PROCEED: A Pareto Optimization-Based Circuit-Level Evaluator for Emerging Devices

Shaodi Wang, Andrew Pan, *Student Member, IEEE*, Chi On Chui, *Senior Member, IEEE*,
and Puneet Gupta, *Member, IEEE*

***Abstract*— Evaluation of novel devices in the context of circuits is crucial to identifying and maximizing their value. We propose a new framework, Pareto optimization-based circuit-level evaluator for emerging device (PROCEED), that uses comprehensive performance, power, and area metrics for accurate device-circuit coevaluation through optimization of digital circuit benchmarks. The PROCEED assesses technology suitability over a wide operating region (megahertz to gigahertz) by leveraging available circuit knobs (threshold voltage assignment, power management, sizing, and so on). It improves the benchmark accuracy by 3× to 115× compared with the existing methods while offering orders of magnitude improvements in runtime over full physical design implementation flows. To illustrate the PROCEED's capabilities, we deploy it to assess emerging technologies, including novel tunneling field-effect transistors, compared with conventional silicon CMOS. As a further illustration, we extend PROCEED to evaluate future heterogeneous integration of varied devices onto the same silicon substrate.**

***Index Terms*— Circuit-level device evaluation, Pareto optimization, silicon-on-insulator (SOI), simulation-based optimization, tunneling field-effect transistor (TFET).**

I. Introduction

AS TRADITIONAL silicon devices approach their fundamental scaling limits, it is important to explore additions or alternatives to CMOS. To do so, emerging devices must be assessed within the context of the circuits they might be used to build. Many technology-benchmarking methods have been proposed to meet this need [1]–[16]; as summarized in Table I, all these methods neglect a number of essential circuit features, any one of which can dramatically alter the results. Because of their variety and complexity, modern devices and circuit designs must be carefully chosen to complement each other before assessing viability; this requires a level of flexibility in the benchmarking process that has not existed until now.

Device-circuit assessments must consider several factors to draw realistic conclusions. First of all, any effective power and delay evaluation of modern circuits should cover several orders of magnitude, since their operating frequencies range from kilohertz to gigahertz. Second, chip area, ignored in all current evaluation methods, should be simultaneously considered because of its impact on manufacturing cost and interconnect length. Third, the crucial tuning knobs, such as logic gate sizing and supply voltage (V_{dd}) or threshold voltage (V_t) selection, must be optimized for proper use of a particular circuit. Fourth, since circuit performance depends critically on the device operating point, benchmarks should consider the full device current–voltage (I–V) characteristics rather than only simplified metrics such as saturation current (I_{ON}) or OFF-state leakage (I_{OFF}). Fifth, a given device may not be suitable for all circuit architectures because of variations in logic depth histogram (LDH) patterns, and logical or physical structure. Sixth, as technologies scale down, device variability due to ambient process fluctuations becomes ever more important and impacts circuit viability. Seventh, the benefits to circuit designs of cooperatively using several device types through heterogeneous integration (HGI) are strongly dependent on the design adaptability and circuit topology, which must be considered in any assessment. All the aforementioned complexities would mandate a complete circuit design flow for performance evaluation, which is nevertheless impractically time-consuming. Therefore, an alternative evaluation method must be developed that accounts for the above factors with reasonable computational run time.

In this paper, we propose a new device evaluation framework, called Pareto optimization-based circuit-level evaluator for emerging devices (PROCEEDs), for fully circuit-aware benchmarking [17]. It incorporates typical circuit design flow flexibilities and tunes physically adjustable device and circuit parameters to generate realistic conclusions about the overall performance. PROCEED remedies the flaws enumerated above in the following ways.

1) It uses Pareto curves to analyze power-delay (PD), area-power, and area-delay (AD) tradeoffs over a practically wide range of power or performance. The corresponding output circuit benchmark metrics are design power (including dynamic and leakage power), minimum working clock period (equal to the critical delay), and design area (total area for all gates).
2) It allows for a range of logic gate sizes, multiple and adjustable V_t, and one or multiple V_{dd} to be used in the

Manuscript received July 12, 2014; revised November 1, 2014; accepted December 28, 2014. This work was supported by the IMPACT+ Project.

The authors are with the Department of Electrical Engineering, University of California at Los Angeles, Los Angeles, CA 90095 USA (e-mail: shaodiwang@g.ucla.edu; pandrew@ucla.edu; chui@ee.ucla.edu; puneet@ee.ucla.edu).

Color versions of one or more of the figures in this paper are available online at http://ieeexplore.ieee.org.

Digital Object Identifier 10.1109/TVLSI.2015.2393852

TABLE I

COMPARISON OF VARIABLES CONSIDERED IN BENCHMARK METHODOLOGIES IN THE LITERATURE

Methodologies:		**Ref. [1]**	**Ref. [2,4]**	**Ref. [3-5]**	**Ref. [6]**	**Ref. [7]**	**Ref. [8]**	**Ref. [9-11]**	**Ref. [12]**	**Ref. [13-14]**	**Ref. [15]**	**Ref. [16]**	**PROCEED**
Metrics		*CV/I*, CV^2, I_{on}/I_{off}	*PD*[1] Pareto Curves, I_{on}/I_{off}	*PD* Pareto Curves	Clock Frequency	*Sub-threshold Swing*, I_{on}	Energy, Clock Frequency	*CV/I*, CV^2,	Clock-cycle Per Instruction	PD Pareto Curves, Yield	PD Pareto Curves	PD Pareto	***PD*[1], *AP*[2], *and AD*[3] Pareto Curves**
Benchmark Circuit		Latch, Inverter Chain	µP[4]	µP[4]	µP[4]	Device	Inverter Chain	Small Logic Elements	µP[4]	µP[4]	Small circuits	µP[4]	Arbitrary Circuit (µP here)
Power management		✗	✗	✗	✗	✗	✗	✗	✗	✗	✗	✗	✓
Optimization Knobs	V_{DD}, V_t	✓	✓	✓	✗	✗	✓	✗	***V_{DD} only***	✓	✗	✓	✓
	Size	✗	✗	✗	✓	✗	✗	✗	✓	✗	✓	✓	✓
	Multiple V_{DD}, V_t	✗	✗	✗	✓	✗	✗	✗	✗	✗	✗	✗	✓
Circuit Conditions	**Interconnect**	✓	✓	✓	✓	✗	✓	✓	✓	✓	✓	✓	✓
	LDH[5]	✗	✗	✗	✗	✗	✗	✗	✗	Architecture simulation	Logic synthesis	**Logic synthesis, Architecture simulation**	✓
	Activity	✓	✓	✗	✓	✗	✓	✓	✗				✓
Device Model	**Current**	I_{on}, I_{off}	I_{on}, I_{half}, I_{off}	I_{on}, I_{half}, I_{off}	I_{on}, I_{off}	TCAD	Model	Model	R_{on}, R_{off}	Lookup table	Lookup table	Lookup table	**Compact Model**
	Capacitance	Fixed	Fixed	Full *C-V*	Fixed	N/A	Fixed	Full *C-V*	-	Lookup table	Lookup table	Lookup table	**Full *C-V***

[1]PD: Power-Delay. [2]AP: Area-Power. [3]AD: Area-Delay.[4]Micro Processor, [5]LDH: Logic Depth Histogram (estimated using slack histogram).

evaluation circuit benchmarks, in accord with realistic designs.

3) It utilizes compact or lookup table-based full device models to properly account for device operation at each bias.
4) To assess circuit topology, it considers the full chip characteristics, including LDH, interconnect loads, activity factor (i.e., average gate toggle rate), and average fan-out. It also realistically models interconnect loads as functions of gate size, rather than treating them as constants.
5) It analyzes the circuit impact due to device variability, e.g., random dopant fluctuation, and parasitic voltage drops by calculating delay for logic gates evaluated at different variation corners.
6) For computational efficiency, it adopts scalable Pareto optimization techniques.
7) It models power gating and dynamic voltage and frequency scaling (DVFS) to assess power management and scaling.
8) To explore the benefits of HGI of emerging technologies, it allows evaluation of digital circuits built with multiple types of devices onto the same chip.

In this paper, we introduce the PROCEED framework and demonstrate its efficacy by using it to evaluate several technology options, including: 1) traditional silicon-on-insulator (SOI) devices; 2) novel tunneling FETs (TFETs); and 3) HGI combinations of these devices. TFET is a new device concept capable of steep subthreshold switching and is therefore drawing intense interest for highly energy-efficient operation [19]. Previous works have investigated TFETs at the circuit and architecture levels, relying on simulated device characteristics to show that it is potentially more power efficient than CMOS in low-power applications [13]–[16]. In this paper, we use TFETs as a vehicle to perform a microprocessor-level study by comparing experimental Si TFETs with benchmark SOI technologies and elucidate their respective strengths and disadvantages. We outline the methodology behind PROCEED in Section II, and explain details of the Pareto optimization procedure in Section III. We present the results of our PROCEED study on TFET and SOI devices in Section IV. Finally, the conclusions are drawn in Section V.

II. OVERVIEW OF PROCEED FRAMEWORK

As shown in Fig. 1, typical inputs to PROCEED include interconnect information, such as average wire resistance (R) and capacitance (C) and chip size, circuit benchmark design (i.e., design LDH and average fan-out), variability (through V_{dd} drops, V_t shifts, and so on), full device I–V models, and operating activity, as well as optional constraints on V_{dd}, V_t, chip area, and the ratio of average to peak throughput (i.e., clock cycles per second). Simulation blocks with interconnect loads are generated for canonical circuit construction through a feedback process using input from the Pareto optimizer (through tuning parameters like V_{dd}, V_t, and gate sizes). Optimized results are generated in the form of the PD Pareto curve. Finally, the power management analysis, including DVFS and power gating, is performed using this Pareto curve. In its present implementation, the PROCEED can evaluate an arbitrary logic device candidate as long as it does not cause a dramatic change in circuit topology. For instance, the multistate logic devices fall outside PROCEED's current

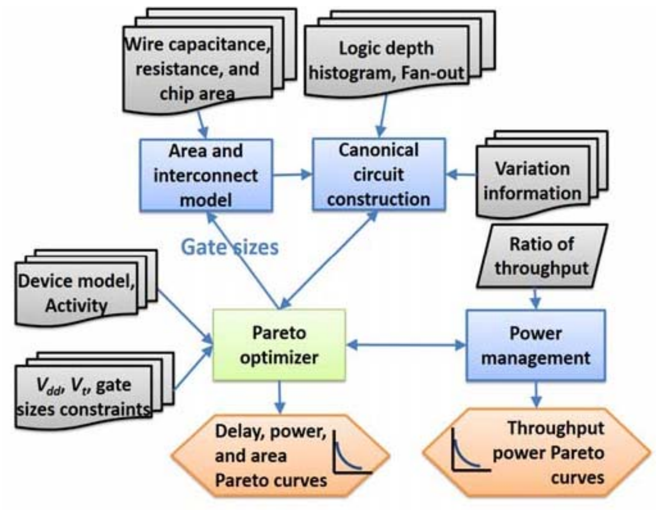
Wire capacitance, resistance, and chip area
Logic depth histogram, Fan-out
Area and interconnect model
Canonical circuit construction
Variation information
Gate sizes
Ratio of throughput
Device model, Activity
Pareto optimizer
Power management
V_{dd}, V_t, gate sizes constraints
Delay, power, and area Pareto curves
Throughput power Pareto curves

Fig. 1. Overview of PROCEED framework.

