Near-Threshold Computing in FinFET Technologies: Opportunities for Improved Voltage Scalability

Nathaniel Pinckney[1], Lucian Shifren[2], Brian Cline[3], Saurabh Sinha[3], Supreet Jeloka[1], Ronald G. Dreslinski[1], Trevor Mudge[1], Dennis Sylvester[1], David Blaauw[1]
[1]University of Michigan, Department of Electrical Engineering and Computer Science, Ann Arbor, MI.
[2]ARM Inc., San Jose, CA. [3]ARM Inc., Austin, TX.
npfet@umich.edu

ABSTRACT

In recent years, operating at near-threshold supply voltages has been proposed to improve energy efficiency in circuits, yet decreased efficacy of dynamic voltage scaling has been observed in recent planar technologies. However, foundries have introduced a shift from planar to FinFET fabrication processes. In this paper, we study 7nm FinFET's ability to voltage scale and compare it to planar technologies across three dynamic voltage scaling scenarios. The switch to FinFET allows for a return to strong voltage scalability. We find up to 8.6× higher energy efficiency at NT compared to nominal supply voltage (vs. 4.8× gain in 20nm planar).

1. INTRODUCTION

Transistor threshold voltages have stagnated since the 90nm technology node, deviating from constant-voltage scaling theory and directly limiting supply voltage scaling. This has directly influenced processor design, shifting architectures from increased clock speeds to increased number of cores with each generation. Recent work has observed that we are at a point where not all cores can be simultaneously active at full voltage and clock frequency without exceeding thermal design budgets [1]. Consequently, at any given time large sections of a chip will remain inactive in order to not exceed thermal limits of the package and cooling system. This scenario, dubbed *dark silicon* [1], has shown that the percent of chip inactivity is increasing each generation and the majority of chip area in a CPU could be dark within the next few years.

Stagnation in technology scaling demands aggressive circuit and architectural advancements to meet future performance goals. Voltage scaling is a key circuit technique directly impacting architectural decisions [2], but lowering a core's power supply voltage can no longer be limited to periods of idle workloads. Instead voltage scaling must be aggressively leveraged to regain system performance, as processors become increasingly power constrained, by running much closer to a transistor's threshold voltage than in years past [3–6]. Operating at "near-threshold computing" (NTC) supply voltages achieves sizable energy gains with moderate performance loss. Lost performance, due to reduced clock frequency from increased logic delay, can be regained by parallelizing across cores [1, 5, 5, 7].

Foundries have introduced a fundamental switch from planar transistors to FinFET at the 22 - 16nm node and below, opening a new chapter in Moore's law. However, NTC in FinFET has not been explored, unlike previous planar studies [5]. In this paper, we quantify improved dynamic voltage scalability (benefits of operating at near-threshold as oppose to nominal supply voltage) in FinFET technology nodes. With three dynamic voltage scaling scenarios we examine how to maximize eneryg efficiency or minimize task completion latency in six technology nodes, three FinFET (7nm, 10nm, 14nm) and three planar (20nm, 28nm, 40nm), using transistor models developed by ARM. Unlike previous near-threshold studies, we also include area constraints in our analysis to understand what is achievable within a reasonable area budget.

We find FinFET has significant voltage scaling advantages over planar technologies that allows improved energy efficiency through dynamic voltage scaling. In 7nm, $6.3-8.6\times$ energy efficiency gains (near-threshold compared to planar operation) are possible for performance sensitive and insensitive tasks, compared with $2.7-4.8\times$ in 20nm planar, when area unconstrained. When area constrained, achievable energy efficiency gains drop to $2.4-2.5\times$ for 7nm and $1.3\times$ for 20nm. If single task performance is prioritized over energy efficiency, then 40% faster latency is possible in 7nm FinFET compared to no gains in 20nm planar, as long as the task is sufficiently parallelizable.

2. SCALABILITY ANALYSIS

2.1 Methodology

This analysis uses a similar framework to that used by Pinckney et al. [5] for estimating energy and performance when voltage scaling in planar technologies. Circuit simulations use HSPICE models for 7nm, 10nm, and 14nm FinFET, and for 20nm, 28nm, and 40nm planar, of which all sets were developed by ARM based on published numbers, historical trends, and informed assumptions and calculations. The canonical circuit simulated to model circuit effects is a chain of thirty-one inverters, which emulates reasonably deep processor pipelines. Though actual critical paths are composed of more complex gates, we found inverters provided sufficient accuracy for comparing performance and energy between operating voltages and technologies. Within our circuit model, we also included back-end-of-line (BEOL) parasitics, which are additional capacitors and resistors from wires that interconnect gates on a chip. Back-end-of-line is important to model as achievable energy gains when BEOL is included are lower than with ideal wires. Lastly, impact from across die mismatch variation is accounted for by derating a percentage of the total energy and minimum clock pe-

DAC '16, June 05-09, 2016, Austin, TX, USA

DOI: http://dx.doi.org/10.1145/2897937.2898049

riod, proportional to increases in variation, to penalize low voltage operation.

