}) = q[n_{on}(\mathbf{r}) - n_{off}(\mathbf{r})]$, where $n_{on}(\mathbf{r})$ and $n_{off}(\mathbf{r})$ are the electron density profiles when the device is in the on- and off-states, respectively. The induced charges in the source-channel-drain region and the gate region are equal in magnitude and opposite in signs, so that the integration

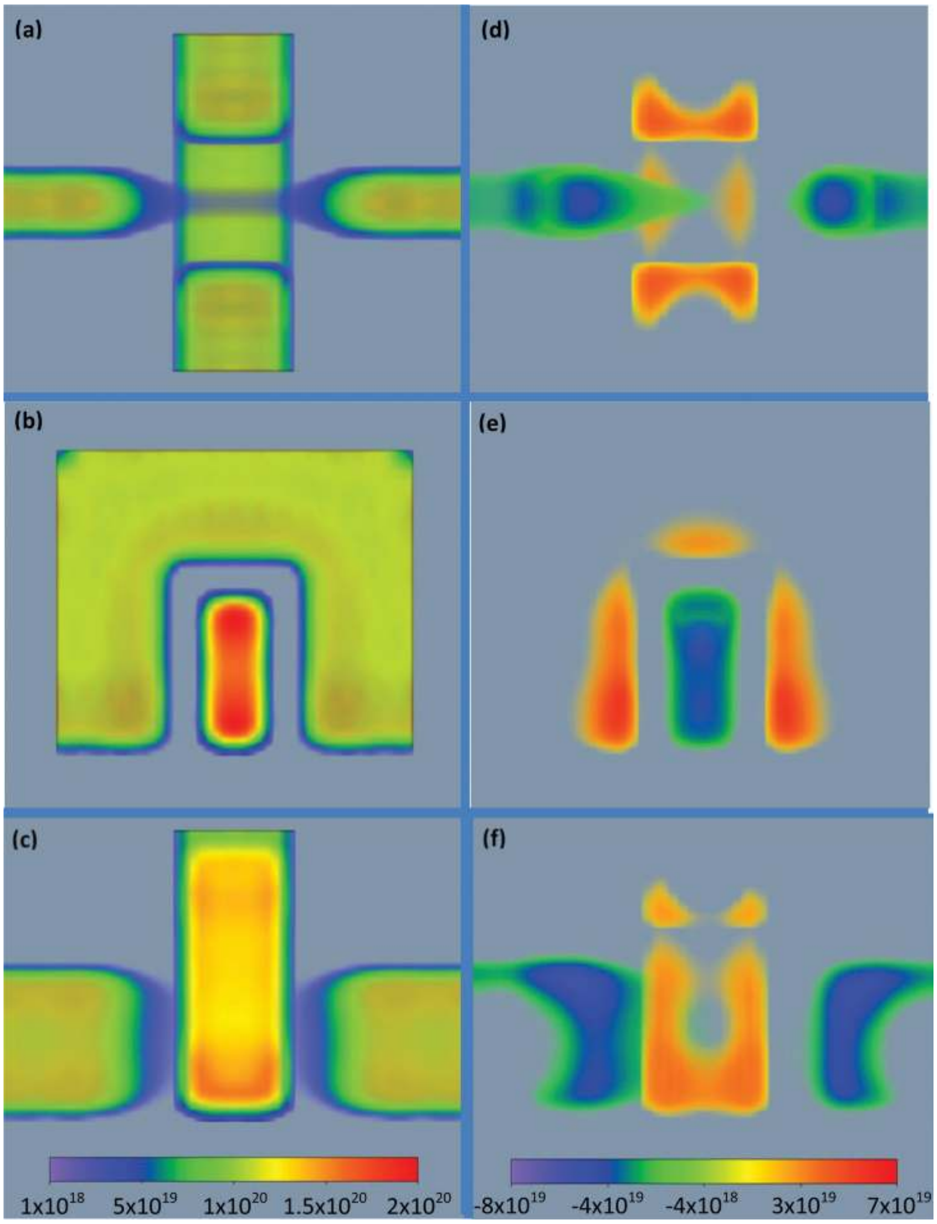

FIG. 5. Projected views of the on-state ($V_g = 0.5$ V) electron density (left column) and the corresponding induced charge distribution (right column) for the optimized 6-nm MuGFET device. (a) and (d) Bottom views, (b) and (e) side views, (c) and (f) front views.

over the entire device volume is zero to satisfy the total charge neutrality condition in the device. The capacitance is computed as $C_g = \int_{\Delta Q(\mathbf{r})>0} \Delta Q(\mathbf{r}) d\mathbf{r}/V_g$, from which the switching energy E_s is calculated as $E_s = C_g V_g^2$.

