Generic System-Level Modeling and Optimization for Beyond CMOS Device Applications

Victor Huang, Chenyun Pan, and Azad Naeemi
Georgia Institute of Technology, 791 Atlantic Drive NW, Atlanta, GA 30332, USA
E-mail: vhuang@gatech.edu

Abstract

In this work, a fast, generic system-level design and optimization methodology is presented for futuristic devices. This work evaluates GaN Heterojunction TFET, WTe_2 Two-dimensional heterojunction interlayer TFET (ThinTFET), and WTe_2 Transition Metal Dichalcogenide TFET (TMD TFET) in terms of performance and energy-delay product (EDP). This study investigates the impact of device-level performance on the system-level performance and power dissipation. The system-level methodology uses a generic model that utilizes a stochastic wire distribution to estimate system performance. An optimum supply voltage and gate count to achieve maximum throughput is examined for each of the devices using an empirical CPI model under different power budget constraints. Based on this study, the optimal design of each beyond-CMOS device technology is demonstrated to improve EDP. Results in this work delineate an optimal EDP for a given range of power budgets, and provides insightful trends on key design parameters as well as optimal performance and power metrics based on the fast system-level optimization at the early design stage.

Keywords

Beyond-CMOS, empirical CPI, system-level design methodology, tunneling field-effect transistor

1. Introduction

As traditional IC technology reaches fundamental scaling limits due to electron thermal energy and undesired tunneling currents, new classes of devices are being explored as potential alternatives to achieve optimal device performance and energy consumption. For low-power applications, in recent literature, the tunneling field-effect transistor (TFET) device architecture has gained a lot of popularity due to its low leakage properties.

TFETs switch by modulating a barrier width, controlling tunneling currents through a barrier instead of modulating barrier height as in conventional field-effect transistors. They promise low leakage, steep subthreshold slope and low supply voltage, but also have low on-current. In addition, TFETs generally have a larger footprint compared to CMOS and have unidirectional current flow. With these potential benefits and drawbacks, it becomes important to understand how these devices would perform compared to conventional CMOS technology.

Recent efforts in benchmarking these new devices have evaluated performance of the energy and delay for 32-bit adders and Arithmetic Logic Unit (ALU) for Beyond CMOS Benchmarking (BCB) [1, 2]. Current benchmarking models are good for 32-bit ALU, but it is architecture and circuit specific and does not consider area constraints and power budgets. A system level approach will extend this study to evaluate and optimize system performance for a single logic core. A more general system model is useful to uniformly compare different technologies with different architecture and complexity, allowing it to easily scale to more complex logic cores where these devices will ultimately be used. With a flexible system model, throughput can be optimized by finding optimal supply voltage and number of gates, which represents a system architecture complexity and functionality. Using this optimization process, the impacts of different power budgets on the optimized throughput performance in terms of energy-delay product (EDP) can be evaluated.

Using the generic system model approach, this paper uniformly models and optimizes three promising TFET devices, GaN Heterojunction TFET, WTe_2 Two-dimensional heterojunction interlayer TFET (ThinTFET), and WTe_2 Transition Metal Dichalcogenide TFET (TMD TFET), and compares their system performance with ITRS projections for CMOS high performance (CMOSHP) and low voltage (CMOSLV) devices. The rest of the sections are organized as follows. Section 2 details the generic system modeling methodology and optimization. Section 3 discusses the simulation results and trends, and section 4 concludes the paper.

2. System Modeling Methodology

2.1. Generic System Model

A generic system model is used to quickly estimate the system level performance of different technology nodes. Existing system model IntSim is modified and used to model the power performance for each device technology. Intsim is an interconnect CAD tool that estimates pitch for different wiring levels, co-optimizes signal, power, and clock interconnects, and stochastically derives wiring distributions. It also provides estimates for the system level power consumption for a given set of system parameters [4]. For a given operating frequency target, interconnect networks are optimized to obtain metal pitches on different metal levels for a set of system parameters.

The advantage of using such a model is that it is a fast, generic model, captures system parameters and power, and has been validated with commercially available CPUs. While it is not as accurate as physical design models, it provides insightful trends and starting design parameters.

