

Impact of FinFET on Near-Threshold Voltage Scalability

**Nathaniel Pinckney, Supreet Jeloka,
Ron Dreslinski, Trevor Mudge,
Dennis Sylvester, and David Blaauw**
University of Michigan

Lucian Shifren, Brian Cline, and Saurabh Sinha
ARM Inc.

Editor's note:

Near-threshold operations provide a powerful knob for improving energy efficiency and alleviating on-chip power densities. This article explores the impact of newest FinFET CMOS technologies (from 40 to 7 nm) on near-threshold computing in terms of performance and energy efficiency.

– Muhammad Shafique, Vienna University of Technology

A previous analysis [7] of the near-threshold (NT) region across planar nodes (180–32 nm) showed NTC energy improvement is becoming less effective with each generation, with only a $4\times$ energy gain in

■ **THE DEMISE OF DENNARD** [1] scaling in recent years has led to increasing power densities with each generation of technology. The ability to cool, or extract waste energy from, a processor has remained relatively constant, and subsequently the ability to activate all portions of a chip multiprocessor (CMP) simultaneously is becoming successively limited. Consequently, at any given time, large regions of a chip will remain inactive in order to not exceed thermal design budgets of the package and cooling system, dubbed dark silicon [2]. To help overcome dark silicon, there has been proposals to aggressively voltage scale and operate at near-threshold computing (NTC) supply voltages [3]–[5], thus improving energy efficiency at the cost of moderate performance loss. Slower clock frequencies can be balanced through parallelizing a workload across additional cores, transforming dark silicon to dim silicon [6] by trading high single-core performance for energy-efficient many-core operation.

Digital Object Identifier 10.1109/MDAT.2016.2630303

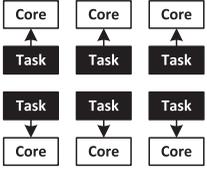
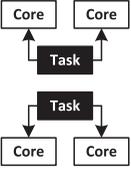
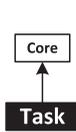
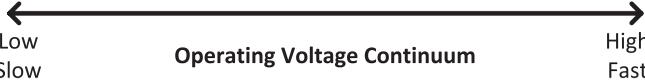
Date of publication: 17 November 2016; date of current version: 23 February 2017.

32 nm for performance sensitive workloads. While this gain is not insignificant, NTC is needed most in new technology nodes because of increased power densities, yet energy gain in 32 nm is nearly half of the gain in 180 nm.

Foundries have initiated a fundamental switch from planar to FinFET transistors at the 22–16-nm node and below, opening a new chapter in Moore's law. However, NTC in FinFET is largely unexplored. FinFET differs significantly from planar technology, with much improved channel characteristics, which have the potential to dramatically improve NT performance. This paper examines FinFET's impact on NT.

First, we present an analytical model of NT's energy gain and, by using this model along with the methodology in the Methodology section, key device characteristics that are responsible for NT performance in FinFETs are analyzed in the Device characteristics section. With this knowledge, we then compare NTC performance in three planar technologies and three FinFET technologies from 40 to 7 nm.

Table 1 Voltage scaling operating scenarios, from ultralow supply voltages to traditional nominal-voltage operation.

Voltage:	Ultra-Low	Near-Threshold	Nominal
Goal:	Minimize Energy (Latency Insensitive)	Minimize Energy (Latency Sensitive)	Maximize Single-Core Performance
Latency:	Unconstrained	Fixed to Latency of 1 Core @ Nominal	Minimized
Cores within TDP:	Many Cores	Some Cores	Few Cores
System Configuration:			
			

Analytical model

A simple analytical model is introduced to better understand underlying effects on device parameters. Though this model does not have high accuracy, especially for recent technology nodes, it is beneficial in understanding the effects of device parameters on NTC performance.

Voltage scaling can be viewed as a continuum of three operating scenarios from traditional nominal voltage operation to ultralow-voltage operation (Table 1). Nominal voltage operates a core at its peak clock frequency, therefore single-threaded performance is maximized. However, nominal voltage also consumes the most power and thus limits the system to the fewest number of cores that can be active within a thermal design power budget. Scaling down voltage to the ultralow, subthreshold region greatly reduces power demands, allowing for more cores to operate within a power budget. However, voltage scaling also significantly degrades clock frequency, so ultralow voltage is not suitable for workloads that are latency sensitive.