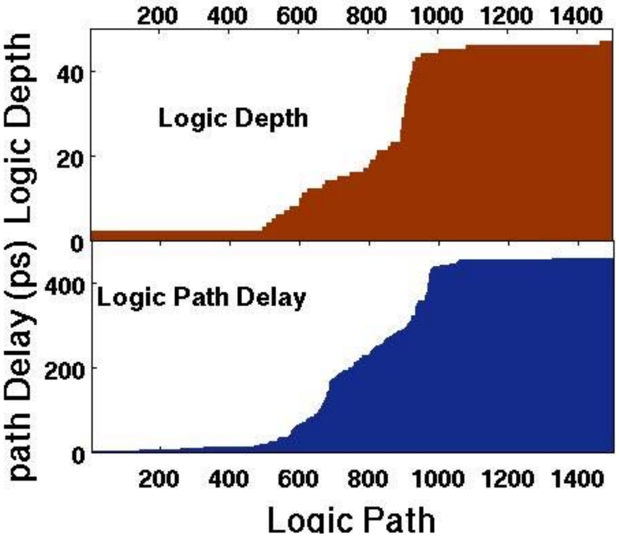

Fig. 2. Typical logic path depth distribution and logic path delay extracted from a synthesized CortexM0.

scope because of the unconventional circuit architectures in which they operate.

A. Canonical Circuit Construction

A complete and exact optimization is an impossible job for large digital circuits. Since the goal of PROCEED is to predict the best performance and power tradeoffs for emerging devices, detailed circuit design is not our target and it contributes little to evaluation. We therefore use only essential design information to maximize performance and determine the optimal V_{dd}, V_t, and gate sizes for a given power-consumption limit. Typical circuit designs contain both short and long logic paths with the path delay roughly proportional to the logic depth, as shown in Fig. 2, for a CortexM0 design. Hence, we derive the LDH by extracting endpoint slacks from benchmark designs and estimating logic paths.

In Fig. 3, we show an example of the simulation blocks used to construct a specific circuit. For simplicity, we first divide logic paths into n bins based on logic depth; in Fig. 3(a), for instance, $n = 5$. A larger number of bins improve accuracy at the expense of computation time. Each bin is modeled by corresponding simulation blocks S_i [S_1–S_5 in Fig. 3(a)], which are in turn made of i gate stages. We use the gate design for S_i to construct logic paths belonging to a given bin i.

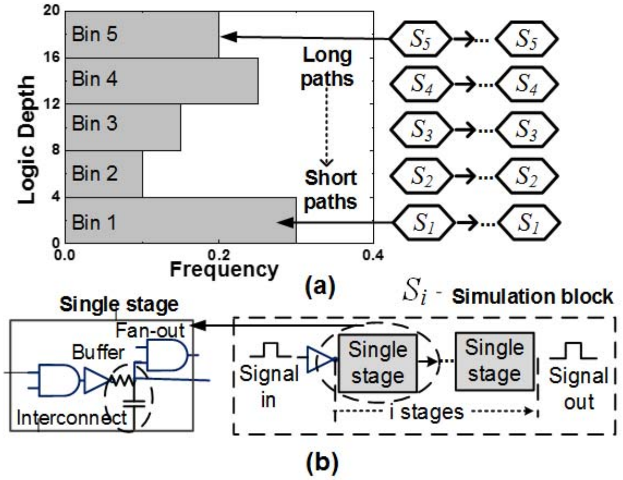

Fig. 3. (a) Example of simulation block allocation in PROCEED based on logic depth. (b) Circuit schematic used for simulation and optimization.

The LDH is divided in such a way that the longest path in each bin has the same delay if all these blocks have the same delay. Fig. 3 shows an example of this with five evenly spaced bins for logic paths from one to twenty stages such that the first bin contains one to four stage paths, the second holds paths with five to eight stages, and so forth. The delay weight W_D is the number of copies of S_i needed to construct the longest path in bin i ($W_D = 4$ in Fig. 3). The logic gate and interconnect used for a single stage in the simulation blocks is shown in Fig. 3(b). The gate can be NAND, NOR, or something more complicated like XNOR, depending on the average number of transistors per gate in the chosen benchmark. The gate choice can also differ from bin to bin, though in the examples in this paper, we use NAND gates for all bins. An inverter or buffer is inserted after the gate to drive the fan-out (which is a replica of the chosen gate sized to the average fan-out) as well as interconnects, which are represented by the RC elements.

B. Delay and Power Modeling

Delay, power, and area are the three most important gross metrics in the design of digital circuits, but usage constraints lead to tradeoffs between them that must be balanced to maximize the overall efficiency of the design. Hence, we use them as evaluation metrics in PROCEED. As described in the beginning of Section II-A, the circuit benchmarks are mapped to canonical circuits that are used to estimate these metrics without time-consuming large-scale simulations. The delay and power of the canonical circuits are extracted from SPICE simulations, which are then scaled, summed, and minimized to obtain the corresponding values for the given benchmark. We have verified that the values predicted by this method agree well with those calculated using commercial synthesis tools

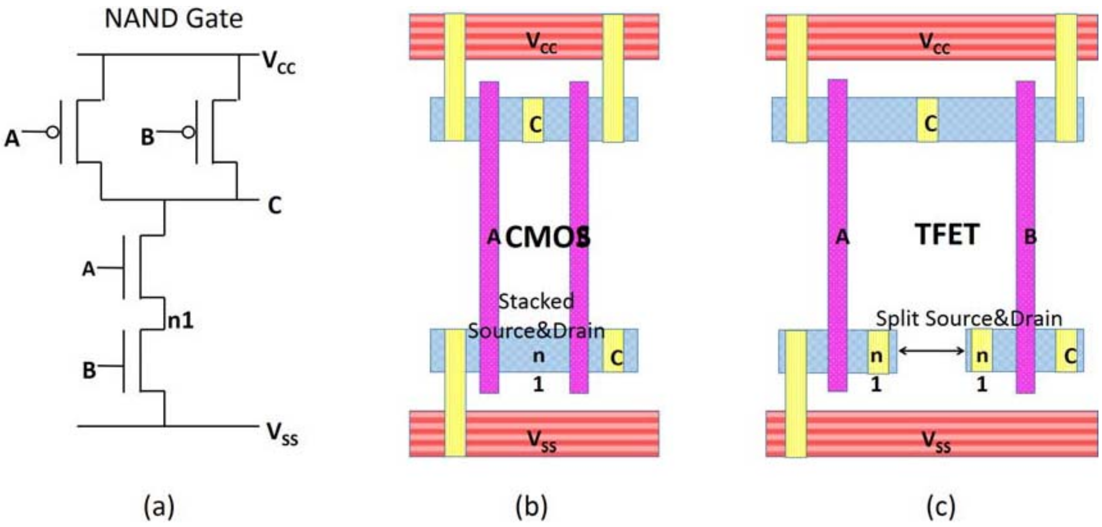

Fig. 4. (a) NAND gate. Schematic and layouts for (b) CMOS and (c) TFET.

(as described in Section IV-A), demonstrating PROCEED's accurate reproduction of realistic circuit behavior.

C. Area Modeling

The area of the gates used in canonical circuit constructions is simulated using UCLADRE[1] [18], where they are minimized in accordance with input design rules and gate netlists.

Unlike traditional CMOS devices, which have interchangeable source and drain, some emerging technologies such as TFETs [19]–[21], have asymmetric structures where current can only flow in one direction. This asymmetry prohibits stacking of transistors by sharing the source and drain. Fig. 4(a) shows a NAND gate logic schematic, where adjacent transistors share a source/drain at node $n1$. Fig. 4(b) stacks two nMOS devices to create a compact layout for traditional CMOS technology. However, due to the source/drain asymmetry, a TFET layout for the same circuit must split the stack, leading to additional area overhead, as shown in Fig. 4(c). To account for this effect, we modify UCLADRE such that it generates area-optimal TFET layouts for any input circuit netlist. The cell area of CMOS-based and TFET-based NAND gate as a function of gate width is shown in Fig. 5. The additional area overhead of TFET is clearly significant.

Design rules can differ depending on the technology; for instance, for nanotransfer HGI, the additional separations between p-wells and n-wells may be needed to eliminate overlay errors depending on the material choices for NFETs and PFETs [22]. Similarly, two devices with different design rules will result in different areas even if they are sized to the same gate width and length. For technology evaluations, we calibrate the design rules, sweep gate width in UCLADRE, and fit linear models of gate area to the simulation results. An example of the model's accuracy is shown in Fig. 6(a). The chip area is calculated using the following procedure: 1) the area of gates in each bin is obtained from the fitted area model; 2) the chip area is calculated as the weighted sum of gate areas in all bins; and 3) the weights are decided during canonical circuit construction stage.

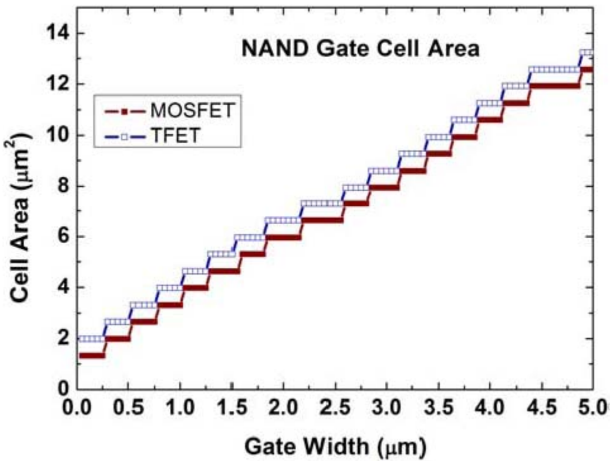

Fig. 5. Cell area of CMOS-based and TFET-based NAND gates as a function of gate width.

D. Process Variation and Voltage Drop

As devices scale to ever smaller technology nodes, the device variations due to process and ambient variations are becoming more important and should not be neglected in PD evaluation. In circuit design, the slow corner devices are commonly used to estimate the upper bound on minimum working clock period (critical delay) and create a safe design with sufficient delay margin. We define the slow corner as a device with reduced effective V_{dd} and increased V_t due to variability and parasitic effects, and the corresponding voltage shifts are inputted into PROCEED. Separate models for additional variability effects may be incorporated as needed. During circuit optimization, the worst case scenario is considered by calculating delay using the slow corner device and power using the normal device. An illustration of how this may affect the operating point of real devices is given for TFETs and SOI MOSFETs in Section IV-G.