2.2. Empirical CPI Model

For this work, an empirical cycles per instruction (CPI) model is used in conjunction with the generic system model

to calculate the system throughput based on the number of transistors used in the system. The empirical CPI is based on the observations that a power law relation exists between CPI and the number of logic transistors. Previous works have shown a power-law relationship between number of logic gates and cycles per instructions. This has been verified based on 8 Intel processors using data extraction from existing CPU and CPU benchmarking specification SPECint. An updated CPI model is used in this study for the Intel microprocessor family [10].

$$CPI_{logic} = 2.466(N_{transistor})^{-0.420} \quad (1)$$

where $N_{transistor}$ is the number of logic transistors in millions.

The functionality of the system can be improved by increasing the number of transistors, leading to a smaller CPI; however, at the same time the more complex system requires more interconnects, which imposes more constraints on the maximum frequency at which the system can operate. Therefore, when the empirical CPI model is combined and the area is fixed, there is a tradeoff between system's operating frequency and CPI that gives an optimal throughput.

2.3. Optimization Methodology Flow

For the system model in this study, IntSim is used to predict the optimal operating frequency for a given supply voltage (V_{dd}) and number of gates (N_{gates}). The maximum frequency that can be successfully routed while staying within power budget is estimated and the maximum throughput is calculated using the empirical CPI model.

At the core of the parameter optimization is supply voltage and number of logic gates. Supply voltage controls the on-current for the device and governs the system operating frequency, while the number of gates impacts our CPI. For a fixed supply voltage and number of gates, the highest operating frequency will give us the highest throughput for these two design points. The goal is to find a valid system model that operates at the highest frequency within a given power budget.

By sweeping V_{dd} and N_{gates}, the parameters that maximizes throughput for a given power budget and design space can be found. When looking at different power budgets, the different constraints impact the performance and a comparison is made with different device technologies.

2.4. Input Data and Device Technologies

The system model requires input data for on-current, off-current and input capacitance for different device inputs. This work evaluates sidewall-gated GaN/InN heterojunction TFET (GaNTFET) [5, 6], WTe_2 Two-dimensional heterojunction interlayer TFET (ThinTFET) [7], and WTe_2 Transition Metal Dichalcogenide TFET (TMDTFET) [8, 9] and compares system performance with conventional CMOSHP and CMOSLV devices. All data are kept consistent with the physical dimensions presented in previous Beyond CMOS Benchmarking (BCB) works [1, 2] and ITRS Roadmap for the 2018 node [3].

The IV curves (Fig. 1) and input capacitances for the evaluated devices are taken from published sources [5-9]. To optimize V_{dd} for a given power budget, the full IV curve for I_{on} and I_{off} along with voltage dependent input capacitance for multiple supply voltage data points are extracted from these works.

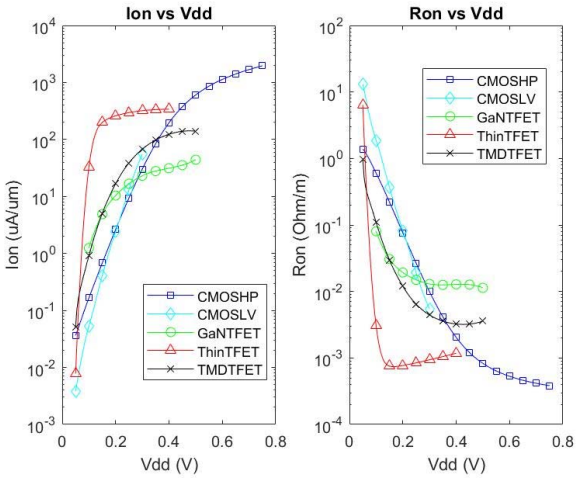

Figure 1: Input on-current and on-resistance data for different device technologies

3. Simulation Results

3.1. System Model Comparison of ALU and Single Logic Core

A comparison between the BCB 3.0 simulator [12] and the generic system model is made for a 32-bit ALU. See Fig. 2 for the comparison of energy and delay between the two models. In general, the two models show similar trends, with the generic system models more optimistic in energy for less complex systems.

