Optimal Choice of FinFET Devices for Energy Minimization in Deeply-Scaled Technologies

Mohammad Saeed Abrishami, Alireza Shafaei, Yanzhi Wang, and Massoud Pedram
Department of Electrical Engineering
University of Southern California
Los Angeles, CA 90089
{abri442, shafaeib, yanzhiwa, pedram}@usc.edu

Abstract—**FinFET devices are considered to be the device substitute for bulk CMOS in sub-20nm technology nodes due to the reduced short-channel effects, improved ON/OFF current ratio, and improved voltage scalability. This paper investigates the problem of optimal selection of deeply-scaled FinFET technology to achieve minimum energy consumption for different applications such as sensor applications, smartphones, embedded micro-processors, or server micro-processors, which differ in the required performance and duty ratio. For each application space, different FinFET technologies (with different V_{th} and gate length biases) are compared in term of minimum energy consumption for both logic circuits and cache memories. A device-circuit-architecture cross-layer framework has been developed to facilitate this technology selection. This optimal technology selection procedure demonstrates up to $11\times$ energy saving compared to poorly selected technologies.**

I. INTRODUCTION

The steady down-scaling of feature size of bulk CMOS technology has resulted in various *short-channel effects* (SCEs), such as *Drain Induced Barrier Lowering* (DIBL) and V_{th} roll-off effect [1]. The SCEs degrade the expected power efficiency achieved by the further scaling of bulk CMOS transistor in deep-submicron regions [1], [2]. The multi-gate or tri-gate transistor structures such as FinFETs have been proposed to rejuvenate the chip industry from SCEs [3], [4]. The improved electrostatic integrity of FinFET devices can alleviate SCEs and achieve higher voltage scalability to improve power efficiency [3], [5]. It has been reported that FinFET devices are estimated to be up to 37% faster while consuming less than half the dynamic power or reduce the leakage current by up to 90% compared to bulk CMOS devices [6]. Besides, the absence of channel doping in FinFETs will eliminate the random dopant fluctuation, which is a major source of process-induced variations in conventional CMOS technology [7]. Therefore, FinFETs are promising device candidates for bulk CMOS at the 22nm technology node and beyond [4], [6]. For a specific deeply-scaled FinFET technology, the V_{th} could be adjusted through gate work-function engineering [8] and the gate length could be adjusted by using gate-length biasing technique [9].

Different applications exist ranging from low-power and low-duty ratio sensor applications to smartphone applications, and from embedded micro-processors to high-performance server micro-processors [10]. Various applications differ from each other mainly in two factors: required performance (clock frequency) and duty ratio, in which duty ratio is defined as the ratio of active time to the total time. However, it remains unexplored on the optimal technology selection of deeply-scaled FinFET technologies for different application types, or more specifically, what are the best-suited V_{th} and gate length values of FinFET devices for each type of application and what is the optimal corresponding supply voltage level V_{dd}? For example, the low-performance and low-duty ratio sensor application prefers a higher-V_{th} (due to reduced leakage) and lower-V_{dd} (due to reduced switching power consumption) FinFET technology. On the other hand, a high-performance and high-duty ratio server application prefers a lower-V_{th} and higher-V_{dd} FinFET technology due to the enhanced speed.

In this paper, we investigate the problem of optimal selection of deeply-scaled FinFET technology to achieve minimum energy consumption. We develop a device-circuit-architecture cross-layer framework by (i) designing and optimizing deeply-scaled (7nm) FinFET devices [11] with different V_{th} and gate length biasing values using Synopsys TCAD suite [12], (ii) extracting Verilog-A formats that are compatible to SPICE simulation for each type of FinFET device for fast circuit-level simulation, (iii) and modifying the CACTI tool [13] for cache memory modeling by adding support for deeply-scaled FinFET devices. In order to compare different technologies for optimal selection, we define distinct application spaces according to their required performances and duty ratios. Then all different FinFET technologies (with different V_{th} values and gate length biases) are compared in term of minimum energy consumption for both logic circuits and cache memories. In this comparison, the supply voltage of logic circuits is properly set to meet the required performance level, but cannot be reduced below the *minimum energy point* since it will be resulted in higher energy consumption. This optimal technology selection procedure demonstrates up to $11\times$ energy saving compared to poorly selected technologies.
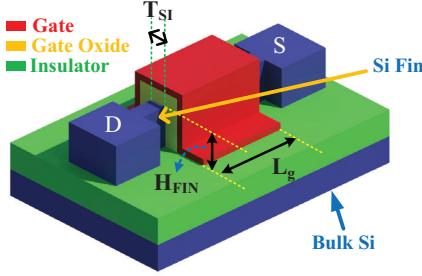
Fig. 1.   Structure of a FinFET device.