[1] Freely available for download at http://nanocad.ee.ucla.edu/pub/Main/DownloadForm/PROCEED.tar.g

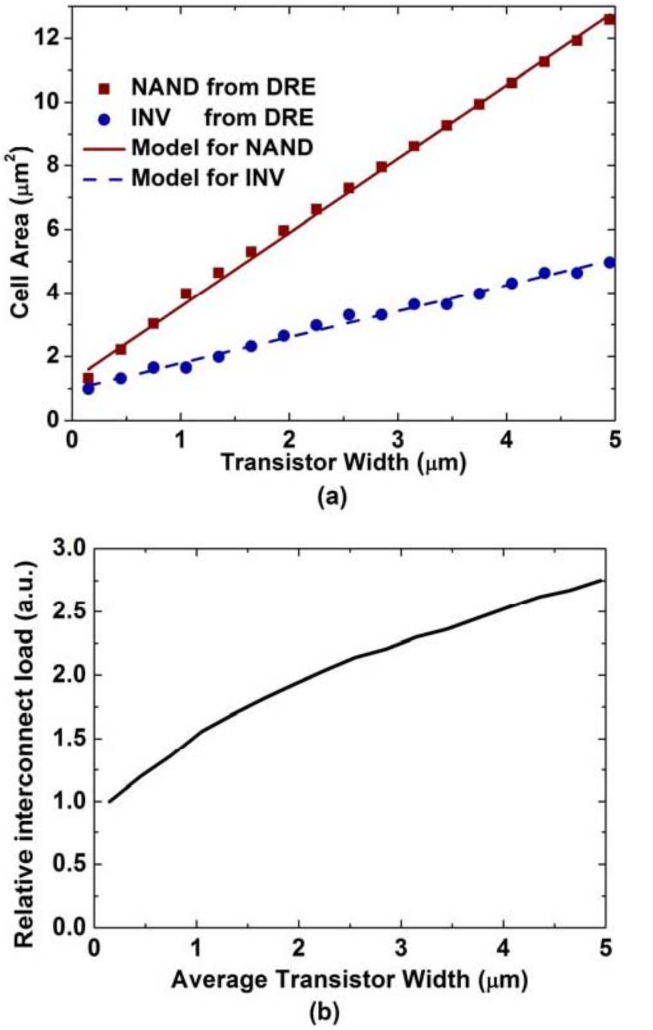

Fig. 6. (a) Cell area and (b) interconnect load as a function of transistor width. In (b), transistor width is the same in inverter and NAND gate.

E. Interconnect Load

We model interconnect loads using a series RC circuit. We assume R and C are linear with interconnect length, so the load will be proportional to the square root of the chip area [23], and can be dynamically changed based on average gate width. Fig. 6(b) shows an example of interconnect load as a function of transistor width, using a combined NAND and INV cell to estimate the cell area. The average RC and extracted gate width are then fed into PROCEED. Even simple considerations of interconnect load will strongly impact the overall evaluation results. In general, gates using devices with low driving ability need to be sized up to achieve the same performance as those with high driving ability. This increases the area of the chip as well as the interconnect loads, which exacerbates the drive demands and requires further gate sizing. The PROCEED correctly describes such cases, as quantitatively demonstrated in Section IV-C.

F. Pareto-Based Optimization

Following canonical circuit construction, all logic paths are replaced by simulation blocks (S_i) which will be optimized. However, these blocks cannot be optimized separately because they usually share a common V_{dd} and V_t, complicating the procedure. As a result, we use a modified form of a general simulation-based Pareto technique to perform the optimization [24], as discussed in more detail in Section III. The simulation target is regarded as a black box with two optimization objectives: any two of the design area, power, and critical delay (minimum working clock period).

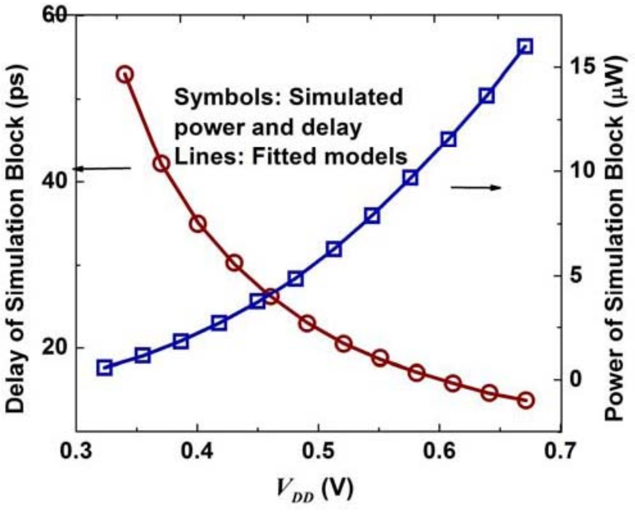

Fig. 7. Model fitting for simulation block's delay and power as a function of V_{dd}.

G. Power Management Modeling

Current technologies usually allow circuits to operate in at least three modes: 1) normal; 2) power saving; and 3) sleep mode. Previous evaluation works consider only the normal mode, where devices continuously work at peak performance. PROCEED allows devices to also operate at a second, lower supply V_{dd2} (DVFS) as well as in the OFF-state (power gating). This allows us to evaluate device PD scalability as a function of V_{dd}, an important circuit feature which, to the best of our knowledge, has been ignored in all previous evaluations.

The ratio of average to peak throughput is another input for PROCEED. To study power management, we choose all designs from the generated Pareto points, which achieve the lowest power and peak throughput. From this, the optimizer selects the best choice for the second power rail and divides the time spent operating at high V_{dd1} (i.e., the original supply) and the new lower V_{dd2}. This is done as follows. Starting from the optimized design (with maximized peak throughput), we carry out circuit simulations by sweeping voltages lower than the original V_{dd1}. The original design may even have multiple V_{dd}, in which case different blocks can use different V_{dd2} values. Delay and power models for every simulation block S_i as functions of V_{dd} are constructed using polynomial functions, as in Fig. 7

$$D_{\mathrm{Si}}(V) = \sum_{j=-2}^{5} a_{i,j} V^i, \quad P_{\mathrm{Si}}(V) = \sum_{j=-1}^{5} b_{i,j} V^i. \qquad (1)$$

We have tested and found this model to be sufficiently accurate; for instance, in our experiments presented in Section IV-F, the relative error of the polynomial fittings is <2%. We then optimize for the weighted power sum $f_1 P_1 + f_2 P_2$, subject to

$$D_2 >= W_D D_{\text{Si}}(V_{i2}),\ P_2 >= \sum_{i=1}^{n} W_{\text{Pi}} P_{\text{Si}}(V_{i2}), \quad i = 1, 2, \ldots, n$$
$$f_1 \cdot 1/D_1 + f_2 \cdot 1/D_2 >= T_{\text{Ave}}, \quad 0 \leq f_1 + f_2 \leq 1. \qquad (2)$$

Here, $D_{1,2}$ and $P_{1,2}$ are the design delay and power using $V_{\text{dd}1,2}$, W_D and W_P are the delay and power weights mapping from simulation blocks to the design, and f_1 and f_2 are the fractions of time spent operating with $V_{\text{dd}1}$ and $V_{\text{dd}2}$, with any remaining time assumed to be spent in the OFF-state. Typically, this step is not a feasible convex optimization problem; however, using the fitted model of (1), an enumeration approach can solve this problem very efficiently with acceptable accuracy. In Section IV-F, we give an example of how PROCEED's power management capabilities are applied in practice.

H. Activity Factor

Activity varies widely with application. In embedded sensing, for instance, factors <1% are observed in car-park management [25], while those for systems like VigilNet exceed 50% [26]. Activity factor can, therefore, dramatically change the evaluation results and is included as an input to PROCEED. In circuit simulations, the dynamic and leakage power are separately extracted and the total power is equal to their weighted sum. From this, the circuit can be optimized for a known activity factor. This can be of primary importance in determining the usability of a given device, as we experimentally show in Section IV-E.

I. Multiple V_{dd} and V_t

In modern circuit designs, the multiple V_{dd} and V_t values are used, as shown in Section IV-B. In our scheme, transistors in each simulation block S_i must be assigned the same voltages, so to optimize a design with integer m different V_{dd} or V_t biases, the number of simulation blocks must be greater than m. In addition, our optimization is an iterative process whereby Pareto points are updated and improved based on previous iterations. Therefore, if the same V_{dd} or V_t is shared by multiple simulation blocks, this assignment cannot be changed during the optimization. A full optimization for multiple V_{dd} and V_t is implemented by considering designs with all sets of reasonable voltage assignments in parallel. For example, if we have five simulation blocks S_1–S_5 and two available V_t, then for i from 1 to 4, blocks S_1 to S_i use the high V_t and S_{i+1} to S_5 use the low V_t. This comprises the set of useful voltage assignments, since simulation blocks with longer logic paths require higher performance (i.e., lower V_t).

J. Heterogeneous Integration

Every emerging device has its own characteristic advantages, such as steep subthreshold slope for TFETs and high mobility and ON-current for III–V or CNT FETs. However, any one of these devices cannot fulfill all the disparate requirements of the various macros in the future circuit applications. HGI combines several types of devices onto a single chip to maximize performance at the expense of cost and area penalties [22]. In PROCEED, we use a quick way to explore the benefits brought by this technology. In general, the slow and low-power devices are useful for nontiming critical macros, while high-performance devices are suitable for high-speed macros. Furthermore, within single circuit blocks, the critical and noncritical paths can be built using different types of devices. In PROCEED, the models for all HGI devices are inputted and the delay and power of logic paths built using different devices are modeled accordingly. Since these devices operate in the same circuit and affect the overall performance, the PROCEED optimizes the HGI gates in a concurrent fashion. Since different devices are apportioned among the available logic path bins, the granularity of HGI optimization results in this approach is set by the number of bins considered. PROCEED, therefore, allows us rapidly evaluate combinations of multiple technologies over a wide range of delay, power, and area requirements. As an example of this, we evaluate the potential of circuits implemented using TFET and SOI HGI in Section IV-H.

