Can Beyond-CMOS Devices Illuminate Dark Silicon?

Robert Perricone, X. Sharon Hu, Joseph Nahas, and Michael Niemier
Department of Computer Science and Engineering, University of Notre Dame
Notre Dame, IN 46556, USA, Email: {rperrico,shu,jnahas,mniemier}@nd.edu

***Abstract*—Throughout the last decade, the microprocessor industry has been struggling to preserve the benefits of Moore's Law scaling. The persistent scaling of CMOS technology no longer yields exponential performance gains due in part to the growth of dark silicon. With each subsequent technology node generation, power constraints resulting from factors such as sub-threshold leakage currents are projected to further limit the number of transistors that can be simultaneously powered. To overcome the limits of CMOS devices, researchers are working to develop "beyond-CMOS" device technologies. To determine the most promising beyond-CMOS devices, it is necessary to benchmark them against CMOS. In this paper, we present the design and validation of an analytical benchmarking model that evaluates CMOS and beyond-CMOS devices at the architectural-level. Our model is built from the device to the architectural/application-level. Our target architecture is a symmetric multi-core processor executing highly parallel applications (i.e., PARSEC). As a case study, we select one class of promising beyond-CMOS devices, tunneling field-effect transistors, to evaluate against CMOS.**

I. Introduction

From its inception, Moore's Law has been the fundamental economic driver of the microprocessor industry [1]. According to Moore's Law, the number of (CMOS) transistors on-chip doubles with each technology generation. Moreover, microprocessor performance also doubles as a result of transistor scaling (i.e., from Dennard scaling [2] and Pollack's rule [3]). Unfortunately, within the last decade, transistor device limitations have abated the performance scaling trend. Increased sub-threshold leakage current and decreased supply voltage scaling have diminished the historically exponential growth in processor performance [4]. Consequently, microprocessor engineers have adopted multi-core architectures in an attempt to preserve processor performance scaling by leveraging parallel processing [3].

While multi-core processors have succeeded in delivering modest (approximately linear) performance gains, projections indicate they will encounter a power wall as transistors continue to scale [5]. Specifically, [5] suggests that as the number of transistors continues to double, the power density will approach the physical and economical limits of a chip's thermal design power (TDP)—thus necessitating the selective activation of on-chip devices. This phenomenon is colloquially referred to as "dark silicon" [5], and it has inspired a wide range of solutions [6], such as "beyond-CMOS" devices [7], "dim silicon" cores [8], near-threshold computing [9], reliability-efficient computing [10], customized accelerators [11], and even combinations of the above to produce heterogeneous architectures [12], [13]. While each of these ideas offers a novel approach to overcoming the "dark silicon" phenomenon, beyond-CMOS devices are the obvious, yet most unpredictable choice for overcoming CMOS limitations[6].

Benchmarking beyond-CMOS devices is an important step for identifying the most promising devices. To ensure accurate benchmarking, it is necessary for all candidate devices to be evaluated using a uniform methodology. In fact, this was one of the goals of the recent Nanoelectronics Research Initiative (NRI) led benchmarking effort [14], [15]. The NRI benchmarking methodology in [14] provided the foundation for a simple, uniform, and transparent approach to benchmarking various beyond-CMOS devices. However, this benchmarking approach only investigated the devices at the circuit-level (e.g., Inverter fanout-of-4, 2-input NAND gate, and 32-bit ripple carry adder). Moreover, this level of benchmarking does not provide direct insight as to how devices will fare at the architectural-level, especially when considering multi-core architectures.

In this paper, we present a methodology to benchmark beyond-CMOS devices at the architectural-level. Our benchmarking model, which we will refer to as the *new Dark Silicon* (nDS) model, is purely analytical and is based on two existing analytical benchmarking methodologies. The first is the "Beyond-CMOS Benchmarking" version 3 (BCBv3) methodology [16], which provides circuit-level benchmarking of beyond-CMOS devices. The second methodology is an architectural-level approach (referred to as the "Dark Silicon model" or DS model) that explores the limits of multi-core scaling within a fixed TDP and area budget for CMOS technology [5]. By combining these two methodologies and introducing three modifications, we are able to benchmark beyond-CMOS devices at the architectural-level for multi-core processors executing parallel workloads (i.e., PARSEC benchmarks [17]).

To illustrate our benchmarking methodology, we first present, in Sec. II, an overview of the two benchmarking methodologies that form the foundation of nDS, as well as the parameters of the beyond-CMOS devices to be benchmarked. Next, in Sec. III, we discuss how the two models are combined to form nDS as well as the necessary modifications that allows us to properly benchmark beyond-CMOS devices at the architectural-level. In Sec. IV, we validate our modifications against empirical processor data and show that our approach results in significant accuracy improvements when compared to projections based on existing models. In Sec. V, we present a case study that evaluates one class of promising beyond-CMOS devices, tunneling field-effect transistors (TFETs). We conclude and summarize our work in Sec. VI.