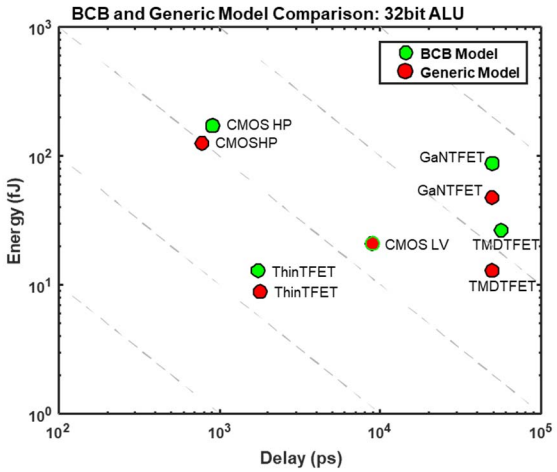

Figure 2: Trend comparison of Energy vs Delay between BCB model with Generic System model for a 32-bit ALU.

3.2. Power Breakdown

The power breakdown for a 32-bit ALU and single logic core is evaluated and shown in Fig. 3. For evaluating the system model extension of the ALU to a single logic core,

the same device input is used. The key parameters used in the model are shown in Table 1. The normalized power breakdown shows that as the circuit becomes larger and more complicated, system overhead starts to take a larger proportion of the power. Interconnect power doubles in proportion compared to the rest of the power breakdown. This comparison using the system model highlights the more critical role of interconnects and repeaters in more complex systems. Having a more flexible model that captures the interconnect network and optimizes it for different design points is important when looking at more complex system.

Table 1: Table of input parameters for system model comparison of ALU and Single Logic Core

Key Parameters	Values
ALU N_{gates}	1500
ALU Area [mm^2]	3.6×10^{-4}
Single Logic Core N_{gates} (Million)	16.3
Single Logic Core Area [mm^2]	3.9
Logic Depth	10
Power Budget Density [W/cm^2]	90
Activity Factor	0.1

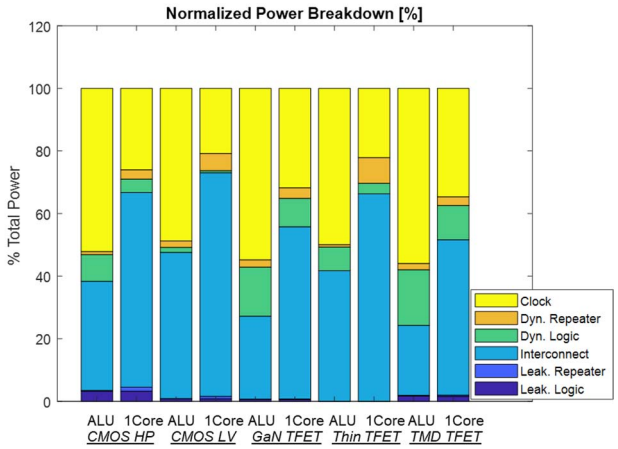

Figure 3: Power breakdown comparison for 32-bit ALU and single core processor (1Core) for different device technologies.

3.3. Throughput Optimization

By using the empirical CPI model, there is a tradeoff between number of gates and higher operating frequency due to large gate widths for a fixed area. This leads to an optimal number of gates. A fixed chip area of $5mm^2$ is used for the single logic core optimization. Throughput is limited at higher supply voltages due to power budget constraints, and an optimal V_{dd} can also be found for a given power budget. The optimization algorithm finds the highest throughput for a given V_{dd} and N_{gates} pair based on the empirical CPI model. This is done by finding the highest operating frequency that meets the power budget constraint for a given N_{gates} and V_{dd}. See Fig. 4 for a typical optimization result for CMOSHP.