NT is applied to reduce the energy of a task when latency sensitive, balancing ultralow and nominal operating modes by parallelizing a task at low voltages to regain lost performance from clock frequency degradation [5]. A previous study [7] developed a systematic methodology for defining the NT operating

point by considering performance sensitivity through fixing the latency of a task to that of the task running on a single core at nominal voltage. As voltage is lowered to NT, clock frequency decreases and subsequently latency increases. However, this latency increase can be balanced through speeding up the task through parallelism (Table 1, middle). This is the definition of NT we use in this work. As an overall metric of system performance the figure of merit (FoM) is a number of tasks within a thermal budget divided by task latency. If a task consumes $1/X$ of the fixed power budget, then X tasks can be run on the system. If each task becomes more power efficient, or latency improves for the same power consumption, then FoM will increase.

The energy of a task can be split up into two categories, dynamic and static, given by

$$E_{\text{total}} = E_{\text{dynamic}} + E_{\text{static}} \quad (1)$$

Dynamic energy is the working energy needed to switch inputs of transistors and values of wires for calculations or communications during a task's execution. Dynamic energy can be modeled as a charge on a capacitor, and thus varies quadratically with supply voltage

$$E_{\text{dynamic}} \propto C_{\text{switch}} V_{dd}^2 \quad (2)$$

Static energy is caused by leakages of a circuit regardless of whether a task is executing. Static energy is usually dominated by subthreshold leakage through a transistor's source and drain and is dependent on the supply voltage and the period of time for a task to run

$$E_{\text{static}} \propto I_{\text{leak}} V_{dd} T_{\text{task}} \quad (3)$$

The time for task completion depends inversely on the clock frequency of a core which, to first order, is inversely proportional to circuit delay and can be modeled using the alpha power law of a transistor [8]

$$T_{\text{task}} \propto 1/f \propto \frac{V_{dd}}{(V_{dd} - V_t)^\alpha} \quad (4)$$

Note that the alpha power law is used as a first-order approximation in this work to understand general trends, but plotted results use more accurate HSPICE models. Additionally, actual task completion scaling depends not only on circuit delay of core logic, but also on the many levels of memory hierarchy on a processor. In this work, we only consider

scaling of core logic gates. From the above relationships, the dynamic energy monotonically decreases with supply voltage while the leakage energy initially decreases because I_{leak} and supply voltage drop. However, the task completion time rises exponentially at NT voltages, and thus static energy increases as V_{dd} continues to be lowered. Energy is minimized when the margin cost of dynamic and static energy are in balance, in essence when the dynamic energy gain of scaling down voltage is equal to the marginal cost of static energy

$$\frac{\delta E_{dynamic}}{\delta V_{dd}} = -\frac{\delta E_{static}}{\delta V_{dd}}. \quad (5)$$

Let V_{opt} be the operating voltage at which energy minimization occurs.

Up until now we have neglected the energy required to maintain task latency. Parallelism overhead of a program can be modeled through Amdahl's law [9], where the speedup of a parallelized program is given by

$$\text{Speedup} = \frac{n}{1 - P_s + P_s n} \quad (6)$$

where n is the number of cores parallelized over and P_s is the percent serial coefficient of the workload. A perfectly parallelizable program has a $P_s = 0\%$, while higher percent serials indicate less of the code is parallelizable, up until $P_s = 100\%$, implying the workload is completely serial. In this work, we consider a fixed latency constraint when performance sensitive, so that the speedup through parallelism has to balance any performance loss from longer circuit delay as a consequence of scaling to NT

$$\text{Speedup} = \frac{T_{task, NTC}}{T_{task, nominal}}. \quad (7)$$

For a task that does have parallelism overhead, and is not perfectly parallel, the energy is derated by a factor of $(n/\text{Speedup})$ compared to the perfectly parallelizable baseline. The total energy when parallelizing is then

$$E_{total, parallel} \propto \frac{n}{\text{Speedup}} E_{total, ideal} \quad (8)$$

where $E_{total, ideal}$ is the energy of a perfectly parallelizable workload. Since Amdahl's law has asymptotic behavior, this ratio worsens when V_{dd} is very low, thus causing $E_{total, parallel}$ to deviate significantly from $E_{total, ideal}$ for workloads that are not perfectly parallelizable. A workload that is latency insensitive does not need to be parallelized across cores and thus does not incur parallelization overhead, hence always has energy usage equal to the ideally

parallelizable (Speedup = n) baseline energy of $E_{total, ideal}$.