II. Benchmarking Models and Devices Overview

In this section, we provide an overview of the devices and the benchmarking models that form the foundation of nDS. First, we briefly introduce the TFET devices and their parameters that will be used in our case study. Second, we summarize the BCBv3 model [16] and provide a sample output of this approach using our selected TFET devices. Lastly, we summarize the DS model introduced in [5].

A. Overview of TFET Devices

From the class of beyond-CMOS devices, we have selected four TFETs to be used in our work. TFETs are a subset of

978-3-9815370-6-2/DATE16/ ©2016 EDAA

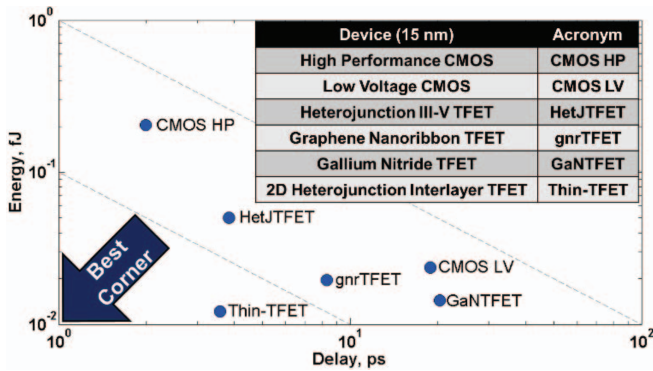

Fig. 1. BCBv3 sample output showing dynamic energy vs. delay of an inverter fanout-of-4 (INVFO4) for different device technologies. Device input parameters are given in Table I.

beyond-CMOS devices, which simulations suggest are competitive with CMOS—especially for low-power applications [18]. TFETs are "steep slope" devices that could offer a sub-threshold slope below the CMOS intrinsic limit of 60 mV/decade. The result is increased on-current and reduced leakage current at low supply voltage as compared to CMOS.

Device Parameters			
Device	V_{DD} (V)	I_{On} (μA/μm)	$C_{G,avg}$ (fF/μm)
CMOS HP [19]	0.73	1805	0.41
CMOS LV [20]	0.3	53	0.06
HetJTFET [18]	0.4	417	0.21
gnrTFET [21]	0.25	120	0.21
GaNTFET [22]	0.2	47	0.34
Thin-TFET [23]	0.2	263	0.18
Geometric Parameters [16]			
Half-pitch (nm)	EOT (nm)	Gate Len. (nm)	Gate Wid. (nm)
15	1.08	12.8	60

TABLE I. Benchmarking parameters for CMOS and TFET devices. Device parameters are taken from the indicated references in the leftmost column.

The TFET devices we selected represent four of the most promising devices based on the results of the Nikonov and Young BCB methodology [16]. Specifically, we have selected the heterojunction III-V TFET (HetJTFET) [18], graphene nanoribbon TFET (gnrTFET) [21], gallium nitride TFET (GaNTFET) [22], and the two-dimensional heterojunction interlayer TFET (Thin-TFET) [23]. To benchmark these devices at the circuit-level using the BCB methodology (discussed further in Sec. II-B), only three parameters are necessary for each device: supply voltage ($V_{DD} = |V_{DS}|$), on-current ($I_{On} = I_{Dsat}$), and average gate capacitance ($C_{G,avg}$).

In Table I, we summarize the CMOS and TFET device parameters that are used in the BCB model and ultimately our model. Note that most of the device parameters are consistent with the parameters used in the BCB methodology [14], [16] except for the Thin-TFET, which has been updated to represent its present parameters [23]. Also, the geometric parameters—metal half-pitch, equivalent oxide thickness (EOT), gate length, and gate width—listed in the table are applied to all devices.

B. BCB Summary

The first beyond-CMOS benchmarking effort by Bernstein *et al.* (i.e., BCBv1) attempted to assemble some device researchers to collectively benchmark their devices. However, due to self-reporting, no unifying assumptions were made. To overcome this problem, Nikonov and Young developed an analytical methodology built on BCBv1, i.e., BCBv2,

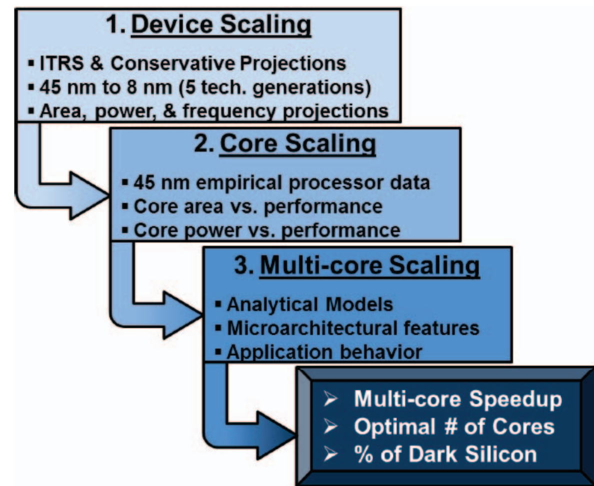

Fig. 2. Overview of the dark silicon model's hierarchy from [5].

by which all device parameters would be derived from the same uniform assumptions. Furthermore, they provided detailed computations used to benchmark each device at the logic circuit-level. These computations yielded estimates for performance, area, switching delay, and energy [14]. BCBv3 was recently released to reflect an improved understanding of the beyond-CMOS devices and their operations. In addition to updated device parameters, BCBv3 also included new logic circuit configurations (e.g., sequential logic) and computation of standby power for each device [16].