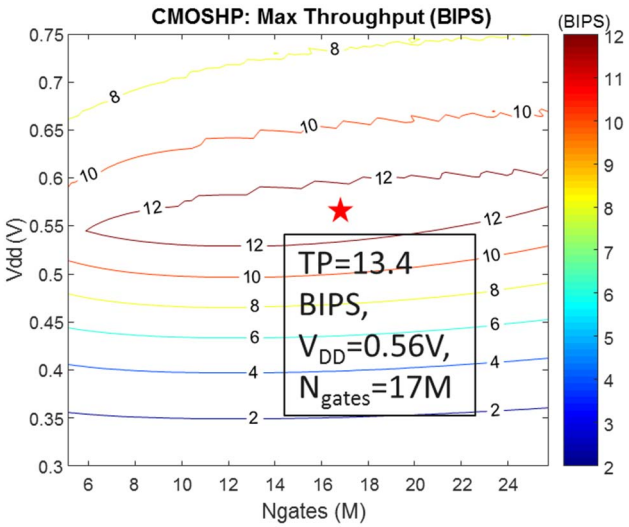

Figure 4: CMOS HP optimal throughput contour plot. Area is fixed at $5mm^2$. Power budget is set to 4.6 W, with a power density of $93W/cm^2$

3.4. Throughput vs Power Budget

The results of optimizing throughput for a range of power budgets are show in Fig. 5. The power budget limits the supply voltage and frequency the system can operate. Low-power TFET devices perform better in terms of throughput when compared to CMOS LV for low power applications (<0.1W, $2W/cm^2$). For high performance applications, CMOSHP still performs the best in terms of throughput at high power budgets (>2.5W, $50W/cm^2$). The optimal V_{dd} and N_{gates} at each point will be shown in the next subsection.

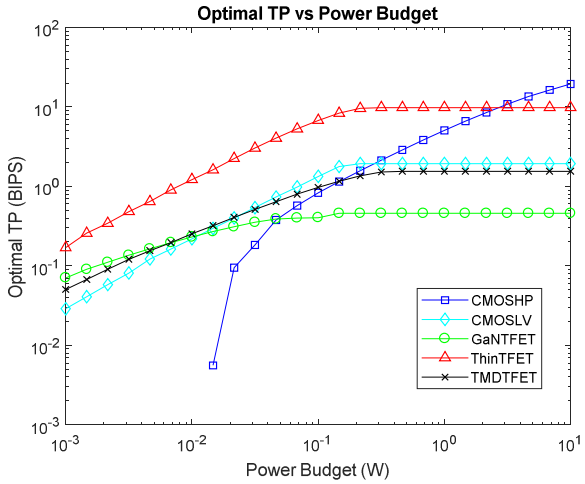

Figure 5: Optimal throughput result versus power budget for different device technologies

3.5. System Optimization Trends for V_{dd} and N_{gates}

When optimizing for throughput for different power budgets, a trend for V_{dd} and N_{gates} emerges. Fig. 6 shows the V_{dd} trend as power budget increases. At low power budgets, the system power is constrained and requires lower supply

voltages to satisfy the requirement. As power budget increases, the optimal supply voltage also increases to allow for higher on-currents and operating frequency. For the lower power TFET and CMOS devices, V_{dd} quickly saturates to the maximum value as throughput saturates. CMOSHP, however, continues to increase due to its larger V_{dd} range and higher on-currents. For all cases, the optimal supply voltage settles to a V_{dd} point that corresponds to its minimum R_{on}.

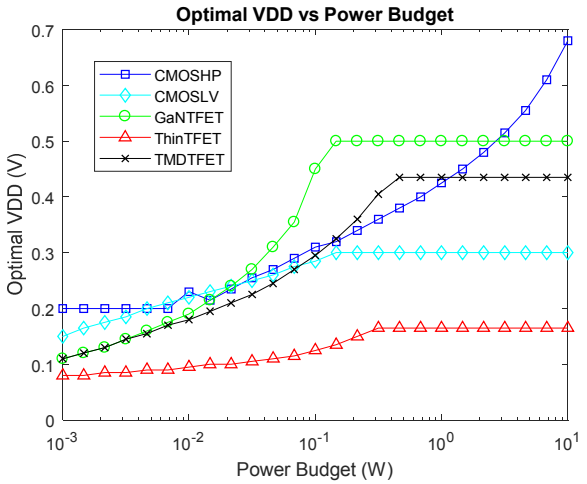

Figure 6: Optimal V_{dd} trends for different device technologies and power budgets

A trend for the optimal number of N_{gates} is shown in Fig 7. At low power budgets, the system model favors more gates for lower CPI, which allows throughput to increase by increasing the functionality of the system without increasing power significantly. As power budget increases, N_{gates} approaches the optimal value associated with the preferred V_{dd} for maximum throughput due to the tradeoff between higher N_{gates} and lower CPI versus lower N_{gates} and higher frequency.