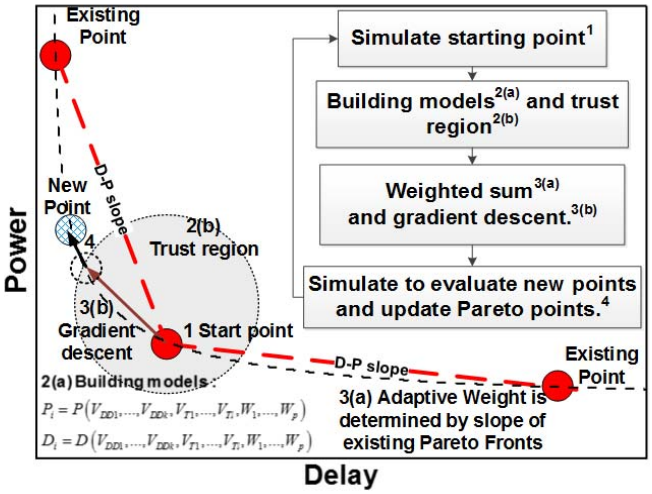
Existing Point
New Point
D-P slope
2(b) Trust region
4
3(b) Gradient descent
1 Start point
D-P slope
Existing Point
2(a) Building models:
$P_i = P(V_{DD1},\ldots,V_{DDk},V_{T1},\ldots,V_{Tl},W_1,\ldots,W_p)$
$D_i = D(V_{DD1},\ldots,V_{DDk},V_{T1},\ldots,V_{Tl},W_1,\ldots,W_p)$
3(a) Adaptive Weight is determined by slope of existing Pareto Fronts
Simulate starting point^1
Building models$^{2(a)}$ and trust region$^{2(b)}$
Weighted sum$^{3(a)}$ and gradient descent.$^{3(b)}$
Simulate to evaluate new points and update Pareto points.^{4}
Power
Delay

Fig. 8. Optimizer overview. Adaptive weight is chosen by slope of existing fronts. Based on starting point, metamodeling is built and gradient descent is used to find potential points. Simulate potential points to get new Pareto points.

III. Pareto Optimization

PROCEED can simultaneously optimize any two metrics out of delay, power, and area, while the third is treated as a constraint; for instance, we can perform a Pareto optimization of delay and power with a maximum area constraint. As described in Section II-B, the chosen area model is linear in gate width and hence easier to optimize than delay and power. Therefore, in the remainder of this section, we will describe in detail the Pareto optimization of delay and power with constrained area. Fig. 8 shows an overview of our Pareto optimization process. PROCEED treats circuit simulations as a black box and uses models to optimize tuning parameters based on the simulation results. Gradient descent is utilized to find minimal objectives in the trust region. Final simulations are performed on designs outputted by the model-based optimization. The vector of tuning parameters $\boldsymbol{X}$ for optimization is represented as

$$\begin{aligned}\boldsymbol{X} &= (x_1, x_2, \ldots, x_m) = (\boldsymbol{y_1}, \boldsymbol{y_2}, \ldots, \boldsymbol{y}_n)\\ \boldsymbol{y}_i &= (V_{\mathrm{dd},i}, V_{t,i}, W_{i1}, W_{i2}, \ldots, W_{i,2i}), \quad i = 1, \ldots, n \end{aligned} \tag{3}$$

where x_j are the variables of $\boldsymbol{X}, \boldsymbol{y_i}$ are vectors of the tuning parameter variables for simulation block S_i, and $V_{\mathrm{dd},i}$, $V_{t,i}$, and W_{ij} are the supply, the threshold voltages, and the sizes of gates and inverters in each block S_i, respectively. The optimization entails the following steps.

A. Picking a Starting Point

Each iteration of the optimization process uses a starting set of variables $\boldsymbol{X_0}$ around which to explore. For the first iteration, any reasonable $\boldsymbol{X_0}$ may be inputted. The choice of the initial point can affect runtime but not final accuracy, since bad points will gradually be eliminated by the optimization process and converge to the true answer. Subsequently, $\boldsymbol{X_0}$ is determined from already existing Pareto points by computing the Euclidean distance between all neighboring points in delay-power coordinates, as shown in Fig. 8. The point with the largest total distance from its two neighbors is chosen as the starting point $\boldsymbol{X_0}$ since it lies in the sparse region, which is usually suboptimal.

B. Building a Local Model Around $\boldsymbol{X_0}$

To accelerate the optimization process, the second-order delay, and the power models are constructed based on the simulation results. The delay and power models D_{Si} and P_{Si} for each block S_i are calculated separately and then combined to obtain the model for the whole canonical circuit. Compared with simultaneously calculating model parameters for all blocks, this approach reduces the number of simulations, as determined by the size of the Hessian matrix (proportional to the number of variables squared). D_{Si} and P_{Si} are represented by the gradient vector $\boldsymbol{G}$ and Hessian matrix $\boldsymbol{H}$ as

$$\begin{aligned} D_{\mathrm{Si}}(\boldsymbol{y_{i,0}} + \boldsymbol{\Delta y}_i) &= D_{\mathrm{Si},0} + \boldsymbol{G_{Di}^T \Delta y_i} + \frac{1}{2}\boldsymbol{\Delta y_i^T H_{Di} \Delta y_i}\\ P_{\mathrm{Si}}(\boldsymbol{y_{i,0}} + \boldsymbol{\Delta y}_i) &= P_{\mathrm{Si},0} + \boldsymbol{G_{Pi}^T \Delta y_i} + \frac{1}{2}\boldsymbol{\Delta y_i^T H_{Pi} \Delta y_i}. \end{aligned} \tag{4}$$

This second-order model is a local estimation near the starting point. To guarantee validity, an adaptive trust region is applied, as shown in Fig. 8, limiting the model range inside the region

$$\boldsymbol{X_0} - \boldsymbol{\lambda}(r) < \boldsymbol{X} < \boldsymbol{X_0} + \boldsymbol{\lambda}(r) \tag{5}$$

where r is the radius of this trust region and $\boldsymbol{\lambda}$ is the range of the tuning parameters, and $\boldsymbol{X}$ is a linear function of r.

C. Model-Based Optimization

In this step, four metrics are used in optimization: 1) D; 2) P; 3) $W_{\mathrm{dl}} \times D + W_{\mathrm{pl}} \times P$; and 4) $W_{\mathrm{dr}} \times D + W_{\mathrm{pr}} \times P$. Minimization of D and P yields the fastest and lowest power designs in the local region, while the weighted sums of delay and power are used to populate the phase space by finding two Pareto points between the starting point and its neighbors. The optimization also needs to satisfy the constraint from the third metric (e.g., area in this case). Since the problem may not be convex, gradient descent with the logarithmic barrier method [27] is used to find these optimal points. The model's region of validity lies in the intersection of the trust region and the inputted bounds for the tuning parameters. The objective function is performed as follows:

$$\begin{aligned} &\text{Minimize } W_D D(\boldsymbol{X}) + W_P P(\boldsymbol{X})\\ &\qquad -t\left(\sum_{j=1}^{m} \log|x_j - x_{j,b}| - \log(-A(\boldsymbol{X}) + A_{\max})\right)\\ D(\boldsymbol{X}) &= D(\boldsymbol{X_0}) + \boldsymbol{G_D}(\boldsymbol{X_0})^T(\boldsymbol{X} - \boldsymbol{X_0})\\ &\quad + (\boldsymbol{X} - \boldsymbol{X_0})^T \boldsymbol{H_D}(\boldsymbol{X_0})(\boldsymbol{X} - \boldsymbol{X_0})\\ P(\boldsymbol{X}) &= P(\boldsymbol{X_0}) + \boldsymbol{G_P}(\boldsymbol{X_0})^T(\boldsymbol{X} - \boldsymbol{X_0})\\ &\quad + (\boldsymbol{X} - \boldsymbol{X_0})^T \boldsymbol{H_P}(X_0)(\boldsymbol{X} - \boldsymbol{X_0}) \end{aligned} \tag{6}$$

where $x_{j,b}$ are the upper and the lower bounds for variable x_j, $A(\boldsymbol{X})$ and $A_{\max}$ are the area model and the maximum area constraint, respectively, and $D(\boldsymbol{X})$ and $P(\boldsymbol{X})$ are the delay and power for the entire design, respectively. The weights for delay and power are defined as follows:

$$\begin{aligned} W_{\mathrm{dl}(r)} &= (P_{l(r)} - P_0)\Big/\sqrt{(P_{l(r)} - P_0)^2 + (D_{l(r)} - D_0)^2}\\ W_{\mathrm{pl}(r)} &= (D_0 - D_{l(r)})\Big/\sqrt{(P_{l(r)} - P_0)^2 + (D_{l(r)} - D_0)^2} \end{aligned} \tag{7}$$

where (D_0, P_0) is the starting point and (D_l, P_l) and (D_r, P_r) are the left and right neighbor points, respectively. The solid points in Fig. 8 are examples of such points. The direction vectors $(W_{\mathrm{dl}}, W_{\mathrm{pl}})$ and $(W_{\mathrm{dr}}, W_{\mathrm{pr}})$ of the weighted sum of objectives are calculated so as to be perpendicular to the connecting lines between the starting point and its neighbors, as shown by the dashed line in Fig. 8. The weights in the weighted sum optimization are used to yield the two new optimal points between the starting point and its neighbors. D and P are given by

$$\begin{aligned} D(\boldsymbol{X}) &= W_D \cdot \max((D_{S1}(\boldsymbol{y}_1), D_{S2}(\boldsymbol{y}_2), \ldots, D_{Sn}(\boldsymbol{y}_n)))\\ P &= \sum_{i=1}^{n} W_i \cdot P_{\mathrm{Si}} \end{aligned} \tag{8}$$

where W_D is the delay weight discussed in Section II-A and $\boldsymbol{W_i}$ is the number of S_i used in the canonical circuit construction. Because the maximizing function does not have a continuous derivative, we use higher order norms to estimate the maximum, so the elements of gradient vector and Hessian matrix for delay are derived as follows:

$$\begin{aligned} D(\boldsymbol{X}) &\approx \|\boldsymbol{D}\|_K,\ \boldsymbol{D} = (D_{S1}(\boldsymbol{y}_1), D_{S2}(\boldsymbol{y}_2), \ldots, D_{Sn}(\boldsymbol{y}_n))\\ G_{D,j}(\boldsymbol{X}) &= \frac{\partial D(\boldsymbol{X})}{\partial x_j} \approx \frac{\partial \|\boldsymbol{D}\|_K}{\partial x_j}\\ H_{D,jk}(\boldsymbol{X}) &= \frac{\partial^2 D(\boldsymbol{X})}{\partial x_j \partial x_k} \approx \frac{\partial^2 \|\boldsymbol{D}\|_K}{\partial x_j \partial x_k} \end{aligned} \tag{9}$$

where K is the order of the norm. Higher K results in more accurate results (we use $K = 100$ in our simulations).