Figure 1 demonstrates how parallelism overheads increase the minimum energy and decrease voltage scaling's efficacy. With a perfectly parallelizable program ($P_s = 0\%$), the minimum energy is 8% of the energy at nominal. A serial coefficient of $P_s = 2\%$ raises minimum energy to 28% of the energy of nominal for this example technology. Figure 1 includes FoM normalized to nominal voltage, showing improvement as energy per task is reduced. The number of cores in which the task is parallelized, n , is 10, 9, and 6, when P_s is 2%, 5%, and 10%, respectively. For $P_s = 0\%$, approximately 30 cores would be required to maintain a fixed latency.

Methodology

The framework in this work is similar to [7] and is split into two components: circuit characterization, to extract circuit delay and energy, and architectural models to account for parallelism overheads. For this work, ARM developed a set of predictive technology HSPICE models that include effects specific to FinFETs. The models were created based on published numbers, historical trends, and informed assumptions and calculations. The circuit simulations in this work use HSPICE BSIM Level 72 for FinFET 7-, 10-, and 14-nm models, and Level 54 for planar 20-, 28-, and 40-nm models.

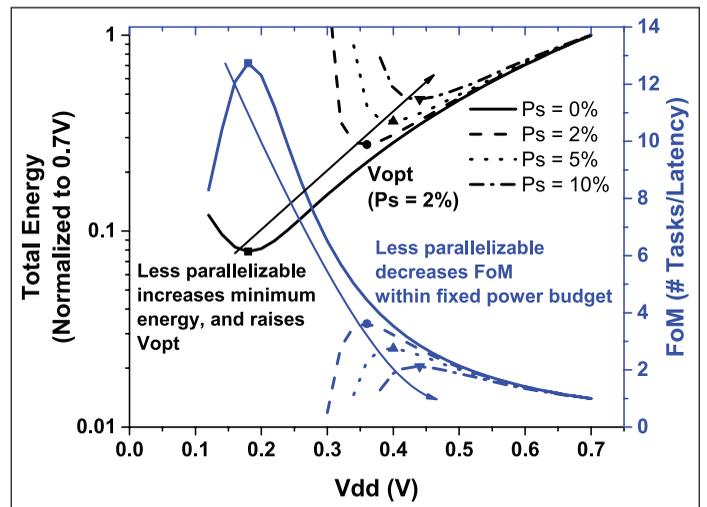


Figure 1. Total energy of a task increases with a higher serial coefficient (P_s) since parallelism overheads limit voltage scalability as task latency is fixed. FoM is improved as energy per task is reduced.

Predicting future technology nodes is difficult as many technological challenges have yet to be overcome. The International Technology Roadmap for Semiconductors (ITRS) provides estimates of many future device parameters. However, ITRS reports are driven by future technology requirements and are not necessarily representative of what is realizable. Therefore, ITRS tends to provide an optimistic outlook while industry estimates are more conservative [2]. Additionally, ITRS does not provide simulation device models, so ITRS data cannot be used directly to gauge efficacy of voltage scaling.

The canonical circuit simulated to characterize voltage scalability is a chain of 31 inverters in fanout-of-4 (FO4), with a 15% activity factor to emulate reasonably deep processor pipelines. Though actual critical paths are composed of more complex gates, we found inverters to be accurate for comparing performance and energy between operating voltages and technology.

Ease of parallelism was modeled through Amdahl's equation [9] to illustrate sensitivity to performance loss from voltage scaling. We chose an Amdahl serial coefficient of 2% which is higher than all but two of the benchmarks [7] from SPLASH-2 [10], a scientific benchmarks intended to evaluate parallel systems.

Device characteristics

Transistor devices have a multitude of interrelated characteristics, but we focus on a few key

parameters relevant to FinFET. We first examine the effects of three basic device characteristics impacting NT performance: drain-induced barrier lowering, subthreshold slope, work function (i.e., threshold voltage). Then, we expand this analysis by including within-cell and back-end-of-line parasitics.