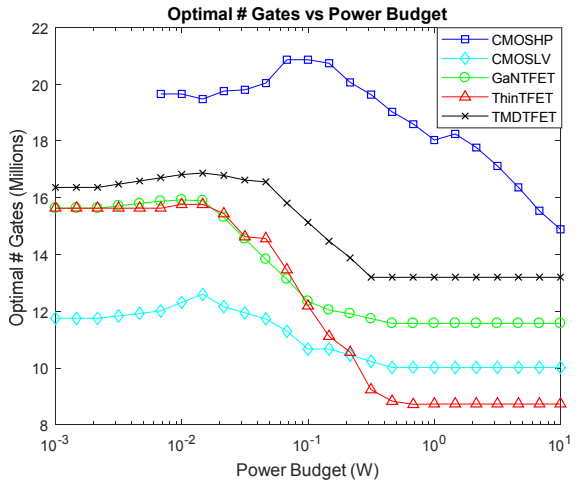

Figure 7: Optimal number of gates (N_{gates}) trend for different device technologies and power budgets.

3.6. Optimization with minimum EDP: Single Core vs Optimized Single Core

Using the system-level modeling approach in conjunction with the empirical CPI, the number of gates is optimized to maximize throughput for a range of power budgets. As the number of gates increases, the device width decreases when the area and gate density is fixed. This decreases the on-current and maximum frequency at which the system can operate. However, if larger devices are used to increase the system operating frequency, the number of gates is reduced, increasing the number of cycles per instruction. This tradeoff leads to an optimal number of gates to maximize system throughput.

The results of the optimization in terms of energy and delay are shown in Fig. 8 and 9. The minimum EDP is evaluated for a range of power budgets and the results are tabulated in Table 2. Overall improvements are made in terms of EDP due to the optimization of number of gates and supply voltage. For TFET devices, GaN TFET benefitted the most from the optimization of the single core with an improvement of 64% in EDP. This is primarily driven by reducing the power budget and operating at a lower frequency and supply voltage.

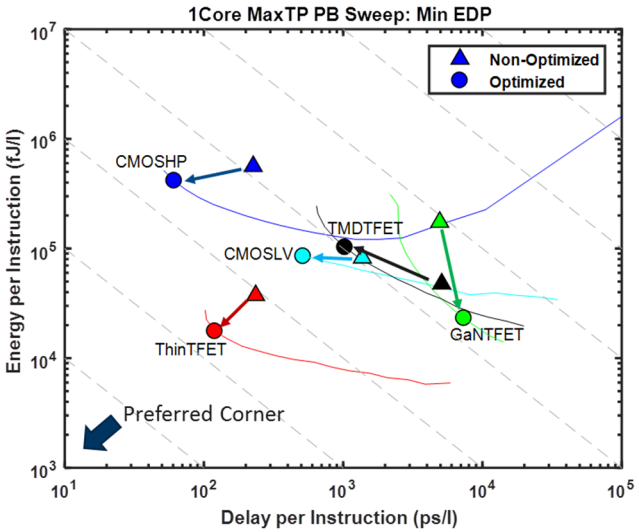

Figure 8: Energy vs Delay per instruction results for different device technologies. The solid lines represent the optimized results for a range of power budget. The circle indicates the optimized results that correspond to the minimum energy delay product for that range of power budgets. The lower left-hand is the preferred corner that corresponds to a lower energy delay product.

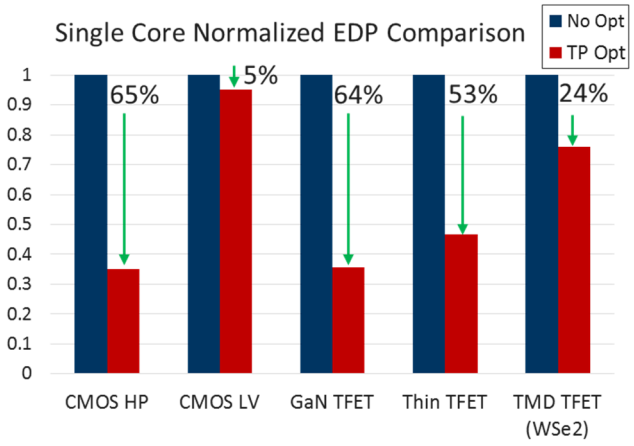

Figure 9: Normalized energy delay product comparison between the non-optimized case and the optimized case from a range of power budgets.