A percent serial coefficient of 2% is used to describe the voltage scaling scenarios, as a representative value that is higher than all but two SPLASH-2 benchmarks [5] (a scientific benchmark suite [8]). However, we provide a table of with three sets of serial coefficients at the end of this section. Final energy and performance estimates were calculated by combining circuit and architectural data using MATLAB. The final energy and performance numbers are used to generate figure-of-merit estimates under different voltage scenarios.

2.2 Voltage Scaling Scenarios

When considering the performance of a near-threshold system, three voltage scaling scenarios are evaluated depending on the prioritization of task latency versus overall system efficiency. The three scenarios are summarized in Table 1. To simplify analysis, all tasks running on a system are assumed to be identical.

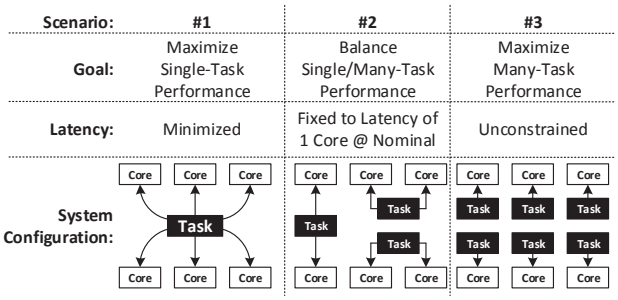

Table 1: Voltage scaling scenarios when latency is minimized, fixed, or unconstrained.

The first voltage scaling scenario is when all processors in a system are utilized by a single task to maximize speedup, and no other tasks are run on the system. Absolute maximum single task performance is achieved, however because adding additional cores to a task may only marginally improve performance, the overall energy efficiency is impacted. The second scenario considers multiple tasks running the system, but the tasks are performance sensitive which is accounted for by constraining a task's latency to that of the task running on a single core at maximum supply voltage and frequency. As supply voltage is lowered and clock frequency degrades, tasks are parallelized across more cores until the fixed latency constraint is met. This definition was used to define the near-threshold region by Pinckney et al. [5]. The final scenario again considers multiple tasks, but each task is assigned to a single core and latency is completely unconstrained. In this scenario tasks may take very long to finish, but run very efficiently by minimizing energy consumption.

Each of these voltage scaling scenarios is described in detail within the following subsections. A percent serial of 2% is used to initially explain each scenario's behavior, but we conclude this section by expanding analysis to 5%, 10%, 15%, and 25% serial.

2.2.1 Minimizing Latency

The figure-of-merit used in this work considers both the latency per task and the total number of simultaneous tasks that can be run. As an illustrative example, consider a single core running at 1V, 100MHz, and consuming 1W as shown in Figure 1 (top). A task runs on this core and completes in a quarter-second, after which it is rerun on the core. This continues over-and-over, so that four tasks complete per second at 1V. The figure-of-merit we use is number of concurrent tasks divided by task latency, while the baseline is of the system running at nominal voltage (1V in this example, though this is technology dependent). The baseline in this example has a figure-of-merit of four tasks per second.

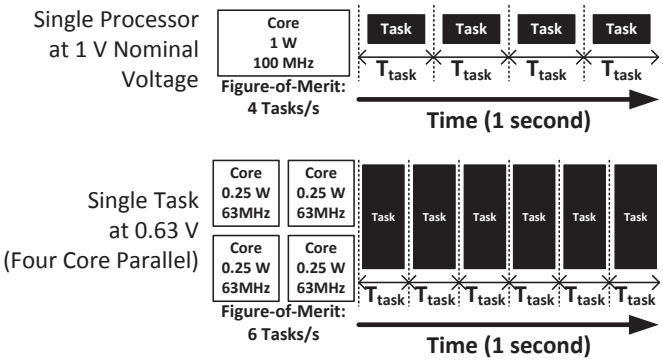

Figure 1: Figure-of-merit of single-task scenario (minimizing latency). All available cores, within power and area budgets, are assigned to run the task. Improving core efficiency and parallelism allows for faster task latency and higher FoM. In this example, FoM is improved from four tasks per second to six ($1.5\times$ gain).