TABLE I.   SPECIFICATIONS OF 7NM FINFET PROCESS TECHNOLOGY.

Parameter	Value	Comment
L	2λ = 7nm	Gate length
T_{SI}	3.5nm	Fin width
H_{FIN}	14nm	Fin height
P_{FIN}	$2\lambda + T_{SI}$ = 10.5nm	Fin pitch
t_{ox}	1.3nm	Oxide thickness
V_{DD}	0.45V	Nominal supply voltage at the super-threshold regime

II.   FINFET BASICS AND OUR FINFET DEVICES

A. 7nm Gate Length FinFET Devices

Figure 1 illustrates the quasi-planar structure of a three-terminal FinFET device. This structure allows FinFET devices to enhance power efficiency, ON/OFF current ratio, as well as random variation and soft-error immunity compared with bulk CMOS counterparts [3]. Consequently, the FinFET technology is currently viewed as the technology-of-choice for technology nodes below 22nm [4], [6]. The major component that distinguishes FinFET devices from bulk CMOS counterparts is the vertical *fin*, which provides the transistor channel. The fin is surrounded by the gate material, and thus, the gate terminal establishes a three dimensional control over the channel, which essentially enhances the gate control and reduces SCEs accordingly. The key geometric parameters of a FinFET device, which are related to the fin, include the fin height H_{FIN}, fin width (also known as silicon thickness) T_{SI}, and fin length L_{FIN} (cf. Figure 1). The effective channel width of a single fin is approximately equal to $2 \times H_{FIN}$, which is the minimum achievable channel width in a FinFET device. In order to increase the width (strength) of a FinFET device, more fins are added.

Due to the lack of industrial data for deeply-scaled FinFETs, we develop and optimize our own 7nm FinFET devices [11] using the Synopsys Sentaurus Tool Suite [12], the advanced multi-dimensional device simulator from the TCAD tool suite. Sentaurus Device utilizes various models such as carrier transport, bandgap, mobility, and quantization models, and accounts for quantum effects in order to simulate electrical and thermal characteristics of semiconductor devices. For this work, we have developed a 7nm FinFET process with geometries and nominal supply voltage listed in Table I, which is considered as the standard (STD) 7nm FinFET device.

B. FinFET Devices with Leakage Power Saving Techniques

Gate-Length Biasing: The nominal gate length L_G of our FinFET devices is 7nm, and in this work, we consider gate length biasing technique with increased gate lengths up to 9nm. The reason to choose 9nm as the upper bound on gate length is that significantly longer gate lengths are not layout swappable with nominal devices and may result in substantial *engineering*

change order overheads during layout design. Similar to the gate length biasing technique for CMOS technology [9], the relatively small gate length biases for FinFET devices can be achieved by slight modification on the layout. FinFETs with a longer gate length than 7nm will be referred to as LC devices in the rest of the paper.

Adjusting V_{th}**:** Unlike changing doping concentration to adjust the V_{th} value for CMOS devices, we engineer the work-function of gate materials to increase V_{th} of the FinFET devices [8]. The V_{th} of our standard FinFET device is 0.235V, and the V_{th} values of the two high-V_{th} versions, called HVT1 and HVT2, are 0.335V and 0.435V, respectively.

To sum up, we have generated standard FinFET devices with 0.235V V_{th} value and 7nm gate length using Synopsys Sentaurus Device. We have also generated a set of FinFET devices with increased (biased) gate lengths up to 9nm and standard V_{th} value, as well as two high-V_{th} FinFET devices with 7nm nominal gate length and increased V_{th} values of 0.335V and 0.435V. The naming conventions of all types of generated FinFET devices along with the characteristics of each device are summarized in Table II. Finally, we generate SPICE compatible Verilog-A models for all types of FinFET devices listed in Table II, which act as the interface between SPICE and the aforesaid FinFET device models. These SPICE-compatible Verilog-A models compared with the extremely slow device-level simulations allow us to perform relatively fast gate- and circuit-level simulations, and are subsequently utilized for our technology selection procedure to minimize energy for logic circuits and cache memories.