Similarly, the elements of the gradient vector and Hessian matrix for power are given as

$$G_{P,j} = \frac{\partial P(\boldsymbol{X})}{\partial x_j} = \sum_{i=1}^{n} W_i \cdot \frac{\partial P_{\mathrm{Si}}(\boldsymbol{y}_i)}{\partial x_j}$$

$$\mathrm{H}_{P,jk}\frac{\partial^2 P(\boldsymbol{X})}{\partial x_j \partial x_k} = \sum_{i=1}^{n} W_i \cdot \frac{\partial^2 P_{\mathrm{Si}}(\boldsymbol{y}_i)}{\partial x_j \partial x_k}. \qquad (10)$$

D. Addition of New Pareto Points

To correct for model errors, the circuit simulations are performed to evaluate D and P for all remaining potential Pareto points found by the optimization. In Fig. 8, this process is illustrated by the shift of the hatched point to the dotted circle. Finally, points not on the Pareto frontier (such that at least one other point with both lower delay and power exists) are filtered out.

E. Iteration Termination

For each iteration, when choosing the starting point for each step, the radius of trust region around this point is decreased by a factor of $p(p > 1)$. Two termination conditions are applied. The first condition is the existence of a sufficient Pareto point density in the region of interest, defined by the largest gap between any two neighboring points being smaller than a given criterion. This condition is usually used for devices with wide operating regions (i.e., suitable for both high-performance and low-power applications). The second condition is the reduction of the radius of trust region below a given criteria. This usually occurs due to limitations on the device operating region or device model discontinuities.

The PROCEED runtime is of the order $O\ (r \times m^2) + O(r)$, where r is the resolution constraint (number of points in a unit Pareto curve), m is the total number of tuning parameters, $O\ (r \times m^2)$ is the complexity of the simulations for gradient and Hessian matrix calculation, and $O(r)$ is the complexity of simulating potential Pareto points. In our experiments, runtimes are mainly dominated by the resolution constraint; however, for large m, the $O\ (r \times m^2)$ term will dominate. The average PROCEED runtime to generate a full Pareto curve over three orders of magnitude in performance is ~4 h on a single CPU. We use MATLAB in the optimization process and HSPICE for circuit simulations.

IV. Experimental Results

To illustrate PROCEED's capabilities, we use it to assess SOI and silicon TFET devices at the 45-nm node and compare the evaluation results with existing methods. Because of their interband tunneling conduction principle, TFETs are capable of very low leakage and extremely steep subthreshold swing, making them well-suited for low-voltage operation [19]. Currently, however, nonidealities in experimental devices and low ON-current limit their performance. We examine the viability of currently achievable TFETs using a device compact model [28], [29] calibrated against TCAD simulations and experimental SOI devices [30]. While this does not represent the best possible TFET, which may require a different channel material or device structure, it have the advantages of being experimentally validated and structurally comparable with conventional SOI devices and represents a realistic lower bound.

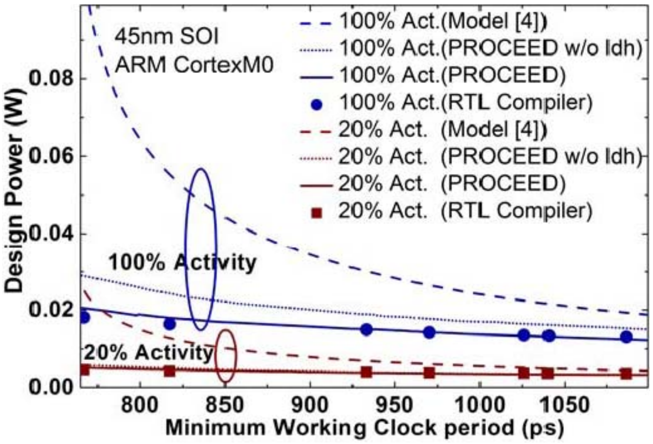

Fig. 9. Pareto curves for delay and power as evaluated using a commercial synthesis tool, Model [4], and PROCEED. V_{dd} and V_t are constants and only gate size is a variable.

For these reasons, we emphasize that our results do not constitute a final judgment on the viability or lack thereof of TFETs in future circuits; rather, they represent both a starting point from which to consider possible uses for present experimental (rather than projected) TFETs as well as a platform to demonstrate PROCEED's unique capabilities. Traditional technologies are represented by 45-nm SOI MOSFETs modeled using commercial characteristics and compact model. Unless otherwise specified, all circuit results are generated with one V_{dd} and two V_t. To easily compare devices, we will frequently refer to the Pareto crossover, defined as the (minimum working) clock period (critical delay) above which the optimized novel device (here, the TFET) consumes less power than the established technology (SOI); lower Pareto crossover means the novel device is more promising for a given case since it has a wider operating range over which it is superior.

A. Framework Evaluation

To validate the PROCEED framework, we use the widely employed evaluation model of [4] (hereafter Model [4]), and a commercial synthesis tool to evaluate the PD Pareto curve for a CortexM0 microprocessor with a commercial 45-nm SOI library and model. The information needed for PROCEED and Model [4] (LDH, average fan-out, and interconnect load) is extracted from a synthesized, placed, and routed netlist at a minimum working clock period of 933 ps. Only single constant values of V_{dd} and V_t are used, as Model [4] does not support multiple voltages and the commercial library has only constant V_{dd} and V_t.

As shown in Fig. 9, the PROCEED predictions are in much better agreement with the comprehensive optimized results from the register-transfer level (RTL) compiler compared with Model [4], which is frequently used for device evaluation [2], [3]. The operating range for comparison is chosen by the synthesis results with the commercial library using one V_{dd} and V_t. We note that using the compiler for evaluation purposes is completely impracticable, since extracting a Pareto curve from kilohertz to gigahertz clock frequencies necessitates libraries with V_{dd} and V_t varying from 0.5 to 1.2 V and 0.1 to 0.5 V, respectively. However, the generation and optimization of these libraries would consume months of runtime, whereas we completed the same study in hours using PROCEED. Meanwhile, the computationally simple Model [4] takes seconds to complete such Pareto curves but grossly overestimates power for two reasons: 1) the neglect of LDH in assuming all gates have the same (large) size used for the critical path and 2) the use of analytical PD models rather than circuit simulations using full device characteristics. The dotted line is the Pareto curve generated by PROCEED while neglecting LDH, illustrating the accuracy improvement contributed by the two aforementioned points. We further note that Model [4] cannot account for adaptability, variability, or multiple V_{dd} and V_t effects. By benchmarking to the RTL results in Fig. 9, we observe PROCEED improves accuracy by 3× to 115× compared with the current standard Model [4].

B. Impact of Multiple V_{dd}, V_t, and Gate Sizing

Additional tuning parameters create a larger design space for design optimization, as shown in Fig. 10, for a 45-nm SOI CortexM0 topology. As more LDH bin divisions are introduced, power is increasingly optimized because of a greater range of gate sizes from which to construct the design. Similarly, the introduction of additional V_{dd} rails and V_t substantially improves power consumption, although the results do not account for the overhead of the voltage shifter used in multiple V_{dd} design. Finer bin division in the LDH also leads to a better optimized DA curve. In PROCEED, the number of V_{dd}'s and V_t's does not impact the DA Pareto curve because they are not associated with area calculation, so they automatically converge to their limiting values to achieve the highest possible performance during area and delay co-optimization. Overall, we observe that the evaluated optimal power at a given minimum working clock period may change by over 50% as gate size tuning and multiple V_{dd} and V_t are introduced, demonstrating the necessity of including these effects in any quantitative comparison.

C. Impact of Interconnect Load

As described in Section II-E, the PROCEED considers interconnect as a function of gate sizes. Large gates result in large chip area and high interconnect load. For instance, when all gates in a design are sized 1× larger, the delay on interconnect improves less than 1× because interconnect load increases with upsizing gate. The impact of interconnect loads is observed in Fig. 11. We compare the area-minimum working clock period tradeoff for the CortexM0 utilizing TFETs with and without the size-scaled interconnect model. Without such model, the interconnect load is set to a constant value independent of gate size, such that sizing leads to an exactly proportional free performance boost. We observe that

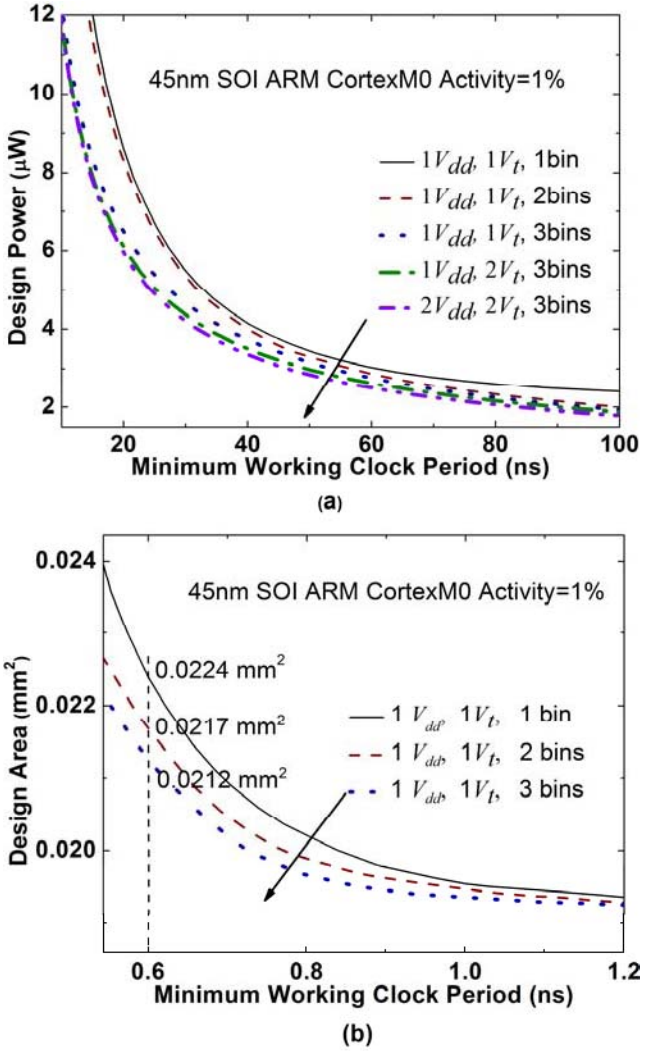

Fig. 10. 45-nm SOI CortexM0 (a) power-minimum working clock period and (b) area-minimum working clock period as tuning parameters are increased.

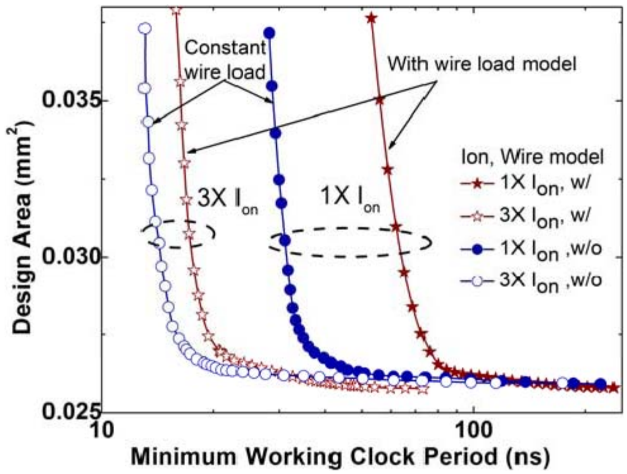

Fig. 11. 45-nm TFET and boosted TFET (3× current) Pareto curves of delay and chip area for the CortexM0 design. Red curves: results using a hypothetical TFET with 3× current boost.

as the minimum working clock period reduces, the longer interconnects necessitated by gate sizing substantially increase the total area and the assumption of constant interconnect load leads to major inaccuracies. For instance, at a 50-ns minimum working clock period, using the gate size-scaled interconnect model increases total area by 50%.