Next, the voltage is scaled from 1V to 0.63V (Figure 1, bottom). In this example, the core slows from 100Mhz to 63Mhz and only consumes 0.25W. Since the original power at 1V is 1W, we allow three additional cores to be added at 0.63V so that the total power consumed at low voltage is identical to that at nominal 1V. All four cores are used to run the task. Since the clock speed only reduced by -37%, but the task is parallelized across four cores, thus the task latency is faster than when running on a single core at 1V. Latency improves by $1.5\times$, increasing figure-of-merit to six tasks per second. In this example, our figure-of-merit gain is $6/4 = 1.5\times$. We use FoM gain and the relative FoM between technologies for the results in this paper.

The FoM of the single-task scenario is shown in Figure 2 while supply voltage is swept across technologies. To include effects of dark silicon, the starting constraint is with a power and area budget sufficient to run one core in 40nm, and subsequent technologies scale using energy estimates of the simulated ARM predictive transistor models, along with area scaling derived from publicly available foundry data. Power budget, with no area budget, only is shown in the left plot, and FoM with power plus area budget is shown on the right plot. Each color represents a different CMOS technology, from 40nm to 7nm. The power budget dominates FoM at high voltage but limits FoM at low voltages, as cores consume less power.

Area is extremely limited in planar nodes and, in this example, can only support one core in 40nm, two cores in 28nm and four cores in 20nm. Along with poor circuit delay scaling, area limits best-case energy efficiency gains in planar nodes. Without constraining area, 40nm is able to improve single-task FoM by $1.4\times$ from operating at 0.7V instead of 1.1V. However, newer planar nodes suffer from poorer circuit delay scalability (degradation in clock frequency at low voltages), exemplified by 20nm and 28nm having negligible figure-of-merit increase when voltage scaling.

FinFET technologies exhibit better circuit delay scalability, and are much less area-limited, thus are able to improve single task performance by 30% in 14nm, 40% in 10nm, and 40% in 7nm, even while constraining area. The power budget allows for 4.9 cores

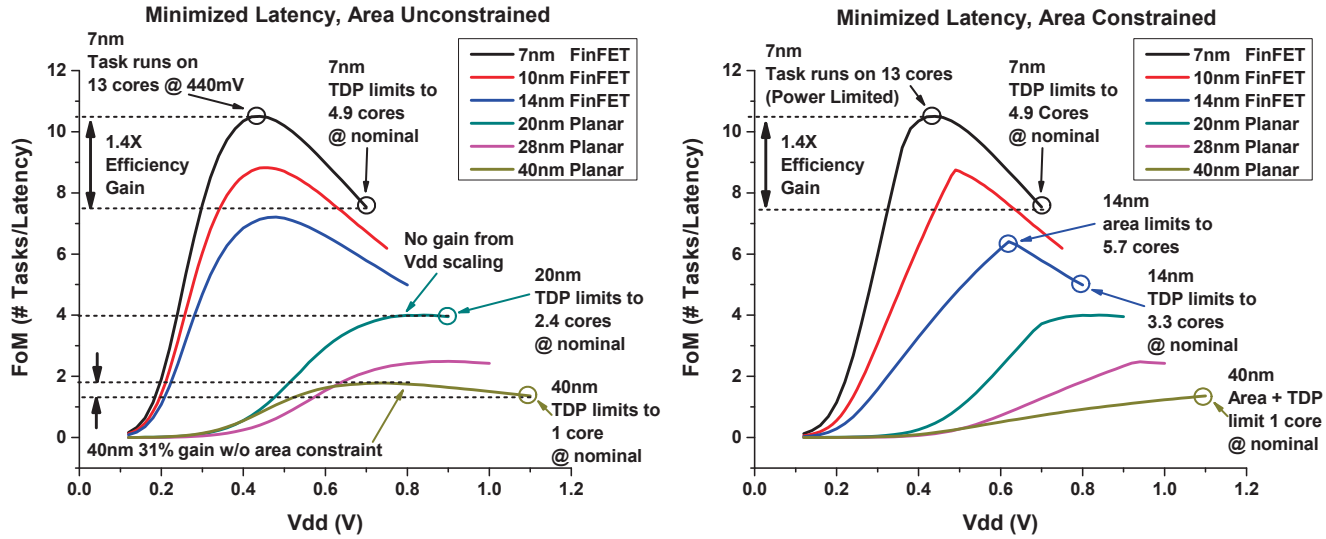

Figure 2: Figure-of-merit of single-task scenario (minimizing latency) across six technologies, without an area budget (left) and with an area budget (right). Amdahl serial coefficient is 2% in resuls shown. In this example, four cores at near-threshold are able to run within the same power budget as a single core at nominal voltage. Therefore, the task is parallelized across those four cores improving latency by 1.5× in this example. FoM (# Tasks/Latency) improved by 1.5× as latency improved but number of tasks remained constant.