The Nikonov and Young BCB methodology uses the basic device parameters given in Table I to analytically calculate the delay, energy, and area of a given circuit, such as an inverter fanout-of-4 (INVFO4), 2-input NAND gate (NAND2), and 32-bit ripple carry adder (32-bit RCA). Fig. 1 is a sample output of BCBv3 that shows dynamic switching energy vs. delay of CMOS and TFET devices for an INVFO4. From the figure, the high performance CMOS (CMOS HP) has the fastest switching speed, but also uses the most energy. The Thin-TFET not only consumes the least amount of energy, but also represent the best device in terms of energy-delay product (EDP).

C. Dark Silicon Model Summary

In [5], Esmaeilzadeh *et al.* explored multi-core scaling limits for five technology node generations from 45 nm to 8 nm in an effort to project how core scaling might impact performance as technology is scaled. The authors developed an analytical model to compute potential performance gains on parallel applications (i.e., PARSEC [17]) for each technology node generation. Ultimately, their model predicts that future multi-core processors will be unable to utilize all available cores due to power constraints. Their model projected that at 22 nm, 21% of the chip will need to be powered off (i.e., "dark"), and the percentage of dark silicon will further increase to 50% at 8 nm.

This Dark Silicon (DS) model is built from the device-level to the architectural-level in three parts (see Fig. 2). First, in the **device scaling model**, the authors consider device scaling trends such as area, frequency, and power from 45 nm to 8 nm based on optimistic and conservative projection schemes[1]. The optimistic projections were based on the updated 2011 ITRS report [19] while the conservative projections were based

[1]45 nm was the current technology node when this work was completed.

on data from S. Borkar of Intel [24]. For each technology node, scaling factors for frequency and power are derived by normalizing their projections against empirical 45 nm CMOS data. It is important to note here that frequency scaling factors derived from ITRS projections are based on INVFO4 simulations from the *Model for the Assessment of CMOS Technologies and Roadmaps* (MASTAR) for each technology node [19]. In this sense, the DS model uses INVFO4 delay to determine the clock frequency of a multi-core processor.

The second part of the DS model is the **core scaling model**. This model provides projections for the maximum performance that a single core can achieve for a given area. Furthermore, it also projects the core power for the selected core performance. This model is derived by creating two scatter plots—core area vs. performance and core power vs. performance—using empirical 45 nm processor data. For both plots, SPECmark is the measure of performance based on single-threaded integer SPEC benchmarks [25]. The authors then use curve fitting on the two charts to plot the Pareto-optimal frontier for the 45 nm processors. Next, for each technology generation, the scaling factors derived in the device model are used to scale the area vs. performance and power vs. performance Pareto frontiers. The result is a set of processor core projections for each technology node.

The third and final part of the DS model is the **multi-core scaling model**. Here the authors consider two multi-core configurations—based on CPUs and GPUs—and four topologies for each configuration: symmetric, asymmetric, dynamic, and composed multi-core. For a selected multi-core configuration and topology, the DS model uses the core-scaling model to analytically compute the best possible speedup, optimal number of cores, and fraction of dark silicon *for each* PARSEC benchmark given a chip area and TDP budget. These computations consider microarchitectural features and application behavior when computing the per benchmark output. The underlying 45 nm microacrchitecture in the DS model is the Intel Nehalem (Core i7) processor. The default parameters for TDP, processor area, cache size, CPI, etc. are derived from this processor.

The DS model takes several inputs including the projection scheme, multi-core organization/topology, TDP budget, TDP percentage of leakage power (the default is 20%), and a specified PARSEC benchmark. As the DS model runs for each technology node, the model discretizes the derived area/performance Pareto frontier and examines each processor design point along the curve. For each design point, a multi-core processor is constructed based on the user-specified organization and topology. As the model iterates, additional cores are added and a new speedup, optimal number of cores, etc. are computed. Power/performance Pareto frontier is used to determine the per core power for each of the selected design points. Once the area or TDP budget is encountered, the model terminates and reports its findings. The results include the optimal core configuration and speedup, which are likely to differ from benchmark to benchmark.