Work function

Work function changes the transistor's threshold voltage V_t , with lower threshold voltages exhibiting increased leakage. If leakage is a significant portion of the total energy, a lower threshold voltage negatively impacts voltage scalability since static energy is more significant and therefore V_{opt} is higher. However, threshold voltage impacts clock frequency scaling through changing transistor ON current $T_{task,NTC} \propto 1/(V_{dd} - V_t)^\alpha$. Figure 2, top, shows normalized FO4 circuit delay (i.e., $T_{task,NTC}$) for five transistor threshold voltages (V_t is 282, 213, 159, 104, and 47 mV for 0.6–2800 nA/ μm). The 0.6-nA/ μm leakage device exhibits the worst circuit delay voltage scalability. For instance, at $V_{dd} = 360$ mV the FO4 delay is 10 \times higher than at the nominal voltage of $V_{dd} = 700$ mV. Therefore, lower V_t devices can voltage scale to lower V_{dd} with the same FO4 degradation.

Better delay scalability allows a task to operate at lower V_{dd} while maintaining performance, as less parallelism is needed for a fixed latency constraint. The FoM across V_t is shown in Figure 2, bottom. The 0.6-nA/ μm device has peak FoM of 360 mV; below this parallelism overhead becomes significant. The 6- and 40-nA/ μm device voltages scale lower and have better FoM since they exhibit less FO4 delay degradation. Despite very low V_{opt} , the 350- and 2800-nA/ μm devices are not good because leakage dominates energy overhead.

For the 2% serial coefficient workload, the 40-nA/ μm device achieved the best FoM, with an energy gain of 8.7 \times at a $V_{opt} = 220$ mV. Devices with this order of magnitude of leakage are known as high-performance (HP) transistors as defined by ITRS. Analysis in the subsequent sections uses the HP device as a baseline in which to compare.

Drain-induced barrier lowering

Drain-induced barrier lowering (DIBL) is a short-channel effect that reduces threshold voltage

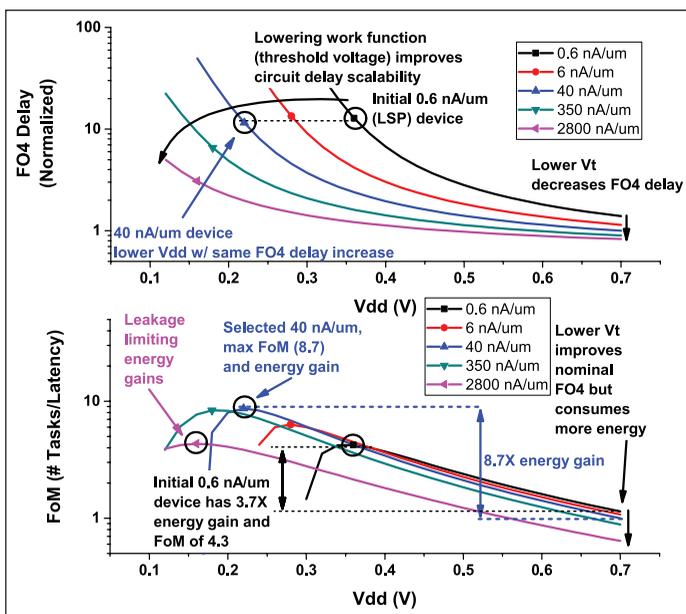


Figure 2. Circuit delay scaling (top) and FoM (bottom) for varying threshold voltage in 7-nm FinFET.

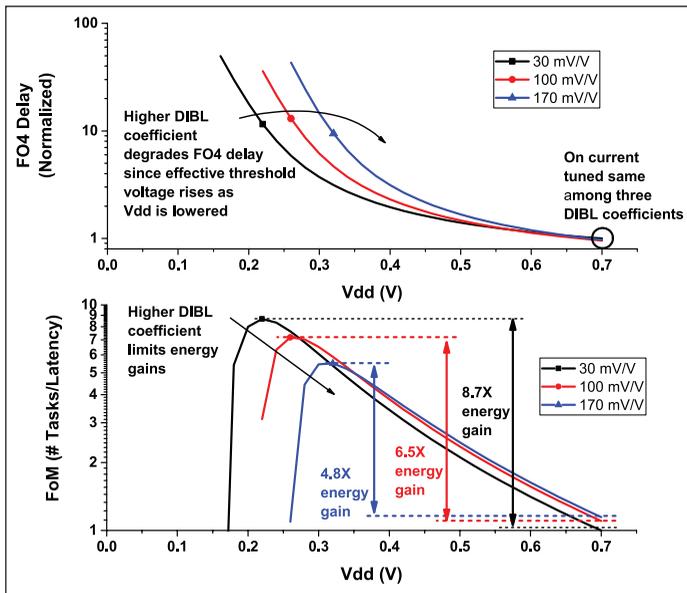


Figure 3. Circuit delay scaling (top) and FoM (bottom) as DIBL coefficient increases in 7-nm FinFET.