Since TFETs often suffer from low drive currents and the interconnect model has a larger impact for high-performance operation, we also performed the area and delay optimization using a hypothetical device with 3× larger current than the experimental TFETs we have been considering [30]. Note that this and even larger performance boosts should be possible through various device improvements, such as channel material choice, doping profile, and geometry. The benefits of larger current are greater when size scaling of interconnects is considered. The wider shift from 1× to 3× happens on curves using the interconnect model (which boosts the performance by more than 3× on average) compared with the constant load cases (where current boost simply reduces the delay by the same 3× factor). This is because the reduced gate sizing of the boosted TFET required to reach a given delay also reduces interconnect lengths. A consistent interconnect model that scales consistently with gate sizing can dramatically change chip area and is, therefore, crucial for evaluating the overall impact of different technologies.

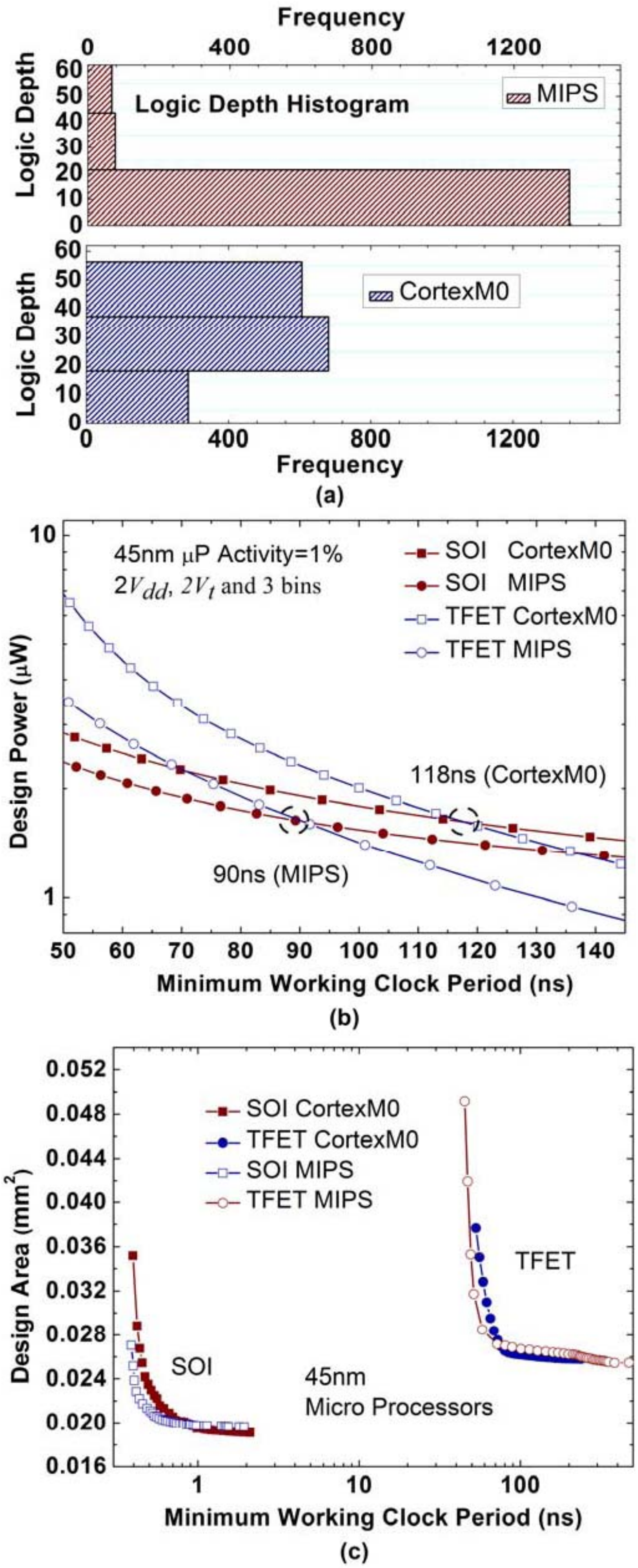

Fig. 12. (a) LDH of MIPS and CortexM0. (b) Power and delay curves and (c) area and delay curves for MIPS and CortexM0 designed with TFET and SOI, respectively. Activity is 1% and one V_{dd}, one V_t and two bins are applied.

D. Impact of Benchmarks on Evaluation—SOI Versus TFET

To show the impact of benchmark selection, we compare the performance of two microprocessors, CortexM0 and microcontroller (MIPS), using SOI and TFET devices and two V_{dd}'s and two V_t's. We choose these benchmarks because, as shown in Fig. 12(a), they have a similar number of critical path stages (56 in CortexM0 versus 62 in MIPS) and total gates (8990 versus 9248), but the CortexM0 has a more evenly distributed LDH. The power consumption in MIPS is dominated by short paths, which means it will be more accommodating of slow devices compared with the CortexM0. Accordingly, in Fig. 12(b), both SOI and TFET achieve better power efficiency in MIPS designs, because the second V_{dd} and V_t can be optimized to save power along the short paths. The crossover points, where the Pareto curves for different devices intersect define their advantageous operating regions; a device changes from being less power efficient on one side of the crossover to being more efficient on the other side. If multiple crossovers are found, then the Pareto curve can be divided into several regions (high performance, low power, and so on) such that in each one, there is only a single crossover point. This allows us to demarcate the (possibly multiple) favorable operating ranges for each device. The Pareto crossover occurs at 90 and 118 ns for MIPS and CortexM0, respectively, showing that TFETs are more acceptable for applications like MIPS, which tolerate slower devices. However, for practical applications, the drive currents for Si TFETs must be increased to reduce sizing and save more dynamic power at high clock rates. As previously mentioned, this may be achieved in practice through a variety of TFET optimization pathways. In Fig. 12(c), SOI beats the existing Si TFETs in area and minimum working clock period curves by a wide margin, with no crossover points.

Again, for both SOI and TFETs, better area efficiency is observed in the high-performance regime for MIPS on account of the concentration of short paths in its LDH. In the low-performance region, all gates are relatively small, so

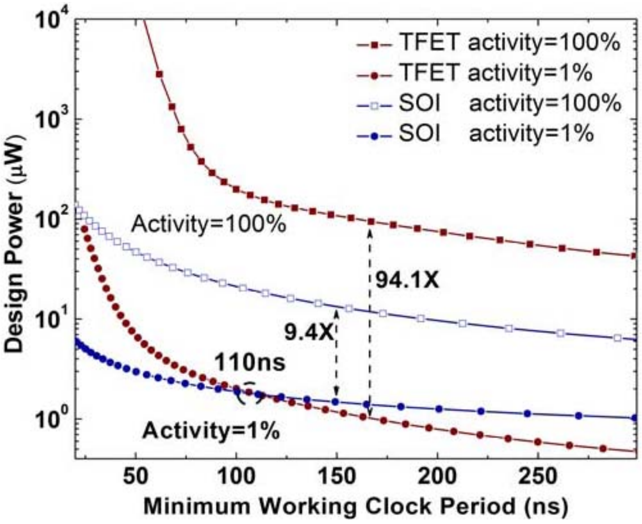

Fig. 13. Activity impact on 45-nm SOI and TFET CortexM0's power-minimum working clock period.

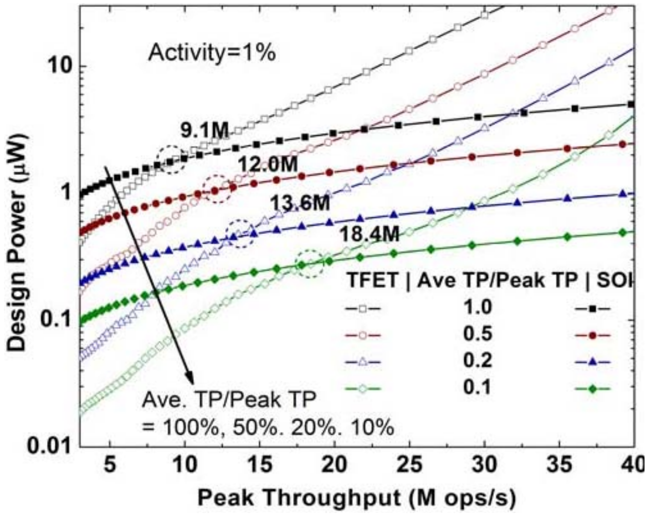

Fig. 14. 45-nm SOI and TFET CortexM0 microprocessors with power management. The ratios of average to peak throughput are 10%, 20%, 50%, and 100%. Curves with ratios of 100% are designs outputted from Pareto optimizer.

MIPS uses more area than CortexM0 on account of its larger number of gates. By contrast, CortexM0, which contains fewer gates, occupies bigger areas for high performance due to the fact that the longer logic paths in its LDH require larger gates. Previous evaluations, like those in Table I, which ignore LDH, are not able to distinguish between benchmarks in this way. These results show how the choice of circuit topology strongly impacts the suitability of emerging devices.

E. Impact of Activity Factor—SOI Versus TFET

We next examine how activity factor affects SOI- and TFET-based CortexM0 processors in Fig. 13. As activity reduces from 100% to 1%, TFET circuit power scales in lockstep by 94.1× due to low device leakage. However, the corresponding SOI designs only see power reduction of 9.4× because of its higher OFF-current. We see that TFETs change from being completely impracticable at 100% activity to being superior to SOI beyond the 110-ns minimum working clock period point at 1% activity; thus activity factor, and hence system use contexts, can drastically alter the device evaluation and must be considered.

F. Power Management Modeling

The results of the previous sections make clear that there is no panacea device and that device-circuit evaluation must be done with specific applications and operating windows in mind. DVFS and power gating are crucial ingredients for such usage-mindful evaluation. In Fig. 14, we show PROCEED-generated Pareto curves at different ratios of average to peak throughputs for SOI and TFET CortexM0 using DVFS and power gating. Power is reduced by operating at the lower supply rail or turned OFF by power gating; the achievable power reduction differs with device and operating region. The peak throughput crossover point for TFETs shifts from 9.1- to 18.4-M operations per second as the ratio of average to peak throughput reduces from 100% to 10%; the relative performance of TFETs effectively doubles as throughput requirements become less aggressive, emphasizing the importance of incorporating power management into device benchmarking.