III. nDS Methodology and Modifications

In the DS model, frequency and power scaling factors are derived from CMOS INVFO4 simulations. As the BCB model produces INVFO4 energy vs. delay output for beyond-CMOS devices (see Fig. 1), we can similarly derive frequency and power scaling factors for beyond-CMOS devices (normalized to 45 nm CMOS)[2]. The derived scaling factors represent how the beyond-CMOS devices compare to 45 nm CMOS from a power and performance perspective. In this sense, the scaling factors represent a technology change from CMOS to the beyond-CMOS technology. Ultimately, we can perform an apples-to-apples comparison between the beyond-CMOS and CMOS devices at the 15 nm technology node.

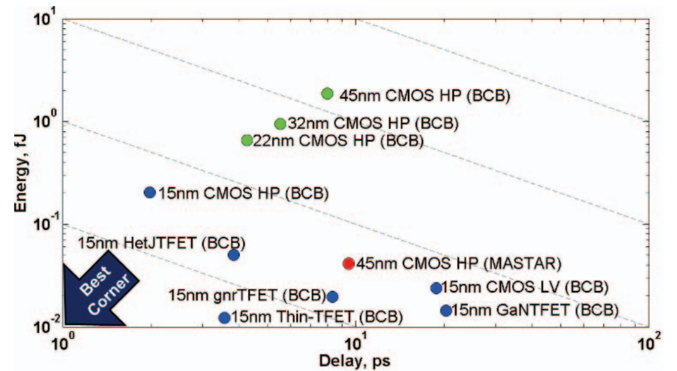

Fig. 3. INVFO4 energy vs. delay for various technology nodes and devices. BCB or MASTAR appears after each device to indicate the computation model for the data point.

However, simply combining the BCB model with the DS model as outlined above can introduce significant inaccuracy that leads to meaningless results. This is due to the fact that some of the assumptions made in the DS model are no longer valid for beyond-CMOS devices. Therefore, it is necessary to modify each level of the DS model to support beyond-CMOS devices. In the following subsection, we explain our modifications and why they are necessary.

A. Modifications to DS Model

To accurately benchmark beyond-CMOS devices at the architectural-level using the DS model, we introduce three modifications—one to each scaling model. First, in the device scaling model, the CMOS frequency scaling factors were derived by normalizing MASTAR INVFO4 simulations to 45 nm (2010 column of 2010 ITRS [19]) for each technology node. The 45 nm power input was analytically computed as $P_{dynamic} = \alpha C V^2 f$ (where $\alpha = 1$) using the same ITRS 2010 report data. Since the 45 nm frequency and power components are based on the MASTAR INVFO4 simulation, simply normalizing our 15 nm BCB-computed frequency and power values by the 45 nm MASTAR-computed ones would introduce inconsistent scaling factors. This problem is illustrated vividly in the INVFO4 data in Fig. 3. Here, in Fig. 3, we show the BCB INFO4 computation and compare it to the MASTAR computation for 45 nm CMOS. From the figure, one can see that there is a large discrepancy between the two computed INVFO4 values (switching energy data points differ by $\sim 50\times$.

To produce consistently normalized values, we applied the BCBv3 model to compute the 45 nm CMOS parameters. We utilized the same 45 nm inputs as used in the MASTAR simulation from the 2010 column of the 2010 ITRS report [19] while adjusting the BCB model for the 45 nm technology node (default is 15 nm metal half pitch). With the newly

[2]The energy component of the INVFO4 plot represents dynamic energy and that average dynamic power is computed for each device as the ratio of its energy to delay. Moreover, the frequency is the inverse of the delay.

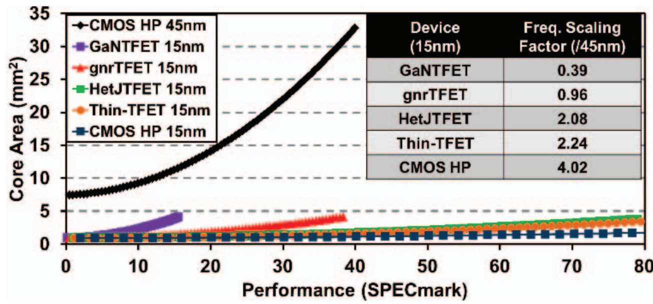

Fig. 4. Erroneous compression of the 45 nm area vs. performance Pareto frontier occurs when a device has a frequency scaling factor of less than one (assuming beyond-CMOS devices and CMOS are equally sized).

computed BCB INVFO4 45 nm data point, we normalize the 15 nm beyond-CMOS and CMOS devices to produce consistent frequency and power scaling factors. These factors can be seen in the inset of Fig. 4.

Our second modification is implemented within the core scaling model. As previously discussed in Sec. II-C, the core scaling model projects the Pareto optimal core area vs. performance and core power vs. performance for each technology node beyond 45 nm. This is achieved by scaling the respective empirical 45 nm Pareto frontiers using the derived scaling factors for each technology node. Note that the area scaling factor is computed from Moore's law using $Area_{scaling} = 2^N$ where N is the number of technology node generations from 45 nm. The performance is scaled by the computed frequency scaling factor in both plots, which is justified in [5] through Pollack's rule (i.e., speedup from microarchitectural improvements using the additional transistors for each subsequent technology node generation).