as drain-source voltage V_{ds} increases. This can be modeled through [11]

$$V_t = V_{t0} - \eta V_{ds} = V_{t0} - \eta V_{dd}$$

where V_{t0} is the threshold voltage with no drain-source potential and η is the DIBL coefficient (typically around 100 mV/V [11]). As V_{dd} is lowered, DIBL causes V_t to increase and therefore the transistor overdrive voltage $V_{ov} = V_{dd} - V_t$ rapidly collapses and severely limits voltage scalability. This directly affects task completion time

$$T_{\text{task,NTC}} \propto 1/(V_{dd} - V_t)^\alpha = 1/V_{ov}^\alpha.$$

Therefore, DIBL also affects the speedup needed to maintain a latency constraint. As the DIBL coefficient η increases, $T_{\text{task,NTC}}$ lengthens, as shown in Figure 3. Each device is tuned to match both OFF current and the ON current at nominal supply voltage of the other devices.

For workloads that are sensitive to performance, the poor voltage scalability in clock frequency translates to limited energy gains and an increasing V_{opt} as more parallelism is required (Figure 3, bottom). Therefore, an improved (lower) DIBL coefficient directly improves NT energy gains and performance in NT for latency sensitive applications.

For latency insensitive workloads, energy and V_{opt} do not change significantly and, in fact, leakage

can be slightly improved at low voltages because of increased threshold voltages.

Subthreshold slope

Subthreshold slope (S_S) is the reduction in drain-source current when threshold voltage is raised, typically expressed as mV's per decade of reduction, which can be modeled as [11]

$$I_{\text{leak}} = I_{ds0} \exp\left(\frac{-V_t}{n V_T}\right).$$

The denominator $n V_T$ sets the subthreshold slope of the device. Subthreshold slope is given in units of mV/dec, and a steeper slope (smaller mV/dec) allows for a lower threshold voltage to achieve the same leakage, as less mV's are required to decimate source-drain leakage currents. Lower threshold voltage is desirable to increase transistor headroom. Two things happen as subthreshold slope increases (becomes shallower): 1) for the same threshold voltage, leakage increases; and 2) the current drivability of the transistor improves (i.e., the transistor is better able to drive a load at lower voltages).

Leakier transistors cause the total energy at NT to increase, however increased drivability improves $T_{\text{task,NTC}}$ scaling with V_{dd} , as shown in Figure 4,

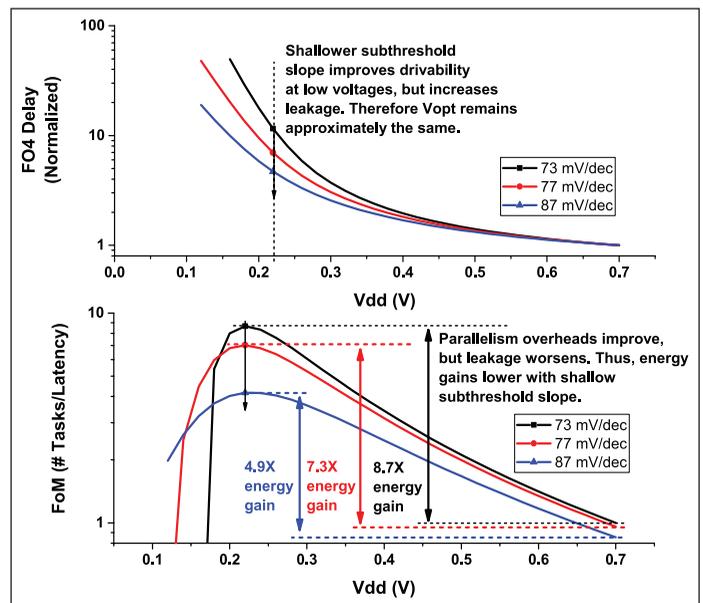


Figure 4. Circuit delay scaling (top) and FoM (bottom) as subthreshold slope becomes less steep in 7-nm FinFET.