G. Variation-Aware Evaluation

To illustrate how variability might impact conclusions drawn using nominal devices, we show in Fig. 15 how the SOI and TFET Pareto curves are changed when slow corner devices are used. We define the slow corner as a device with 5% effective voltage reduction and 50 mV V_t shift; total power is simulated using the nominal device, while delay is evaluated with the slow corner. We observe in Fig. 15(a) that the TFET is more vulnerable to variability effects than SOI, as the Pareto crossover of minimum working clock period is shifted from 110 to 205 ns when variation is considered. In Fig. 15(b), the area and minimum working clock period curve in the presence of variability is shifted more for TFETs compared with SOI. This is due to the TFET's steep subthreshold swing and its low-operating voltage, leading a high sensitivity of drive current to voltage [31], [32]. This suggests that the TFETs need to show substantial nominal device advantages in order to buffer this sensitivity and demonstrates that even a simple consideration of variability is important in device evaluation and selection.

H. Heterogeneous Integration Evaluation

In this section, we evaluate three types of technologies: 1) integrated circuits using SOI only; 2) integrated circuits using TFET only; and 3) integrated circuits using HGI of both SOI and TFET. From the previous results, the low leakage of TFETs makes them suitable for circuits with low activity or LDH dominated by short paths. HGI offers the chance to merge the strengths of TFETs with the higher performance of SOI to maximize their benefits. TFET is used to build gates in short logic paths, which can save more leakage power without

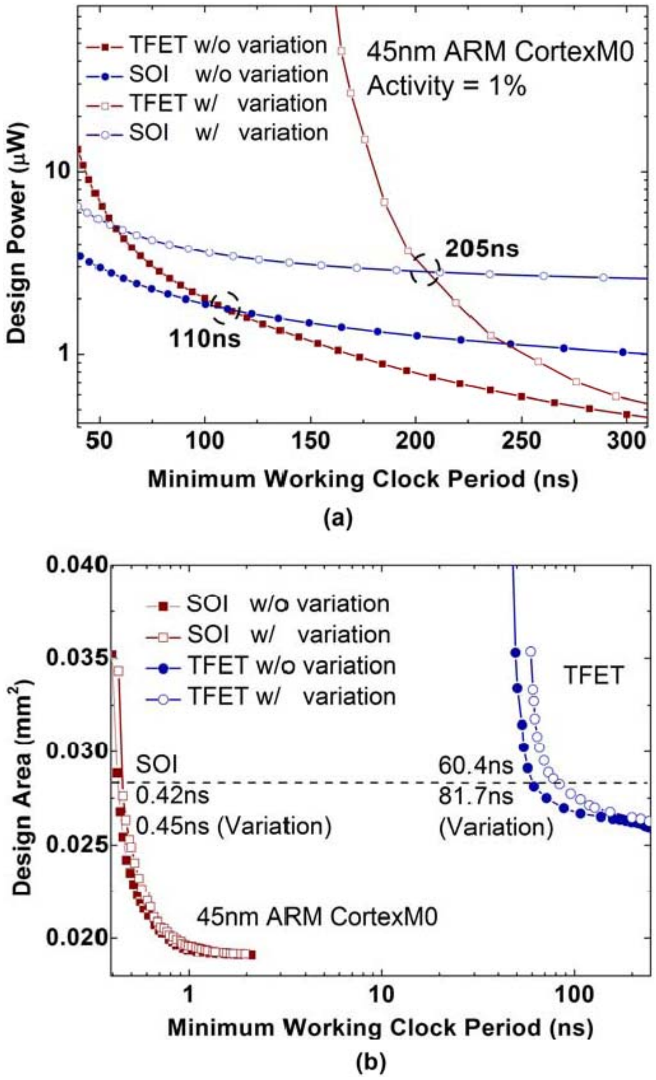

Fig. 15. Variation-aware (a) PD and (b) AD evaluations of 45-nm technologies. Assumed voltage drop is 90%, and V_t shift is 50 mV.

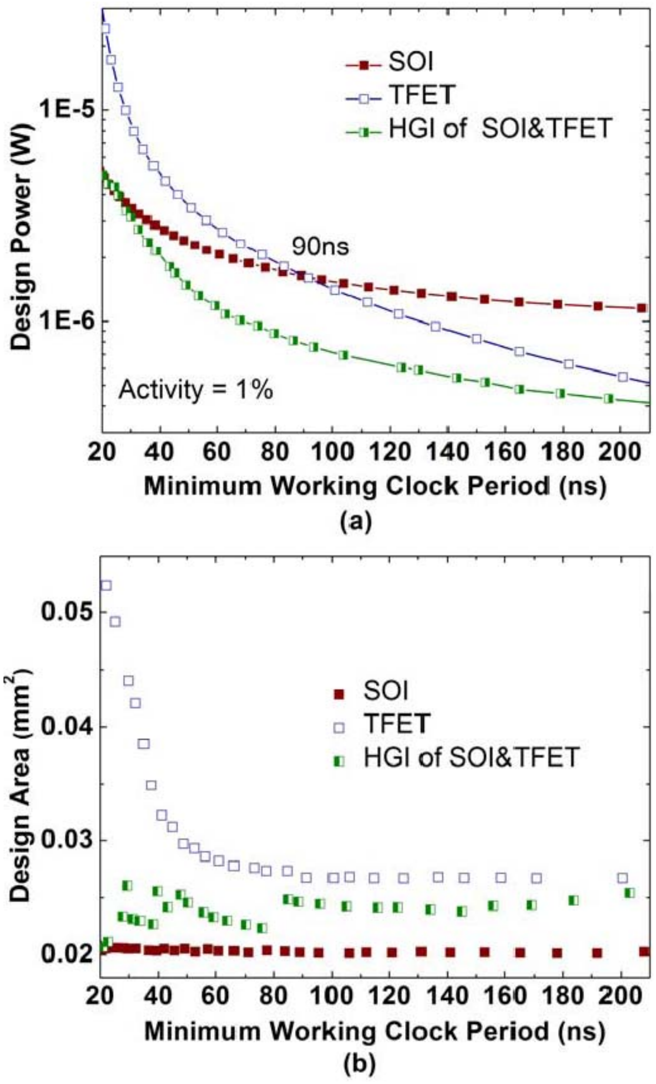

Fig. 16. (a) PD optimization for 45-nm HGI and non-HGI MIPS and (b) its corresponding design area. The fluctuations in the latter arise because the optimization is carried out for power and delay, not design area. Two sets of V_{dd} and V_t are adjusted during optimization, one for SOI devices and the other for TFETs.

sacrificing performance, because minimum working clock period is decided by long paths where SOI is applied. The optimal DP curves for these three technologies are compared in Fig. 16(a). Fig. 16(b) shows the corresponding design area and delay for the DP optimized designs for each technology, assuming 45-nm MIPS. Owing to accuracy constraints of the SPICE simulation tool, HGI is evaluated in PROCEED using four bins (which are divided and assigned to either TFET or SOI). In MIPS, the gates are mostly distributed along short paths, where devices are mainly idle and leakage power is more significant. The much lower leakage of TFETs gives them a big advantage when designing slow gates, while their performance constraints (due to low current) are mitigated using SOI for gates along critical paths.

Accordingly, we see in Fig. 16(a) that HGI outperforms non-HGI circuits between the minimum working clock periods of 20 and 200 ns. In this intermediate region, the respective advantages of TFET and SOI can be combined to give significantly better overall performance. In the (leftmost) high-performance region, the high drive capabilities of SOI dominate circuit operation, such that the optimized HGI designs converge toward the all-SOI counterparts. Similarly, the low leakage of TFETs brings the most benefit for very slow designs lying to the right of the DP curve, so that the HGI incorporation of SOI brings negligible benefits. We note that the finite number of bins in our study discretizes the usage of different devices in our HGI designs, limiting the resolution of the latter. For this reason, the HGI results merge with those of non-HGI circuits in the high- and low-performance limits of Fig. 16(a). If the LDH is divided into a larger, near continuous number of bins, allowing for finer grained designs, the advantages of HGI would become manifest for all operating regions because any small design improvements due to incremental TFET or SOI usage can be evaluated. However, our four bin results and the intuitive arguments above suggest that only small improvements in the very high- and low-performance regimes would come from HGI for the particular devices under study. Moreover, any performance improvements must be weighed against the accompanying fabrication costs of the more complicated HGI process flow, such that the substantial performance enhancements are necessary to justify HGI in practice.

Although we observe obvious advantages for HGI during delay and power optimization, the tradeoff becomes more complicated if the design area is also considered. The areas corresponding to the designs in Fig. 16(a) are shown in Fig. 16(b). We observe that compared with all-SOI designs, HGI requires more area even when it consumes less power at a given delay. This is because SOI devices have strong driving current and can, therefore, be sized relatively smaller. For the same reason, HGI designs require lesser area than all-TFET designs by utilizing some proportion of the smaller SOI gates. By contrast, for very long and short delay periods (corresponding to the leftmost and the rightmost regions of Fig. 16), the HGI optimization leads to all-SOI and all-TFET designs, so the corresponding design area also converges with the non-HGI cases. By quantifying the design area tradeoffs, PROCEED shows that HGI designs receive few benefits in this case because most of the area is consumed by the large number of slow TFET gates.

V. Conclusion

The proposed circuit-device co-evaluation framework[2] accounts for circuit topology, adaptability, variability, and use context using efficient Pareto optimization heuristic. Device evaluation ignoring one or more crucial factors, such as multiple supply and threshold voltages, power management, logic depth, and variability, can give misleading results. For instance, we find that including power management in our evaluation can effectively double the usable operating range for TFETs, and that choice of activity factor can dictate whether TFETs are acceptable at all in a given application. The metrics applied in PROCEED, including delay-power and delay-area tradeoffs, enables a comprehensive comparison of the benefits and shortcomings of various devices. In addition, we demonstrate how PROCEED enables fast, realistic evaluation of HGI using TFET and SOI technologies as an example. These observations are made possible by PROCEED's scope and computational efficiency in studying several orders of magnitude in possible device-circuit performance, and demonstrate the capability and flexibility of our new methodology.

References

[1] L. Wei, S. Oh, and H.-S. P. Wong, "Performance benchmarks for Si, III–V, TFET, and carbon nanotube FET—Re-thinking the technology assessment methodology for complementary logic applications," in *Proc. IEEE Int. Electron Devices Meeting (IEDM)*, Dec. 2010, pp. 16.2.1–16.2.4.

[2] L. Wei and D. Antoniadis, "CMOS device design and optimization from a perspective of circuit-level energy-delay optimization," in *Proc. IEEE Int. Electron Devices Meeting (IEDM)*, Dec. 2011, pp. 15.3.1–15.3.4.