Figure 5 shows three orthogonal projections of the on-state ($V_g = 0.5$ V) electron density (left column) and the corresponding induced charge distribution (right column) for the optimized 6-nm MuGFET device. Similar electron density and induced charge profiles were also obtained for the optimized 5- and 4-nm devices. From the electron density in the left column, two important effects can be observed: (i) the electronic channel is located in the center of the intrinsic silicon region, instead of in the surface region close to the gate, due to the full volume inversion achievable at sub-10 nm gate lengths;[21] (ii) the electron densities in the source, channel, drain, and *gate* regions are all set back from the surfaces, due to quantum confinement. The induced charge distribution in (d) shows that its maximum density is located not in the channel, but near the source/drain-channel junction regions (blue color), implying that at 6-nm gate lengths, the gate capacitance is dominated by fringing effects. This is different from an optimized 10-nm FinFET device in Ref. 21, where the induced charge density is still peaked in the channel, as expected for MOSFET and larger MuGFET devices. Another important feature of the capacitive charge distribution is that the induced positive charge in the Π-shaped gate region has a complex spatial distribution, which exhibits highest densities (red color) near the surfaces that are close to the source, drain, and channel, because of stronger interactions of the gate with these regions. Capture of this complex interactions in CBR3D simulator was made possible, because electron transport in the TaN gate was modeled using the same quantum mechanical approach as the body, assuming TaN as a highly doped semiconductor with an electron effective mass equal to that of free electron and an effective "doping level" determined by fitting simulations to the tunneling current measurements in HfSiON/TaN systems.[22] We note that these interactions between the gate and the source/drain/channel regions would be lost, if one used a standard gate treatment (e.g., Ref. 23), which neglects electron transport and quantum confinement effects and assumes equipotential boundary condition in the gate.

The switching energies have been extracted for the optimized MuGFET devices at the 6-, 5-, and 4-nm gate lengths, and are plotted as filled blue circles in the inset of Fig. 1. The switching energies obtained from the CBR3D simulator are about 10% smaller than the values calculated using ITRS projection data, likely because CBR3D captures the effects of quantum confinement which effectively reduces the gate capacitance. Our CBR3D quantum transport simulation results clearly indicate that the switching energy of an optimized MuGFET at the 5-nm node crosses the threshold of $100k_BT$ and it becomes even smaller at the 4-nm node. This confirms our initial observation, based on ITRS data, that Si MuGFET devices would reach a fundamental downscaling limit around 5 nm, below which the switching energy required to turn a FET device on/off becomes sufficiently close to the energy of thermal fluctuations, preventing the device from performing suitably reliable logic operations.

To investigate how the switching energy downscaling limit may be affected by the channel material, crystallographic wafer/channel orientations, and gate dielectric, we performed CBR3D simulations on a group of other MuGFET devices, which source/drain/channel regions were made of Si(110)/[001], Si(110)/$[1\bar{1}0]$, Ge(100)/[001], Ge(111)/$[\bar{2}11]$, and GaAs, respectively, using gate dielectric of HfSiON with the dielectric constant of 14.0 from Ref. 22 and a yet unknown material with the dielectric constant of 20.0 assumed in the most recent edition of ITRS report.[24] The switching energies were extracted for all these devices and are plotted in Fig. 6 in unit of $100k_BT$. At the 6-nm gate length, the switching energy for MuGFETs using Si channel is still sufficiently above the $100k_BT$ switching threshold and has little dependence on the crystallographic orientation. As the gate length approaches 5 and 4 nm, the switching energy becomes less than the $100k_BT$ threshold, and still shows insignificant dependence on the crystallographic orientation. On the other hand, the channel material and gate dielectric show discernible effect on the switching energy. As seen from Fig. 6, using Ge channel leads to higher switching energy than Si channel, because Ge has a higher dielectric constant of 16.0 compared to 12.0 of Si, which results in—nearly proportionally—higher gate capacitance. At the same time, the switching energy of a GaAs MuGFET, is much smaller than that of Si MuGFET. This effect is due to much lower electron density of states in ultra-scaled, thin channel devices with low effective mass: the corresponding number of electrons in the channel in on-state is reduced compared with the Si channel, thus decreasing the gate capacitance.

In conclusion, we outline three possibilities for the industry after the thermal fluctuation limit is reached and the density of FETs will be impossible or impractical to increase for room temperature operation: (A) accept the end of Moore's law and concentrate efforts on reducing power dissipation with the adiabatic or reversible computing; (B) use