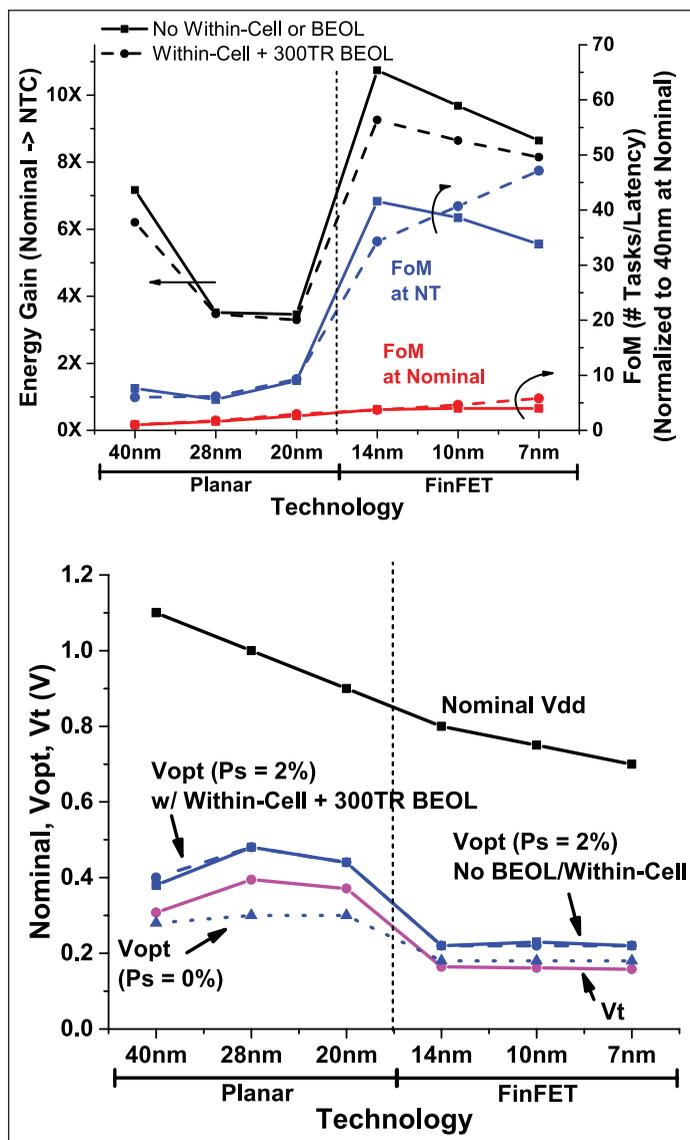


Figure 5. Energy gain and FoM across technology (top) and V_{opt} across technology (bottom).

top, with increasing subthreshold slope (higher mV/dec). These two effects (higher leakage and better drivability) oppose each other for performance sensitive workloads, thus V_{opt} stays relatively constant (Figure 4, bottom). However, for latency insensitive workloads, where $P_s = 0\%$, improved circuit delay scaling has no impact on energy and thus increases both V_{opt} and total energy, limiting achievable energy efficiency gains.

Back-end-of-line and within-cell parasitics

Transistors are interconnected through wires and vias, thus back-end-of-line (BEOL) parasitic capacitance and resistance needs to be considered when

analyzing voltage scaling performance. Within-cell parasitics are added to the characterization simulations in this study by extracting representative standard cell layouts of 1 \times -sized inverters in the predictive technologies provided by ARM. These parasitic models include source, drain, and gate resistance due to trench contacts and local interconnects introduced at sub-20-nm nodes and the corresponding coupling capacitances between input/output pins and power rails. Within-cell parasitics contribute up to half of the total gate delay.

Wire parasitics are modeled in our HSPICE simulations through π -models [12] of predicted resistance and capacitance per unit length of a low-level metal wire with minimum width and spacing. The wire length was swept across multiples of minimum track pitch, from 150 tracks to 1200 tracks. Fanout-of-4 circuit delay was measured and, though the absolute delay increases as within-cell and wire load is added, energy-efficiency gain is nearly identical across the different wire lengths in 7 nm. The absolute FoM is worse with longer wire lengths because of added capacitance and FO4 delay. When FoM is normalized for each individual wire length at 0.7 V, the FoM is relatively unchanged, thus BEOL does not significantly impact voltage scaling analysis in 7-nm FinFET. Older technologies are impacted more by BEOL as we show in the next section.