Table 2: Optimized minimum EDP results for a range of power budgets for different device technologies

Technology	**Optimal PD [W/cm^2]**	**Optimal V_{dd} [V]**	**Optimal N_{gates} [M]**
CMOS HP	92.83	0.56	16.9
CMOS LV	6.32	0.3	9.3
GaN TFET	0.06	0.15	19.1
Thin TFET	2.94	0.14	13.1
TMD TFET (WSe_2)	2	0.3	18.4

4. Conclusion

In this study, a fast system-level model is applied to three beyond CMOS devices and the system level performance is evaluated. The system model is compared with the beyond CMOS benchmarking approach and shows good agreement for the 32-bit ALU. The system level approach is applied for a single logic core evaluation, and the interconnect bottleneck is shown through the doubling in power in proportion to overall system. Optimization is performed for a single logic core analysis, and EDP is shown to improve up to 64% in the case for the sidewall-gated GaN/InN heterojunction TFET. In optimizing throughput for a range of power budgets, a trend in V_{dd} shows an increase as the optimal point becomes less constrained by the power density limits. Higher N_{gates} is favored at lower power budgets before decreasing in favor of high frequency at higher power budgets.

5. References

[1] Nikonov, Dmitri E., and Ian A. Young. "Benchmarking of beyond-CMOS exploratory devices for logic integrated circuits." IEEE Journal on Exploratory Solid-State Computational Devices and Circuits 1 (2015): 3-11.

[2] Nikonov, Dmitri E., and Ian A. Young. "Overview of beyond-CMOS devices and a uniform methodology for their benchmarking." Proceedings of the IEEE 101.12 (2013): 2498-2533.

[3] ITRS International Technology Roadmap for Semiconductors (ITRS) (2011). [Online]. Available: http://www.itrs2.net/.

[4] D. C. Sekar, A. Naeemi, R. Sarvari, J. A. Davis, and J. D. Meindl, "IntSim: A CAD tool for Optimization of Multi-level Interconnect Networks," in Proc. IEEE Int. Conf. on Computer-Aided Design, 2007.

[5] Li, Wenjun, et al. "Polarization-engineered III-nitride heterojunction tunnel field-effect transistors." IEEE Journal on Exploratory Solid-State Computational Devices and Circuits 1 (2015): 28-34.

[6] Lu, Hao, et al. "Universal charge-conserving TFET SPICE model incorporating gate current and noise." IEEE Journal on Exploratory Solid-State Computational Devices and Circuits 2 (2016): 20-27.

[7] Li, Mingda Oscar, et al. "Two-dimensional heterojunction interlayer tunneling field effect transistors (Thin-TFETs)." IEEE Journal of the Electron Devices Society 3.3 (2015): 200-207.

[8] Ilatikhameneh, Hesameddin, et al. "Tunnel field-effect transistors in 2-D transition metal dichalcogenide materials." IEEE Journal on Exploratory Solid-State Computational Devices and Circuits 1 (2015): 12-18.

[9] Ilatikhameneh, Hesameddin, Gerhard Klimeck, and Rajib Rahman. "2D tunnel transistors for ultra-low power applications: Promises and challenges." Energy Efficient Electronic Systems (E3S), 2015 Fourth Berkeley Symposium on. IEEE, 2015.

[10] Pan, Chenyun, and Azad Naeemi. "A fast system-level design methodology for heterogeneous multi-core processors using emerging technologies." IEEE Journal on Emerging and Selected Topics i Circuits and Systems 5.1 (2015): 75-87.

[11] Predictive Technology Model (PTM) [Online]. Available: http://ptm.asu.edu, 2012

[12] Nikonov, D. E. Benchmarking of devices in the nanoelectronics research initiative. (2014). [Online]. Available: https://nanohub.org/tools/nribench/browser/trunk/src/.