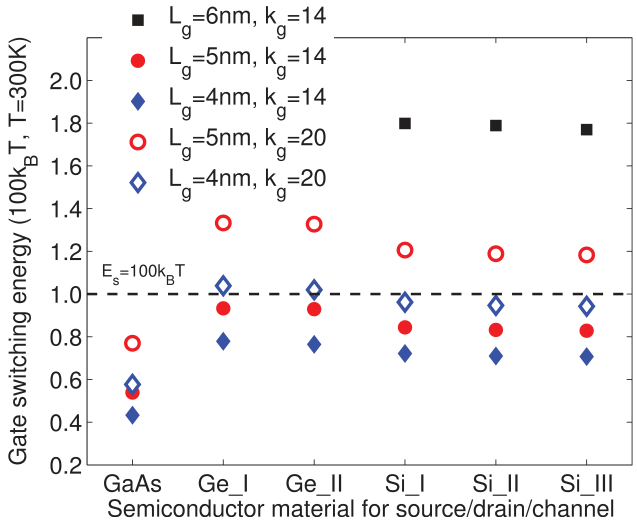

FIG. 6. Gate switching energy in unit of $100k_BT$ (T = 300 K) for MuGFETs using different source/drain/channel materials, crystallographic orientations, and gate dielectrics, at the 6-, 5-, and 4-nm gate lengths. L_g and k_g in the legend represent gate length and gate dielectric constant, respectively. The black dashed line indicates the $100k_BT$ switching energy threshold. Ge_I, Ge_II, Si_I, Si_II, and Si_III represent Ge(100)/[001], Ge(111)/$[\bar{2}11]$, Si(100)/[001], Si(110)/[001], and Si(110)/$[1\bar{1}0]$, respectively.

non-FET alternatives: memristors, super-conducting logic, etc.; (C) continue Moore's law using single-electron transistors (see, e.g., Ref. 25 and references therein): their switching energy trend vs. gate/island capacitance is opposite to that of all other FETs, which may allow their downscaling to sub-5 nm gate lengths.

This work was supported by the Laboratory Directed Research and Development program at Sandia National Laboratories. Sandia is a multiprogram laboratory operated by Sandia Corporation, a Lockheed Martin Company, for the United States Department of Energy's National Nuclear Security Administration under Contract No. DE-AC04-94AL85000.

[1] F. M. Wanlass and C. T. Sah, in *Digest of Technical Papers, 1963 IEEE International Solid-State Circuits Conference* (*ISSCC*), pp. 32–33 (1963).
[2] H. Iwai, in *Proceedings of the 17th International Conference on VLSI Design* (*VLSID04*) (2004), pp. 30–35.
[3] T. Skotnicki, J. A. Hutchby, T. J. King, H. S. Philip Wong, and F. Boeuf, IEEE Circuits Devices Mag. **21**, 16–26 (2005).
[4] N. Z. Haron and S. Hamdioui, in *Proceedings of the 3rd International Design and Test Workshop* (2008), pp. 98–103.
[5] ITRS Reports: 2011, 2012-editions for HP logic devices, see http://www.itrs.net/Links/2012ITRS/Home2012.htm, Table PIDS2.
[6] V. V. Zhirnov, R. K. Cavin III, J. A. Hutchby, and G. I. Bourianoff, Proc. IEEE **91**, 1934–1939 (2003).
[7] R. Landauer, IBM J. Res. Dev. **5**, 183–191 (1961).
[8] M. P. Frank, Comput. Sci. Eng. **4**, 16–26 (2002).
[9] E. DeBenedictis, P. E. Dodd, A. L. Lentine, and K. Kee Ma, "What's beyond Moore's law," Sandia Technical Report SAND2009-2325, April 2009.
[10] G. L. Snider, E. P. Blair, C. C. Thorpe, B. T. Appleton, G. P. Boechler, A. O. Orlov, and C. S. Lent, in *12th IEEE Conference on Nanotechnology IEEE-NANO* (2012), pp. 160–165.
[11] D. Mamaluy, X. Gao, and B. Tierney, Proc. IWCE **2014**, 220–221.
[12] D. Mamaluy, D. Vasileska, M. Sabathil, T. Zibold, and P. Vogl, Phys. Rev. B **71**, 245321 (2005).
[13] H. R. Khan, D. Mamaluy, and D. Vasileska, IEEE Trans. Electron Devices **54**, 784 (2007).
[14] L. V. Keldysh, Sov. Phys. J. Exp. Theor. Phys. **20**, 1018 (1965).
[15] A. Trellakis, A. T. Galick, A. Pacelli, and U. Ravaioli, J. Appl. Phys. **81**, 7880 (1997).
[16] 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).
[17] H. Khan, D. Mamaluy, and D. Vasileska, J. Vac. Sci. Technol., B **25**, 1437 (2007).
[18] R. Venugopal, M. Paulsson, S. Goasguen, S. Datta, and M. S. Lundstrom, J. Appl. Phys. **93**, 5613 (2003).
[19] See http://www.ecs.umass.edu/∼polizzi/feast/ for FEAST algorithm and eigensolver.
[20] See http://www.caam.rice.edu/software/ARPACK/ for ARPACK algorithm and eigensolver.
[21] H. Khan, D. Mamaluy, and D. Vasileska, IEEE Trans. Electron Devices **55**, 743 (2008).
[22] X. Gao-Bo and X. Qiu-Xia, Chin. Phys. B **18**, 768 (2009).
[23] G. Klimeck and M. Luisier, Comput. Sci. Eng. **12**, 28 (2010); S. Steiger, M. Povolotskyi, H. H. Park, T. Kubis, and G. Klimeck, IEEE Trans. Electron Devices **10**, 1464 (2011).
[24] ITRS Report: 2013-edition (released April 2014), see http://www.itrs.net/Links/2013ITRS/Home2013.htm.
[25] Y. Sun, Rusli, and N. Singh, IEEE Trans. Nanotechnol. **10**, 96 (2011).