at nominal voltage (700 mV) in 7nm FinFET, despite having area for 19 cores. Lowering the voltage in 7nm to 440 mV maximizes figure-of-merit by allowing 13 cores to operate within the power budget. Below 440 mV, figure-of-merit reduces because of degrading clock frequency.

Despite low voltage operation's conventional use only during periods of minimal processor load, FinFET's superior circuit delay scalability and the shift to dark silicon, has introduced a new opportunity for voltage scaling to improve energy efficiency even for high-performance applications, so long as a task is sufficiently parallelizable.

2.2.2 *Fixed Latency Constraint*

Fixing latency balances task performance with overall system efficiency, shown in Figure 3. In this example scenario, task latency matches that of the baseline (nominal 1V operation) by parallelizing a task across two cores at 0.63V. Since only two cores are required to meet the latency constrain, a second task is added to the system and run concurrently with the original task. Thus, the figure-of-merit (# Tasks/Latency) is doubled as two tasks run on the system but latency is constant. The FoM gain is $8/4 = 2\times$.

The second scenario is when figure-of-merit of the system is maximized, subject to latency and power constraints, shown in Figure 4 without (left) and with (right) area constraints. Latency is constrained to that of the task running at nominal supply voltage on a single core. As voltage is lowered the clock frequency degrades, however latency can be maintained by parallelizing. If parallelism overhead is sufficiently small, energy efficiency can be improved while maintaining fixed latency for the task. In other words, for energy savings, the energy overhead needed to parallelize the task across more cores should be less than the energy gain from running at the lower voltage. Recall that power is proportional to energy times rate (rate inversely proportional to latency in this example). Because energy for a task improves, while latency remains constant, additional tasks can be run within the same power budget.

Without an area constraint, 40nm planar had the best figure-of-merit gains across the three planar nodes, and gains become pro-

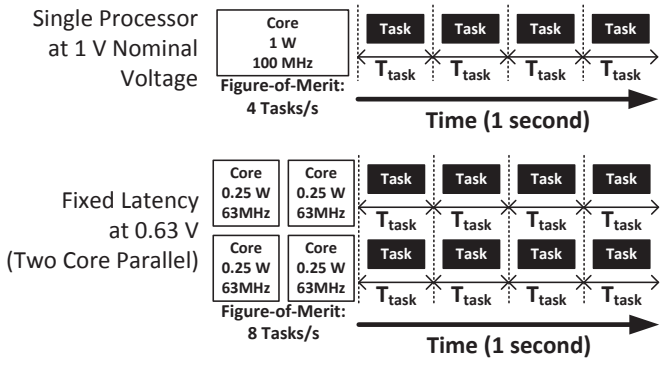

Figure 3: Figure-of-merit of many tasks when latency constrained (fixed latency). As voltage is lowered, tasks are parallelized to match their latency to that of the nominal-voltage, single-core latency. In this example, two independent tasks (each parallelized across two cores) can run within the same power budget as a single core at nominal voltage, while matching latency.

gressively worse in 28nm and 20nm (4.7× gain in 40nm, 2.8× in 28nm, and 2.7× in 20nm). FinFET exceeds all planar nodes in figure-of-merit, with gains of 6.9×, 6.6×, and 6.3× for 14nm, 10nm, and 7nm, respectively. Including an area constraint further limits planar nodes to a maximum gain of 1.3× in 20nm, while FinFET can achieve 2.4× in 7nm because of increased number of cores that fit within the area budget. A practical implementation will fall between the area constrained and unconstrained estimates ($2.4-6.3\times$ energy efficiency gain in 7nm) depending on the size of the core and room dedicated to cores versus other peripherals on an SoC.

2.2.3 *Unconstrained Latency*

Imposing a latency constraint causes task energy efficiency to depend on benchmark parallelizability, since a task is parallelized as supply voltage is lowered in order to meet the latency constraint. When latency unconstrained, a task continues running on a single core despite performance loss, as shown in Figure 5. As a consequence, four tasks are run (one per core) in the example shown and achieves the highest overall figure-of-merit despite increased task latency. The FoM gain in this case is this example is $10/4 = 2.5\times$.