When normalized to 45 nm, the frequency scaling factor of some beyond-CMOS devices is less than one—that is the device is slower than 45 nm CMOS. Scaling the 45 nm area/performance Pareto frontier by these factors leads to erroneous results. For example, if we assume that CMOS and beyond-CMOS devices are similarly sized and scale equally, then the area component of the area/performance Pareto frontier does not differ between CMOS and beyond-CMOS devices at 15 nm. However, when performance is scaled by a frequency scaling factor of less than one, the area/performance Pareto frontier becomes more compressed compared to the 45 nm curve (e.g., see the GaNTFET data in Fig. 4). The result of this compression violates the premise of the area/performance Pareto frontier since performance is not increasing from 45 nm to 15 nm (i.e., Pollack's Rule). In fact, Pollack's rule cannot be correctly applied in this sense when an underlying technology change is made.

To address the area scaling problem, we start by assuming that the area/performance Pareto frontiers of all 15 nm beyond-CMOS devices are identical to the projected 15 nm CMOS area/performance Pareto frontier. This initial assumption guarantees that Pollack's rule for the model is preserved. More specifically, the area/performance Pareto frontier captures the microarchitectural performance improvements (e.g., CPI) gained from additional devices added on-chip—as performance increases, core area (i.e., complexity) increases too. Therefore, performance improvements from microarchitectural changes are agnostic to their underlying technology. Lastly, to address the difference in size between beyond-CMOS and CMOS devices, we add a new area scaling factor that accounts

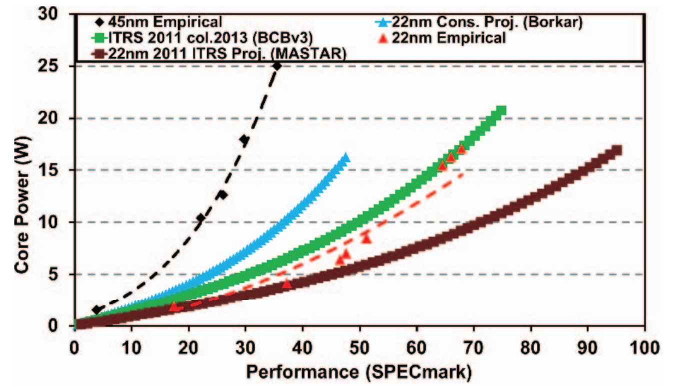

Fig. 5. Comparison of core power vs. performance projections for 22 nm technology for different projection methodologies. Results show our approach (green curve) is closer to the empirical data (red curve) than the projection schemes used in [5].

for the relative size difference between the beyond-CMOS and CMOS devices. For example, TFETs are approximately $1.25\times$ the size of CMOS [14]. Therefore, the total CMOS chip area budget is reduced by a factor of 1.25.

Our third and final modification is implemented within the multi-core scaling model. When run out-of-the-box, the DS model removes 20% of the total TDP to account for leakage power across all CMOS technology node generations. To account for differences in leakage power for different technologies, we modified the multi-core scaling model to consider per device leakage power. To compute the percent leakage power of each device, we leverage the BCB model, which also computes the dynamic and static power of a 32-bit RCA for each beyond-CMOS device. The percentage of leakage power is simply the ratio of static power to total power. Our justification for using the 32-bit RCA is twofold. First, in the DS model, the cores that compose a multi-core architecture are primarily pipelined logic[3]. Second, we use the 32-bit RCA to compute the leakage power not only for the 15 nm beyond-CMOS devices, but also 15 nm CMOS to ensure consistency.

IV. Validation of nDS

Our nDS model addresses the shortcomings of the original DS model when applied to beyond-CMOS devices. To ensure nDS still works well for advanced CMOS technology nodes, we validated the model using empirical 22 nm processor data from Intel's Ivy Bridge and Haswell processors. Specifically, SPECmark scores were collected from the available SPEC CPU2006 data along with each processor's thermal TDP from data sheets to create an empirical power vs. performance plot at the 22 nm technology node. Following the approach by Esmaeilzadeh *et al.* [5] for 45 nm processor data, we used power regression to fit a curve to the 22 nm data in a Pareto-optimal sense. In Fig. 5, the two dotted-line curves represent empirical power/performance Pareto frontiers. The black dotted line with diamond markers represents the original 45 nm empirical Pareto frontier from [5] while the red dotted line with triangle markers represents our newly derived empirical 22 nm Pareto frontier.

Using the 22 nm empirical Pareto frontier, we first wanted to assess the accuracy of the two projection schemes—ITRS and conservative—from [5]. As predicted in [5], the optimistic ITRS projection overestimates the 22 nm technology node as

[3] Shared caches are removed from the input area and power budgets.

