How much time does FET scaling have left?

D. Mamaluy, X. Gao, B. Tierney
Sandia National Laboratories, Albuquerque, New Mexico, USA
E-mail: mamaluy@sandia.gov

The ultimate end of CMOS scaling was predicted almost immediately after the now ubiquitous technology was invented by Frank Wanless 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 (See Fig. 1, green curve).

In this work, we compute the device switching energy, $C_g V_g^2$, for several representative FinFET/MuGFET devices, and explore the role of this quantity as a fundamental physical scaling limitation, which we predict will occur around 2030. In doing so, ITRS downscaling projection data [6] is utilized for reference. MuGFET switching energies are plotted as the blue curve in Fig. 1, in units of $100k_BT$ (T=300K), as FET gate lengths are scaled to 6-nm and below. The inset of Fig. 1 represents our extrapolation of ITRS data. This new way of plotting switching energy reveals that as gate lengths are scaled below about 5nm, the switching energy approaches that of thermal fluctuations.

Scaling of FET technology is thus likely to continue for at least another 15 years, provided that UV lithography, gate dielectric/work function engineering and other technological challenges can be addressed. Below 5-nm gate length however, reliable FET-based logic operations will not be possible due to the thermal noise induced logic errors. Though the ITRS projection data is estimated for Si FETs, 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-CMOS" challenge?*

The tremendous deflationary influence of Moore's law on the global economy is yet to be fully appreciated [7], 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.* 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.

To investigate the reality of the ITRS projected data within the constraint of the thermal fluctuation limit, we employ a fully 3D charge-self-consistent quantum transport simulator, CBR3D, based on the Contact Block Reduction (CBR) method [8,9] of simulating the electrical performance of Si FinFETs at gate lengths of a few nanometers. The CBR method

ISBN 978-2-9547858-0-6

allows us to calculate the source-drain and gate leakage currents utilizing a 3D quantum-mechanical treatment. The effective gate capacitance C_g is extracted using the quasi-static approximation: i.e., the corresponding capacitive (i.e. induced) charge distribution is given by $c(r) = q\Delta n(r)/\Delta V_g$, with $\Delta n = n(V_g{=}V_{dd}) - n(V_g{=}0)$, as shown in Fig. 2 for an optimized 6-nm Si FinFET. Recent CBR3D development allows for nearly linear speed-up with respect to the number of CPU cores, as shown in Fig. 3 (it is also clear from Fig. 3 that the FEAST eigensolver is faster than the ARPACK eigensolver). This linear speed-up has allowed us to simulate a very large number of FinFET devices with different gate lengths and doping profiles to perform device optimization. Several FinFET structures with Si channels, state-of-the-art gate dielectrics, and TaN metal gates, for a variety of doping profiles and gate lengths, including 6-, 5-, and 4-nm, are simulated. We obtain optimized devices at each gate length, and extract the switching energies for these candidates to compare with and confirm the ITRS projected data. It is shown that an alternative choice of channel material does not alter our finding that thermal fluctuations set the fundamental downscaling limit for FETs.

In conclusion, we discuss some theoretical possibilities for *overcoming* the thermal fluctuation limit and the consequent downscaling of charge-based logic devices to sub-5nm dimensions at room temperatures.

This work was supported by the Laboratory Directed Research and Development program of Sandia National Laboratories. Sandia is a multi-program laboratory operated by Sandia Corp., a wholly owned subsidiary of Lockheed Martin Corp., for the United States Department of Energy, National Nuclear Security Administration, under Contract DE-AC04-94AL85000.

[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: 20112012-editions for HP logic devices: http://www.itrs.net/Links/2011ITRS/Home2011.htm, Fig. PIDS2.

[7] http://www.itrs.net/Links/2010ITRS/IRC-ITRS-MtM-v2%203.pdf

[8] D. Mamaluy, M. Sabathil, T. Zibold, D. Vasileska, P. Vogl, Phys. Rev. B **71**, 245321 (2005).

[9] H. Khan, D. Mamaluy, D. Vasileska, IEEE Tran. El. Dev. **54**, pp. 784-796 (2007).

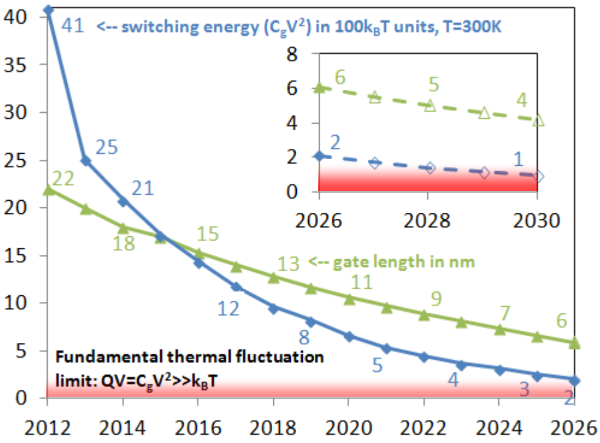

Fig.1. ITRS gate length projection (green) for high performance Si FinFET devices and associated calculated switching energy (blue). Inset shows extrapolated data.

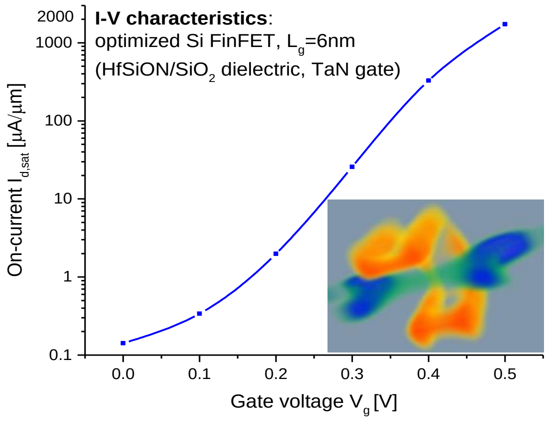

Fig.2. IV characteristics (note the influence of s-d tunneling at low voltages) and the induced capacitive charge distribution in an optimized 6-nm Si FinFET (red color represents positive charge; blue color - negative charge).

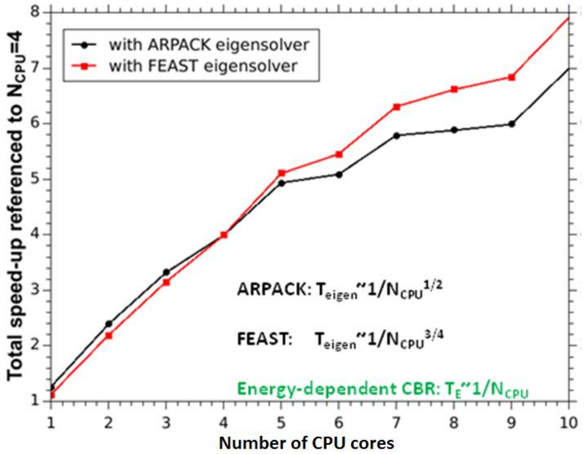

Fig.3. CBR3D speed-up scaling as a function of the number of CPU cores for two different eigensolvers.