[3] P. M. Solomon, D. J. Frank, and S. O. Koswatta, "Compact model and performance estimation for tunneling nanowire FET," in *Proc. DRC*, Jun. 2011, pp. 197–198.

[4] D. J. Frank, W. Haensch, G. Shahidi, and O. H. Dokumaci, "Optimizing CMOS technology for maximum performance," *IBM J. Res. Develop.*, vol. 50, no. 4.5, pp. 419–431, Jul. 2006.

[5] D. E. Nikonov and I. A. Young, "Overview of beyond-CMOS devices and a uniform methodology for their benchmarking," *Proc. IEEE*, vol. 101, no. 12, pp. 2498–2533, Dec. 2013.

[6] D. Sylvester and K. Keutzer, "System-level performance modeling with BACPAC—Berkeley advanced chip performance calculator," in *Proc. SLIP*, 1999, pp. 109–114.

[7] M. Luisier, M. Lundstrom, D. A. Antoniadis, and J. Bokor, "Ultimate device scaling: Intrinsic performance comparisons of carbon-based, InGaAs, and Si field-effect transistors for 5 nm gate length," in *Proc. IEEE Int. Electron Devices Meeting (IEDM)*, Dec. 2011, pp. 11.2.1–11.2.4.

[8] H. Kam, T.-J. King-Liu, E. Alon, and M. Horowitz, "Circuit-level requirements for MOSFET-replacement devices," in *Proc. IEEE Int. Electron Devices Meeting (IEDM)*, Dec. 2008, p. 1.

[9] C. Augustine, A. Raychowdhury, Y. Gao, M. Lundstrom, and K. Roy, "PETE: A device/circuit analysis framework for evaluation and comparison of charge based emerging devices," in *Proc. ISQED*, Mar. 2009, pp. 80–85.

[10] M. Cotter, H. Liu, S. Datta, and V. Narayanan, "Evaluation of tunnel FET-based flip-flop designs for low power, high performance applications," in *Proc. 14th Int. Symp. Quality Electron. Design (ISQED)*, Mar. 2013, pp. 430–437.

[11] G. Cho, Y.-B. Kim, and F. Lombardi, "Assessment of CNTFET based circuit performance and robustness to PVT variations," in *Proc. 52nd IEEE Int. Midwest Symp. Circuits Syst. (MWSCAS)*, Aug. 2009, pp. 1106–1109.

[12] C. Pan and A. Naeemi, "System-level optimization and benchmarking of graphene PN junction logic system based on empirical CPI model," in *Proc. IEEE Int. Conf. IC Design Technol. (ICICDT)*, May/Jun. 2012, pp. 1–5.

[13] K. Swaminathan, H. Liu, J. Sampson, and V. Narayanan, "An examination of the architecture and system-level tradeoffs of employing steep slope devices in 3D CMPs," in *Proc. ACM/IEEE 41st Int. Symp. Comput. Archit. (ISCA)*, Jun. 2014, pp. 241–252.

[14] Z. Li, J. Tan, and X. Fu, "Hybrid CMOS-TFET based register files for energy-efficient GPGPUs," in *Proc. 14th Int. Symp. Qual. Electron. Design (ISQED)*, Mar. 2013, pp. 112–119.

[15] K. Swaminathan *et al.*, "Modeling steep slope devices: From circuits to architectures," in *Proc. Design, Autom. Test Eur. Conf. Exhibit. (DATE)*, Mar. 2014, pp. 1–6.

[16] K. Swaminathan, H. Liu, X. Li, M. S. Kim, J. Sampson, and V. Narayanan, "Steep slope devices: Enabling new architectural paradigms," in *Proc. 51st ACM/EDAC/IEEE Design Autom. Conf. (DAC)*, Jun. 2014, pp. 1–6.

[17] S. Wang, A. Pan, C. O. Chui, and P. Gupta, "PROCEED: A Pareto optimization-based circuit-level evaluator for emerging devices," in *Proc. 19th Asia South Pacific Design Autom. Conf. (ASP-DAC)*, Singapore, Jan. 2014, pp. 818–824.

[18] R. S. Ghaida and P. Gupta, "DRE: A framework for early co-evaluation of design rules, technology choices, and layout methodologies," *IEEE Trans. Comput.-Aided Design Integr. Circuits Syst.*, vol. 31, no. 9, pp. 1379–1392, Sep. 2012.

[19] A. M. Ionescu and H. Riel, "Tunnel field-effect transistors as energy-efficient electronic switches," *Nature*, vol. 479, no. 7373, pp. 329–337, 2011.

[20] W.-C. Wang and P. Gupta, "Efficient layout generation and evaluation of vertical channel devices," in *Proc. IEEE/ACM Int. Conf. Comput.-Aided Design*, Nov. 2014, pp. 550–556.

[21] H. Liu, S. Datta, and V. Narayanan, "Steep switching tunnel FET: A promise to extend the energy efficient roadmap for post-CMOS digital and analog/RF applications," in *Proc. ISLPED*, Sep. 2013, pp. 145–150.

[22] C. O. Chui, K.-S. Shin, J. Kina, K.-H. Shih, P. Narayanan, and C. A. Moritz, "Heterogeneous integration of epitaxial nanostructures: Strategies and application drivers," *Proc. SPIE*, vol. 8467, p. 84670R, Oct. 2012.

[23] J. A. Davis, V. K. De, and J. D. Meindl, "A stochastic wire-length distribution for gigascale integration (GSI). II. Applications to clock frequency, power dissipation, and chip size estimation," *IEEE Trans. Electron Devices*, vol. 45, no. 3, pp. 590–597, Mar. 1998.

[24] J.-H. Ryu, S. Kim, and H. Wan, "Pareto front approximation with adaptive weighted sum method in multiobjective simulation optimization," in *Proc. Winter Simulation Conf. (WSC)*, Dec. 2009, pp. 623–633.

[2] PROCEED is available at http://nanocad.ee.ucla.edu/pub/Main/DownloadForm/PROCEED.tar.gz

[25] J. P. Benson *et al.*, "Car-park management using wireless sensor networks," in *Proc. 31st IEEE Conf. Local Comput. Netw.*, Nov. 2006, pp. 588–595.

[26] T. He *et al.*, "Achieving real-time target tracking using wireless sensor networks," in *Proc. RTAS Symp.*, Apr. 2006, pp. 37–48.

[27] S. Boyd and L. Vandenberghe, *Convex Optimization*. Cambridge, U.K.: Cambridge Univ. Press, 2004.

[28] A. Pan and C. O. Chui, "A quasi-analytical model for double-gate tunneling field-effect transistors," *IEEE Electron Device Lett.*, vol. 33, no. 10, pp. 1468–1470, Oct. 2012.

[29] A. Pan, S. Chen, and C. O. Chui, "Electrostatic modeling and insights regarding multigate lateral tunneling transistors," *IEEE Trans. Electron Devices*, vol. 60, no. 9, pp. 2712–2720, Sep. 2013.

[30] K. Jeon *et al.*, "Si tunnel transistors with a novel silicided source and 46mV/dec swing," in *Proc. Symp. VLSI Technol. (VLSIT)*, Jun. 2010, pp. 121–122.

[31] G. Leung and C. O. Chui, "Stochastic variability in silicon double-gate lateral tunnel field-effect transistors," *IEEE Trans. Electron Devices*, vol. 60, no. 1, pp. 84–91, Jan. 2013.

[32] G. Leung and C. O. Chui, "Interactions between line edge roughness and random dopant fluctuation in nonplanar field-effect transistor variability," *IEEE Trans. Electron Devices*, vol. 60, no. 10, pp. 3277–3284, Oct. 2013.

Shaodi Wang received the B.S. degree from Peking University, Beijing, China, and the M.S degree in electrical engineering from the University of California at Los Angeles (UCLA), Los Angeles, CA, USA. He is currently pursuing the Ph.D. degree with the Department of Electrical Engineering, UCLA.

His current research interests include emerging memory and device technology and modeling for manufacturing.

Andrew Pan (S'12) received the B.S. degree in physics and the M.S. degree in electrical engineering from the University of California at Los Angeles, Los Angeles, CA, USA, where he is currently pursuing the Ph.D. degree in electrical engineering.

His current research interests include semiconductor device modeling, transport phenomena, and solid-state physics.

Chi On Chui (S'00–M'04–SM'08) received the B.Eng. degree in electronic engineering from the Hong Kong University of Science and Technology, Hong Kong, in 1999, and the M.S. and Ph.D. degrees in electrical engineering from Stanford University, Stanford, CA, USA, in 2001 and 2004, respectively.

He joined Intel Corporation, Santa Clara, CA, USA, in 2004, as a Senior Device Engineer. During his tenure with Intel, he also served as a Researcher-in-Residence with the University of California at Berkeley, Berkeley, CA, USA, and Stanford University. From 2005 to 2006, he was also appointed as a Consulting Assistant Professor of Electrical Engineering with Stanford University. Since 2007, he has been with the University of California at Los Angeles (UCLA), Los Angeles, CA, USA, where he is currently an Associate Professor of Electrical Engineering and Bioengineering, and a member of the California NanoSystems Institute, UCLA. He has authored and co-authored over 125 peer-reviewed and invited archival journal and conference papers, and six book chapters. He holds nine issued patents. His current research interests include nanoelectronic device-circuit interaction, heterogeneous integration technology, and biomedical devices.

Dr. Chui has served on the Technical Program Committees and International Advisory Committees of several conferences on electronic devices and circuits. His work has received three Best Paper Awards. He received the Okawa Foundation Research Grant in 2007. He was the first recipient of the IEEE Electron Device Society Early Career Award in 2009, which is regarded as one of the society's highest honors. In 2011, he received the Chinese-American Faculty Association Robert T. Poe Faculty Development Award, the UCLA Faculty Career Development Award, and the University of California at San Diego's von Liebig Entrepreneurism Center Regional Health Care Innovation Challenge Award. He also received the UCLA Henry Samueli School of Engineering and Applied Science Northrop Grumman Excellence in Teaching Award in 2011.

Puneet Gupta received the B.Tech degree in electrical engineering from IIT Delhi, New Delhi, India, in 2000, and the Ph.D. in 2007 from University of California at San Diego, San Diego, CA, USA.

He is currently a Faculty Member with the Department of Electrical Engineering, University of California at Los Angeles, Los Angeles, CA, USA. He co-founded Blaze DFM Inc. (acquired by Tela Inc.) in 2004 and served as its product architect till 2007. He has authored over 100 papers, 16 U.S. patents, a book, and a book chapter.

He was a recipient of an NSF CAREER Award, the ACM/SIGDA Outstanding New Faculty Award, the SRC Inventor Recognition Award, and the IBM Faculty Award. He currently leads the IMPACT+ Center, which focuses on future semiconductor technologies. His current research interests include building high-value bridges across application architecture implementation-fabrication interfaces for lowered cost and power, increased yield, and improved predictability of integrated circuits and systems.