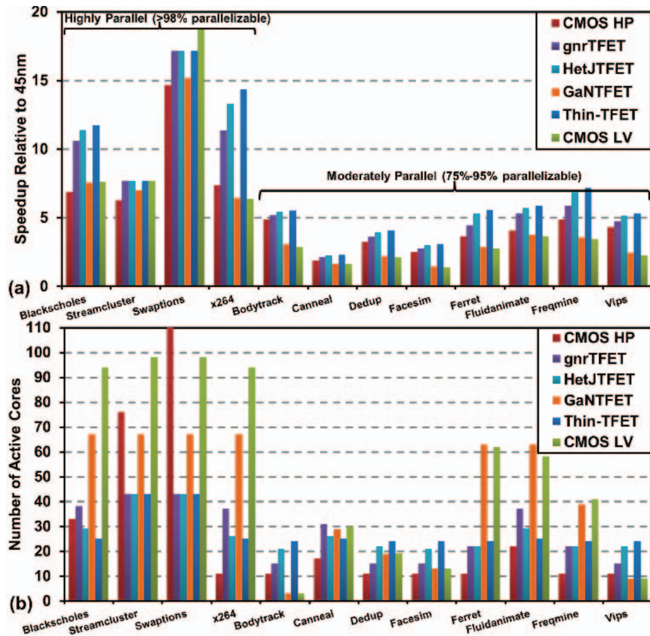

Fig. 6. nDS benchmarking results for CMOS and TFET devices for the high (125 W) TDP case. Results show (a) speedup relative to 45 nm vs. PARSEC benchmark suite. Note these results represent the best possible speedup for each benchmark and (b) number of active cores vs. the PARSEC benchmark suite. Note these results represent the maximum number of active cores to achieve the speedups shown in (a).

shown by the brown curve with square markers. Conversely, the conservative projections from S. Borkar underestimate the 22 nm technology node as shown by the light blue curve with triangle markers. Interestingly, the 22 nm empirical curve appears nearly in the middle of the two projection schemes.

To assess the accuracy of our approach, we first used the BCB model to plot 22 nm CMOS energy vs. delay for an INVFO4 (see Fig. 1). Next, as discussed in Sec. III-A, we computed the frequency and power scaling factors relative to 45 nm by normalizing the BCB-computed 22 nm CMOS against the BCB-computed 45 nm CMOS. Using the BCB-computed 22 nm CMOS scaling factors, we scaled the 45 nm power/performance Pareto frontier, which is shown in Fig. 5 as the green curve with square markers. As one can see, the BCB-based projection produces a nearly identical Pareto frontier compared to the empirical 22 nm. This result is evidence of the accuracy of BCB compared to the MASTAR circuit simulator used by ITRS. For example, at a core power of 15 W, the empirical 22 nm data has a SPECmark of 69. However, the 22 nm MASTAR simulation yields a SPECmark of 90 versus our projection of 63.

V. Benchmarking Results and Discussion

Given the validated nDS model, we now present a case study involving the benchmarking of the TFET devices introduced in Sec. II-A. We first discuss the nDS benchmarking results for the high TDP (125 W) case. Fig. 6(a) represents a typical output of nDS that shows the optimal speedup over 45 nm CMOS per each PARSEC benchmark for each device technology. As indicated in the figure, the four leftmost PARSEC benchmarks are highly parallel ($> 98\%$ parallelizable). These benchmarks benefit the most from more cores placed on-chip to exploit their parallelism. This can be seen by examining the total number of active cores per each benchmark as shown in Fig. 6(b). As expected for highly parallel applications, low power devices achieve higher speedups compared to CMOS HP by utilizing many low power cores.

However, the number of active cores (Fig. 6(b)) is not a reliable indicator of the devices with the best speedups—other factors, such as application behavior, must also be considered. For example, consider CMOS HP versus the GaNTFET and CMOS LV for the x264 benchmark in Figs. 6(a) and 6(b). The CMOS HP has slightly better speedup than the GaNTFET and CMOS LV despite using only 11 cores compared to the 67 cores and 94 cores used by the GaNTFET and CMOS LV, respectively. While x264 is highly parallel, it also has a relatively high L1 and L2 cache miss rate for Intel Nehalem-sized caches (i.e., L1 is 64 KB and L2 is 2 MB) [17], [26]. Consequently, having more cores increases the shared L2 cache miss rate, which reduces the fraction of time that all cores can be doing work (i.e., core utilization is low). To summarize, having more cores enables greater performance gains through exploiting parallelism; however, for applications with higher cache miss rates—especially in the L2—the ultimate speedups are reduced due to low core utilization.

We next examine the results for each device across the entire PARSEC benchmark suite. For each device, the left bars in each plot of Fig. 7 respectively represent the geometric means across PARSEC for (a) speedup relative to 45 nm, (b) number of active cores, and (c) percent of total chip area that is dark silicon for the high TDP case. As one may expect from examining Fig. 1, the Thin-TFET indeed achieves the highest average speedup with the lowest average percentage of dark silicon. However, within this optimistic scenario of selecting the best multi-core configuration for each benchmark, the Thin-TFET falls well short of Moore's Law performance scaling trends. In fact, the Thin-TFET is less than $2\times$ better than both CMOS HP and LV. This result is not readily evident from the circuit-level benchmarking results as shown in Fig. 1. Furthermore, this suggests that moving from CMOS to present TFET device technologies may not be very beneficial—at least for high TDP processors. However, compared to CMOS HP, the Thin-TFET is capable of powering 99% of its cores whereas over one-third of the CMOS HP chip must be powered off.