With the area budget of one core in 40nm, all technologies have approximately the same gains as that of the fixed-latency scenario, since the amount of parallelism is relatively small in the fixed-latency scenario (no more than a couple of cores per task), thus overheads from running on multiple cores are negligible. However, without an area constraint the voltage is pushed lower, achieving gains of 8.6×, 10.2×, and 9.7× in 7nm, 10nm, and 14nm FinFET, and gains of 4.8×, 5.6×, and 8.2× in 20nm, 28nm, and 40nm planar, respectively. Since latency is unconstrained, circuit delay scalability with voltage has less of an impact than the previous scenario, therefore higher FoM gains are achievable.

2.3 Sensitivity to Parallelism Overheads

The previous results were for a relatively parallel task (Amdahl percent serial = 2% or less). A higher percent serial reduces achievable efficiency gains, as more parallelism is needed for the same speedup at lower voltage if latency of a task is fixed or minimized. To account for higher serial percentage we swept percent serial for the three different scenarios and show results for 2%, 5%, 10%, 25% in Table 2 for 7nm FinFET and 20nm planar. With and without area constraints are included, to compare achievable efficiency gains for both large and small cores, respectively. In our scaling example, 20nm only has area budget for four cores and power budget for 2.4 cores at nominal Vdd, versus area for 19 cores and power for 4.7 cores in 7nm. The relative figure-of-merit, number of cores per task, and number of total cores is also included in the table, to compare performance across scenarios and technologies, and to gauge which configurations are practical.

Increasing percent serial decreases the achievable gains, especially for the minimum latency scenario, since serial code can never be sped up through parallelism. In 7nm, a percent serial of 25% shows no gain in the single-task scenario. Area constrained and unconstrained is nearly identical in 7nm for a single task, since the number of cores is always less than the area budget. Because of poor circuit delay scaling in 20nm planar, a single-task can never be improved through voltage scaling.

Larger percent serial also reduces gains in the fixed-latency scenario when area unconstrained. However, this effect is marginalized when area constrained, since in this case the number of cores parallelized across is relatively small. Varying the percent serial does not impact the unconstrained latency scenario, as a task always runs on a single core and is never parallelized.

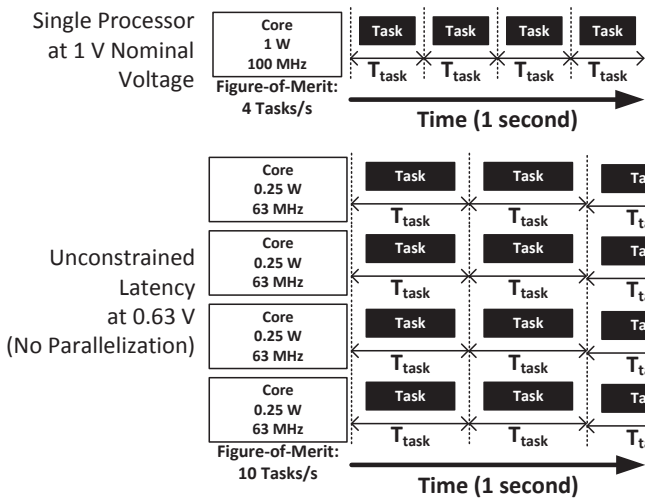

Figure 5: Figure of merit of many tasks when task latency is unconstrained. As voltage is lowered, tasks are not parallelized so many independent tasks are run, one per core, until the power or area budget is reached. This achieves the highest figure-of-merit (# Tasks / Latency) of 10 tasks per second (gain of $2.5\times$) but at the cost of degraded task latency.

The relative figure-of-merits help compare performance across technology. For instance, when latency unconstrained 7nm has a FoM of $70/8.6 = 8.1$ at nominal Vdd, compared to $20/4.8 = 4.2$ in 20nm. Therefore, at nominal Vdd, 7nm has an $8.1/4.2 = 1.9\times$ increase in FoM compared to 20nm, despite three generations of process improvements. Thus, even a $2\times$ gain through dynamic voltage scaling is significant compared to gains from technology improvements. The bottom of Table 2 includes a summary comparing 7nm FinFET to 20nm CMOS planar, by averaging across all values for each scenario.