Technology trends

In this section, we expand the analysis to older FinFET and planar technology nodes. Models for all six technology generations, provided by ARM, target HP transistors (40-nA/ μ m leakage). By consistency targeting the models, better consistency is obtained compared to [7] which used disparate technology models from different foundries.

The energy gain across the six technology nodes is shown in Figure 5, top, both without and with BEOL parasitics. Of the planar nodes, 40 nm has the best energy gain at 6.2–7.2 \times and this reduces in 28 and 20 nm to 3.5 \times and 3.3 \times , respectively, confirming the trends seen in [7]. Energy gain is diminished in newer planar nodes because of stagnated V_t but lower nominal V_{dd} and other device effects causing poor circuit delay scaling.

Transitioning to FinFET in 14 nm shows much better energy gains of 9.3–10.7 \times because threshold voltage has dropped by approximately 210 mV, with the same leakage characteristics, and DIBL coefficient

Table 2 Summary of technology parameters and expected NT energy efficiency with 2% serial workload and BEOL wire models.

Technology	V_{nom} (V)	E_{gain}	V_{opt} (mV)	V_t (mV)	DIBL (mV/V)	S_S (mV/dec)
FinFET						
7nm	0.70V	8.2×	220	159	22	71
10nm	0.75V	8.6×	220	161	28	71
14nm	0.80V	9.3×	220	165	32	72
Planar						
20nm	0.90V	3.3×	440	372	173	101
28nm	1.00V	3.5×	480	395	171	104
40nm	1.10V	6.2×	400	308	102	92

has improved from 173 mV/V in 20 nm to 32 mV/V in 14 nm. In successive FinFET technologies, the energy gain decreases because V_{dd} is dropping by 50 mV per generation while V_t is gradually decreasing (as little as 2 mV). Thus, the delta between nominal V_{dd} and V_{opt} compresses with each generation.

The NT voltage V_{opt} is shown across the six technology nodes in Figure 5, bottom, both with and without BEOL parasitics. With $P_S = 2\%$, V_{opt} is approximately 40–80 mV above V_t . Workloads with higher P_S have higher V_{opt} because of increased parallelism overheads. BEOL wire loads do not significantly change V_{opt} , but do affect FoM and energy gain. A summary of device parameters and NTC efficiency gains is shown in Table 2.

NTC HAS RECEIVED much interest for overcoming power dissipation limits through improving energy efficiency. However, NTC was observed to be less effective in recent planar technology nodes [7]. In this work, we evaluated the impact of device characteristics on voltage scaling and showed how to target a FinFET technology for improved NT operation. FinFET enables significant voltage scaling improvements over planar nodes because of improved channel characteristics, namely less DIBL and steeper subthreshold slopes.

FinFET allows for more effective NT operation than ever before because it achieves better voltage scalability and higher area densities. A deeper analysis of variation and architectural-level modeling would better guide necessary circuit and architectural techniques needed to maximize energy gains. Additionally, performance-sensitive NT operation requires algorithms to readily parallelize over many

cores, requiring further research on building efficient parallel systems.

Acknowledgment

We would like to thank ARM for assistance. This work was supported by the Defense Advanced Research Projects Agency (DARPA) under agreement HR0011–13-2–000. ■

References

- [1] R. H. Dennard, F. H. Gaensslen, V. L. Rideout, E. Bassous, and A. R. LeBlanc, "Design of ion-implanted MOSFET's with very small physical dimensions," *IEEE J. Solid-State Circuits*, vol. 9, no. 5, pp. 256–268, Oct. 1974.
- [2] H. Esmailzadeh, E. Blem, R. S. Amant, K. Sankaralingam, and D. Burger, "Dark silicon and the end of multicore scaling," in *Proc. 38th Annu. Int. Symp. Comput. Architecture*, Jun. 2011, pp. 365–376.
- [3] B. Zhai, R. G. Dreslinski, D. Blaauw, T. Mudge, and D. Sylvester, "Energy efficient near-threshold chip multi-processing," in *Proc. ACM/IEEE Int. Symp. Low Power Electron. Design*, Aug. 2007, pp. 32–37.
- [4] L. Chang et al., "Practical strategies for power-efficient computing technologies," *Proc. IEEE*, vol. 98, no. 2, pp. 215–236, Feb. 2010.
- [5] R. G. Dreslinski, M. Wiecekowsky, D. Blaauw, D. Sylvester, and T. Mudge, "Near-Threshold Computing: Reclaiming Moore's Law through energy efficient integrated circuits," *Proc. IEEE*, vol. 98, no. 2, pp. 253–266, Feb. 2010.
- [6] M. B. Taylor, "Is dark silicon useful?: Harnessing the four horsemen of the coming dark silicon apocalypse," in *Proc. 49th ACM/EDAC/IEEE Design Autom. Conf.*, Jun. 2012, pp. 1131–1136.