Steep slope devices, such as TFETs, are likely to deliver more benefits in power constrained environments. We now re-examine the previously presented results, but for multi-core processors with low TDP (5 W). For each device, the right bars in each plot of Fig. 7 represent the low TDP geometric mean results when the optimal core configuration is selected for each benchmark. We first note that speedup of 15 nm CMOS HP is actually worse than 45 nm CMOS in this case. Leakage power is a concern for CMOS and its effects are especially noticeable when TDP is constrained. On average CMOS HP can only utilize approximately 6 cores while the remaining 93% of the chip is dark silicon. Moreover, CMOS LV is not much better than CMOS HP with $1.5\times$ average speedup and 60% dark silicon.

In this power-constrained scenario, the Thin-TFET again achieves the best overall average speedup. Compared to CMOS HP, the Thin-TFET is more than $5\times$ faster on average with only 30% dark silicon. Compared to CMOS LV, the Thin-TFET is approximately $2.5\times$ faster on average while utilizing 30% more of its cores. Collectively, no TFET is significantly faster than CMOS, but the amount of dark silicon in both the

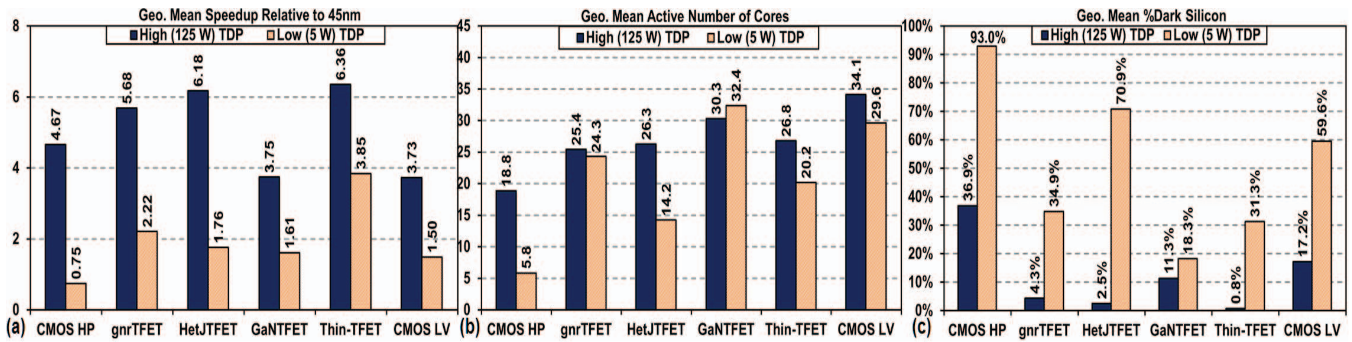

Fig. 7. For each device, the three sub-figures show the geometric means of (a) speedup over 45 nm CMOS, (b) number of active cores, and (c) percentage of dark silicon across all PARSEC benchmarks for a high (125 W) and low (5 W) TDPs.

high and low TDP cases indicates that the majority of the processor's cores can be powered on.

VI. Conclusion

We presented nDS: an analytical architectural-level benchmarking model for beyond-CMOS devices. nDS achieves this level of benchmarking by leveraging and modifying two existing benchmarking models—BCB and the DS model. We validated the accuracy of nDS by demonstrating close correlation between its 22 nm CMOS processor projections and empirical 22 nm CMOS processor data.

To demonstrate the capabilities of nDS, we benchmarked one class of promising beyond-CMOS devices: TFETs. We sought to determine if these devices could sustain Moore's law and/or yield a significant advantage over CMOS. Our results demonstrate that for both high (125 W) and low (5 W) TDP processors, none of the examined beyond-CMOS device technologies currently can sustain the performance scaling trend of Moore's Law. Furthermore, for high TDP processors, the best speedup over CMOS is achieved by the Thin-TFET, but is less than $2\times$. For low TDP processors, the Thin-TFET achieves an average speedup of over 5X compared to CMOS HP, but only $2.5\times$ compared to CMOS LV. Nevertheless, for both high and low TDPs, the TFET devices have less dark silicon compared to CMOS. For the high TDP case, the Thin-TFET has less than 1% dark silicon compared to over 36% and 17% for the CMOS HP and LV, respectively. For the low TDP case, the GaNTFET achieves the lowest dark silicon with approximately 11% compared to the 93% and 60% of the CMOS HP and LV, respectively.

For future work, we plan to utilize nDS to investigate the use of beyond-CMOS devices in emerging architectural innovations such as accelerators. Additionally, we plan to investigate low-power applications to quantify the difference between steep slope and CMOS devices in this space.