7nm FinFET is well within the predicted dark silicon regime, where power density has increased to the point of limiting the majority of cores from operating simultaneously at full voltage and frequency. However, FinFET also offers substantial flexibility in architecture design than planar could not offer. Because of FinFET's improved circuit delay scaling and higher area density, designers may realize energy efficient heavily parallelized systems that work over a wide range of workloads. Traditionally, high single-task performance has been accomplished by running cores at maximum frequency and Vdd, yet in FinFET even single-task performance can be improved through voltage scaling.

3. RELATED WORK

The problem of increasing power density has been referred to as *dark silicon* by Esmaeilzadeh et al. [1], since in this regime it is not possible to run all cores on a processor simultaneously at maximum frequency and voltage. Taylor [7] further looks into issues of dark silicon as it impacts architecture, and discusses four solutions, including heterogeneous architectures and voltage scaling.

Leveraging voltage scaling to regain energy is not new, and traditional dynamic voltage and frequency scaling (DVFS) is used extensively in processors. Early low-power subthreshold architectures were presented by Chandrakasan et al. [9] and Wang et al. [10]. More recent low-voltage work [2, 3, 11, 12] shifts to using voltage scaling (near-threshold or NT) for overcoming dark silicon.

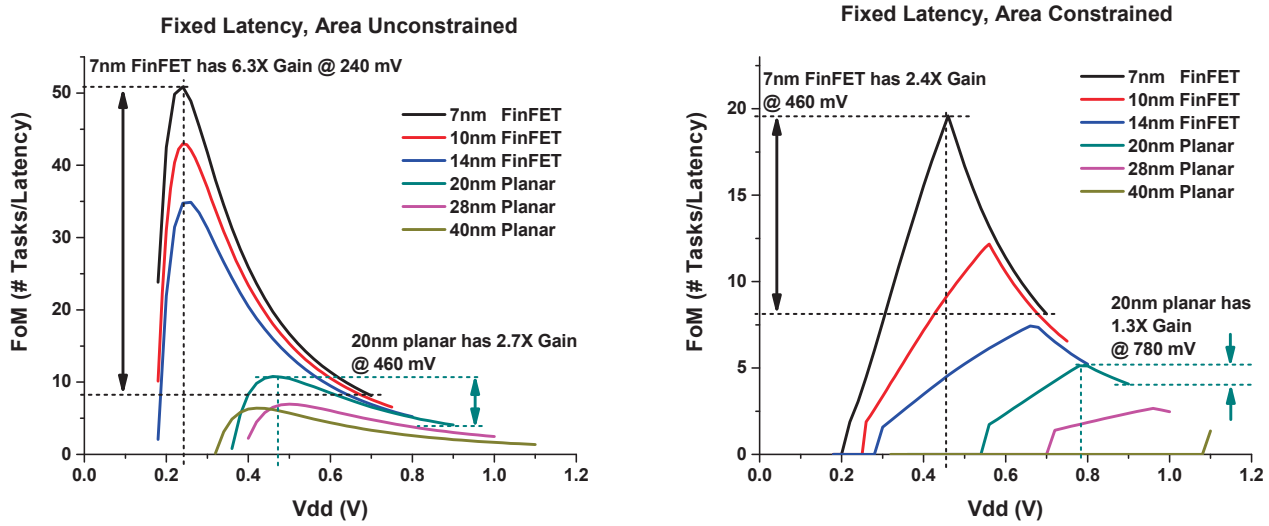

Figure 4: Figure-of-merit of many task system when latency constrained, without an area budget (left) and with an area budget (right). Amdahl serial coefficient is 2% in results shown. In 7nm FinFET a 6.3× FoM improvement is achievable when area unconstrained, compared to 2.7× improvement in 20nm planar. Constraining area limits the number of cores within the system, thus limiting achievable parallelism and reducing FoM gain to 2.4× in 7nm and 1.3× in 20nm, respectively.