- [7] N. Pinckney et al., "Assessing the performance limits of parallelized near-threshold computing," in *Proc. 49th ACM/EDAC/IEEE Design Autom. Conf.*, Jun. 2012, pp. 1143–1148.
- [8] T. Sakurai and A. R. Newton, "Alpha-power law MOSFET model and its applications to CMOS inverter delay and other formulas," *IEEE J. Solid-State Circuits*, vol. 25, no. 2, pp. 584–594, Apr. 1990.
- [9] G. M. Amdahl, "Validity of the single processor approach to achieving large scale computing capabilities," in *Proc. Spring Joint Comput. Conf.*, 1967, pp. 483–485.
- [10] C. Bienia, S. Kumar, and K. Li, "PARSEC vs. SPLASH-2: A quantitative comparison of two multithreaded benchmark suites on chip-multiprocessors," in *Proc. IEEE Int. Symp. Workload Characterization*, Sep. 2008, pp. 47–56.
- [11] K. Roy, S. Mukhopadhyay, and H. Mahmoodi-Meimand, "Leakage current mechanisms and leakage reduction techniques in deep-submicrometer CMOS circuits," *Proc. IEEE*, vol. 91, no. 2, pp. 305–327, Feb. 2003.
- [12] T. Sakurai, "Approximation of wiring delay in MOSFET LSI," *IEEE J. Solid-State Circuits*, vol. 18, no. 4, pp. 418–426, Aug. 1983.

Nathaniel Pinckney is currently a Research Scientist at Nvidia, Austin, TX, USA. Pinckney has a PhD in electrical engineering from the University of Michigan, Ann Arbor, MI, USA. He is a Member of the IEEE.

Supreet Jeloka is currently working toward a PhD at the University of Michigan, Ann Arbor, MI, USA. Jeloka has a BTech in electronics and communication engineering from NIT, Warangal, India and an MS in electrical engineering from the University of Michigan. He is a Student Member of the IEEE.

Ron Dreslinski is currently an Assistant Professor at the University of Michigan, Ann Arbor, MI, USA. Dreslinski has a PhD in computer science

and engineering from the University of Michigan. He is a Member of the IEEE.

Trevor Mudge is on the faculty of the University of Michigan, Ann Arbor, MI, USA. Mudge has a PhD in computer science from the University of Illinois. He is a Life Fellow of the IEEE.

Dennis Sylvester is a Professor of Electrical Engineering and Computer Science at the University of Michigan, Ann Arbor, MI, USA. Sylvester has a PhD from the University of California Berkeley, Berkeley, CA, USA. He is a Fellow of the IEEE.

David Blaauw has been on the faculty at the University of Michigan, Ann Arbor, MI, USA, since August 2001, where he is a Professor. Blaauw has a PhD in computer science from the University of Illinois. He is a Fellow of the IEEE.

Lucian Shifren is currently a Principal R&D Engineer at ARM Inc., San Jose, CA, USA. Shifren has a PhD in electrical engineering from Arizona State University, Tempe, AZ, USA and an MBA from Portland State University, Portland, OR, USA. He is a Senior Member of the IEEE.

Brian Cline is currently a Staff Design Engineer at ARM Inc., San Jose, CA, USA. Cline has a PhD in electrical engineering from the University of Michigan, Ann Arbor, MI, USA. He is a Member of the IEEE.

Saurabh Sinha is a Senior Design Engineer in the R&D division of ARM Inc., Austin, TX, USA. Sinha has a PhD in electrical engineering from Arizona State University, Tempe, AZ, USA. He is a Member of the IEEE.

■ Direct questions and comments about this article to Nathaniel Pinckney, University of Michigan, Ann Arbor, MI 48109, USA; npfet@umich.edu.