Acknowledgment

This work was supported in part by the Center for Low Energy Systems Technology (LEAST), one of the six SRC STARnet Centers, sponsored by MARCO and DARPA.

References

[1] G. Moore, "Cramming more components onto integrated circuits, electronics,(38) 8," 1965.

[2] R. Dennard *et al.*, "Design of ion-implanted mosfet's with very small physical dimensions," *IEEE Journal of Solid-State Circuits*, vol. 9, no. 5, pp. 256–268, Oct. 1974.

[3] S. Borkar, "Thousand core chips: A technology perspective," in *Proceedings of the Design Automation Conference*, 2007, pp. 746–749.

[4] S. Borkar, "Design challenges of technology scaling," *Micro, IEEE*, vol. 19, no. 4, pp. 23–29, Jul. 1999.

[5] H. Esmaeilzadeh *et al.*, "Dark silicon and the end of multicore scaling," in *International Symposium on Computer Architecture*, 2011.

[6] M. B. Taylor, "Is dark silicon useful?: Harnessing the four horsemen of the coming dark silicon apocalypse," in *Proceedings of the Design Automation Conference*, 2012, pp. 1131–1136.

[7] K. Bernstein *et al.*, "Device and architecture outlook for beyond cmos switches," *Proceedings of the IEEE*, vol. 98, no. 12, Dec. 2010.

[8] W. Huang *et al.*, "Scaling with design constraints: Predicting the future of big chips," *IEEE Micro*, vol. 31, no. 4, pp. 16–29, July 2011.

[9] A. Pahlevan *et al.*, "Towards near-threshold server processors," in *Design, Automation and Test in Europe Conference Exhibition*, 2016.

[10] J. Henkel *et al.*, "Towards performance and reliability-efficient computing in the dark silicon era," in *Design, Automation and Test in Europe Conference Exhibition*, March 2016.

[11] G. Venkatesh *et al.*, "Conservation cores: Reducing the energy of mature computations," in *Proceedings of Architectural Support for Programming Languages and Operating Systems*, 2010, pp. 205–218.

[12] L. Wang and K. Skadron, "Implications of the power wall: Dim cores and reconfigurable logic," *IEEE Micro*, vol. 33, no. 5, Sept. 2013.

[13] K. Swaminathan *et al.*, "Steep-slope devices: From dark to dim silicon," *IEEE Micro*, vol. 33, no. 5, pp. 50–59, Sept. 2013.

[14] D. Nikonov and I. Young, "Overview of beyond-cmos devices and a uniform methodology for their benchmarking," *Proceedings of the IEEE*, vol. 101, no. 12, pp. 2498–2533, Dec. 2013.

[15] J. J. Welser *et al.*, "The quest for the next information processing technology," *J. of Nanoparticle Research*, vol. 10, no. 1, pp. 1–10, 2008.

[16] D. Nikonov and I. Young, "Benchmarking of beyond-cmos exploratory devices for logic integrated circuits," *IEEE J. on Exploratory Solid-State Computational Devices and Circuits*, vol. 1, pp. 3–11, Dec. 2015.

[17] C. Bienia *et al.*, "The parsec benchmark suite: Characterization and architectural implications," in *International Conference on Parallel Architectures and Compilation Techniques*, 2008, pp. 72–81.

[18] A. C. Seabaugh and Q. Zhang, "Low-voltage tunnel transistors for beyond cmos logic," *Proceedings of the IEEE*, vol. 98, no. 12, pp. 2095–2110, Dec. 2010.

[19] Itrs report. [Online]. Available: http://www.itrs.net

[20] R. Kim *et al.*, "Source/drain doping effects and performance analysis of ballistic iii-v n-mosfets," *IEEE Journal of the Electron Devices Society*, vol. 3, no. 1, pp. 37–43, Jan. 2015.

[21] M. Luisier and G. Klimeck, "Performance analysis of statistical samples of graphene nanoribbon tunneling transistors with line edge roughness," *Applied Physics Letters*, vol. 94, no. 22, 2009.

[22] W. Li *et al.*, "Polarization-engineered iii-nitride heterojunction tunnel field-effect transistors," *IEEE Journal on Exploratory Solid-State Computational Devices and Circuits*, vol. 1, pp. 28–34, Dec. 2015.

[23] M. Li *et al.*, "Two-dimensional heterojunction interlayer tunneling field effect transistors (thin-tfets)," *IEEE Journal of the Electron Devices Society*, vol. 3, no. 3, pp. 200–207, May 2015.

[24] S. Borkar, "The exascale challenge," Keynote at International Symposium on VLSI Design, Automation and TEST (VLSI-DAT), 2010.

[25] Spec. [Online]. Available: http://www.spec.org

[26] M. Bhadauria *et al.*, "Understanding parsec performance on contemporary cmps," in *IEEE International Symposium on Workload Characterization*, 2009, pp. 98–107.