FoM Improvement, Maximum FoM across Vdd, and Number of Cores at Maximum FoM												
7nm FinFET												
Scenario	**Single-Task (Minimize Latency)**				**Balanced (Fixed Latency)**				**Many-Task (Unconstrained Latency)**			
	Gain	**FoM**	**Cores /Task**	**Cores Total**	**Gain**	**FoM**	**Cores /Task**	**Cores Total**	**Gain**	**FoM**	**Cores /Task**	**Cores Total**
% Serial	**Area Unconstrained**											
2%	1.4×	11	13		6.3×	51	10	310	8.6×	70	1	1,400
5%	1.2×	8	11		4.9×	40	8.6	210	"	"	"	"
10%	1.1×	6	8		3.7×	30	5.6	100	"	"	"	"
25%	1.0×	4	6		2.2×	18	2.9	32	"	"	"	"
% Serial	**Area Constrained (19 Cores Max)**											
2%	1.4×	14	13		2.4×	20	1.6	19	2.5×	20	1	19
5%	1.2×	8	11		2.4×	19	1.6	19	"	"	"	"
10%	1.1×	6	8		2.3×	18	1.7	19	"	"	"	"
25%	1.0×	4	6		2.0×	16	1.9	19	"	"	"	"
20nm Planar												
Scenario	**Single-Task (Minimize Latency)**				**Balanced (Fixed Latency)**				**Many-Task (Unconstrained Latency)**			
	Gain	**FoM**	**Cores /Task**	**Cores Total**	**Gain**	**FoM**	**Cores /Task**	**Cores Total**	**Gain**	**FoM**	**Cores /Task**	**Cores Total**
% Serial	**Area Unconstrained**											
2%	1.0×	4	3		2.7×	11	13	83	4.8×	20	1	1,600
5%	1.0×	4	3		2.1×	9	6.8	35	"	"	"	"
10%	1.0×	4	2		1.7×	7	4.2	18	"	"	"	"
25%	1.0×	3	2		1.3×	5	2.0	5.9	"	"	"	"
% Serial	**Area Constrained (4 Cores Max)**											
2%	1.0×	4	3		1.3×	5	1.3	4	1.3×	5	1	4
5%	1.0×	4	3		1.2×	5	1.3	4	"	"	"	"
10%	1.0×	4	2		1.2×	5	1.3	4	"	"	"	"
25%	1.0×	3	2		1.1×	5	1.4	4	"	"	"	"

Table 2: Summary of figure-of-merit improvement gains from voltage scaling in 7nm FinFET and 20nm planar. Relative figure-of-merit and number of cores per task and total are also listed to compare across scenario or technology.

A key distinction to prior work is that near-threshold is proposed under normal processor load, not just during periods of idleness. Multicore architectures, used to parallize workloads, were also included as parts of these works. Azizi et al. [13] shows that voltage scaling is an effective technique for trading off performance and power, and that a large energy-performance design space can be encompassed using a small core, large core, and voltage scaling. Pinckney et al. [5] provided a methodical definition of near-threshold by defining it as the point where energy is minimized subject to a fixed latency constraint, and examined it across six planar technology nodes (180nm to 32nm). Circuit challenges and solutions in near-threshold were examined by Kaul et al. [14].

This paper differs from prior work primarily by examining differences between FinFET and planar. Compared to prior NT studies, such as [5], we analyze across technologies using a set of models that have been consistently tuned with similar transistor threshold voltage types. We also combine the impact of wire loading, and mismatch variation, directly into the energy optimization, both of which are especially significant in recent technologies. Finally, this work proposes three definitions for voltage scaling scenarios, and includes area constraints, which [5] does not address.

4. CONCLUSIONS

Power and performance improvements in process technology has slowed and systems now, more than ever, need to be co-designed with circuit techniques, such as voltage scaling. Three specific voltage scaling scenarios are examined: (1) single-task system where latency is minimized; (2) many-task system where latency is fixed to that of nominal supply voltage; and (3) many-task system when latency is completely unconstrained. For each of these scenarios, we estimated efficiency gains with differing amounts of parallelism and area budgets. By leveraging voltage scaling we are able to achieve significantly higher figure-of-merit, especially for workloads with low percent serial coefficients running on a many-task system. However, voltage scaling of single-task workloads also achieve gains, which was previously not possible in planar technologies past 40nm.

We show that FinFET offers important advantages over planar CMOS technologies, namely less degradation in circuit performance at low voltages, which translates into sizable energy efficiency gains. For a highly parallelizable workload, 7nm FinFET shows up to 8.6× higher energy efficiency at NT compared to nominal supply voltage (vs. 4.8× higher in 20nm planar). Combined with the ability to pack more cores within the same die area, FinFET offers architects unique opportunities to design futuristic system that leverage voltage scaling for achieving high-performance, even when latency minimization is critical.

It is important to note that many additional factors would impact the efficiency and performance of a complete system, namely interconnects, caches, memory interfaces, and peripherals. Memory and interconnects voltage scale differently than core logic, since they are generally dominated by wire loads and leakage power. Of course, they will also impact the scalability of tasks and further co-optimization is needed. Nevertheless, understanding voltage scaling benefits and limitations is essential in designing futuristic architectures in our post-Dennard scaling world.

5. REFERENCES

[1] H. Esmaeilzadeh, E. Blem, R. St. Amant, K. Sankaralingam, and D. Burger, "Dark silicon and the end of multicore scaling," in *Proceedings of the 38th Annual International Symposium on Computer Architecture*, ISCA '11, (New York, NY, USA), pp. 365–376, ACM, 2011.

[2] B. Zhai, R. Dreslinski, D. Blaauw, T. Mudge, and D. Sylvester, "Energy efficient near-threshold chip multi-processing," in *Low Power Electronics and Design (ISLPED), 2007 ACM/IEEE International Symposium on*, pp. 32–37, Aug 2007.

[3] R. Dreslinski, M. Wieckowski, D. Blaauw, D. Sylvester, and T. Mudge, "Near-Threshold Computing: Reclaiming Moore's Law through energy efficient integrated circuits," *Proceedings of the IEEE*, vol. 98, pp. 253–266, Feb 2010.

[4] L. Chang, D. Frank, R. Montoye, S. Koester, B. Ji, P. Coteus, R. Dennard, and W. Haensch, "Practical strategies for power-efficient computing technologies," *Proceedings of the IEEE*, vol. 98, pp. 215–236, Feb 2010.

[5] N. Pinckney, K. Sewell, R. Dreslinski, D. Fick, T. Mudge, D. Sylvester, and D. Blaauw, "Assessing the performance limits of parallelized near-threshold computing," in *Design Automation Conference (DAC), 2012 49th ACM/EDAC/IEEE*, pp. 1143–1148, June 2012.

[6] D. Fick, R. Dreslinski, B. Giridhar, G. Kim, S. Seo, M. Fojtik, S. Satpathy, Y. Lee, D. Kim, N. Liu, M. Wieckowski, G. Chen, T. Mudge, D. Blaauw, and D. Sylvester, "Centip3de: A cluster-based ntc architecture with 64 arm cortex-m3 cores in 3d stacked 130 nm cmos," *Solid-State Circuits, IEEE Journal of*, vol. 48, pp. 104–117, Jan 2013.

[7] M. Taylor, "Is dark silicon useful? harnessing the four horsemen of the coming dark silicon apocalypse," in *Design Automation Conference (DAC), 2012 49th ACM/EDAC/IEEE*, pp. 1131–1136, June 2012.

[8] S. C. Woo, M. Ohara, E. Torrie, J. P. Singh, and A. Gupta, "The splash-2 programs: Characterization and methodological considerations," in *Proceedings of the 22Nd Annual International Symposium on Computer Architecture*, ISCA '95, (New York, NY, USA), pp. 24–36, ACM, 1995.

[9] A. Chandrakasan, S. Sheng, and R. Brodersen, "Low-power CMOS digital design," *Solid-State Circuits, IEEE Journal of*, vol. 27, pp. 473–484, Apr 1992.

[10] A. Wang and A. Chandrakasan, "A 180mV FFT processor using subthreshold circuit techniques," in *Solid-State Circuits Conference, 2004. Digest of Technical Papers. ISSCC. 2004 IEEE International*, pp. 292–529 Vol.1, Feb 2004.

[11] R. Dreslinski, M. Wieckowski, D. S. Blaauw, and T. Mudge, "Near threshold computing: Overcoming performance degradation from aggressive voltage scaling," in *Proc. Workshop Energy-Efficient Design*, pp. 44–49, 2009.

[12] R. Dreslinski, B. Zhai, T. Mudge, D. Blaauw, and D. Sylvester, "An energy efficient parallel architecture using near threshold operation," in *Parallel Architecture and Compilation Techniques, 2007. PACT 2007. 16th International Conference on*, pp. 175–188, Sept 2007.

[13] O. Azizi, A. Mahesri, B. C. Lee, S. J. Patel, and M. Horowitz, "Energy-performance tradeoffs in processor architecture and circuit design: A marginal cost analysis," in *Proceedings of the 37th Annual International Symposium on Computer Architecture*, ISCA '10, (New York, NY, USA), pp. 26–36, ACM, 2010.

[14] H. Kaul, M. Anders, S. Hsu, A. Agarwal, R. Krishnamurthy, and S. Borkar, "Near-threshold voltage (ntv) design - opportunities and challenges," in *Design Automation Conference (DAC), 2012 49th ACM/EDAC/IEEE*, pp. 1149–1154, June 2012.