III.   APPLICATION SPACES AND MINIMUM ENERGY POINT OF DEEPLY-SCALED FINFET CIRCUITS

A. Application Space Classification

As shown in [10], the application space can be classified based on two metrics, required performance (clock frequency) and duty ratio, in which duty ratio is defined as the ratio of active time to the total time (sum of active time and standby/idle time.) By using these two metrics, the whole application space is classified into six categories as shown in Figure 2. The bottom left application space refers to sensor-type applications with very low duty ratio and low performance, including environmental sensor and implantable biomedical electronic devices. The duty ratio of this type is estimated around 0.001 - 0.01 [10]. The low required performance is likely to set the supply voltage to the *minimum energy point*, denoted by V_{min},

TABLE II.   Characteristics of our generated FinFET devices. STD, HVT, and LC denote the standard, high voltage threshold, and long channel devices, respectively.

| Device Name	Gate Length (nm)	Threshold Voltage (V)	ON Current (A/µm)		OFF Current (A/µm)		ON/OFF Current Ratio	
			NFET	PFET	NFET	PFET	NFET	PFET
STD	7	0.235	8.818e-04	5.504e-04	3.811e-08	5.782e-08	23,140	9,518
HVT1	7	0.335	4.139e-04	2.860e-04	1.627e-09	3.271e-09	254,390	87,444
HVT2	7	0.435	5.100e-05	6.100e-05	7.264e-11	2.060e-10	702,065	296,117
LC1	8	0.235	8.107e-04	5.061e-04	1.802e-08	2.631e-08	44,995	19,234
LC2	9	0.235	7.689e-04	4.768e-04	1.074e-08	1.517e-08	71,576	31,427
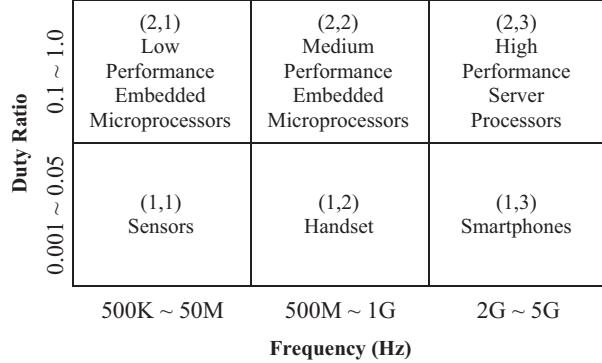
Fig. 2.   Classification of application spaces based on different performance (clock frequency) and duty ratio requirements.
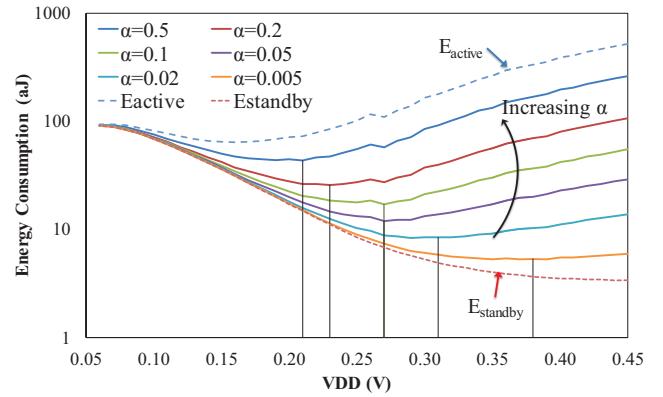

Fig. 3.   Active, standby, and total energy consumptions of 40-stage FO4 inverter chain for different V_{DD} values. Total energy consumption is measured for various activity factors. Vertical lines in the figure point to the V_{min}. Moreover, vertical axis is in logarithmic (base 10) scale.

at which the energy consumption per operation is minimized [7], [10]. Further reducing supply voltage lower than V_{min} will actually increase the energy consumption per operation because of the exponentially increasing delay in the sub/near-threshold region. The other five application spaces refer to handset applications, smartphones, low-performance embedded processors, medium-performance embedded processors, and high-performance server processors.

B. Minimum Energy Point of Deeply-Scaled FinFET Circuits

We test the energy consumption per operation of a 40-stage FO4 inverter chain using the STD device at different supply voltage levels, in order to find the V_{min}. Figure 3 illustrates the minimum energy point of the inverter chain at different activity factors (α). When α is higher than 0.2 (typical activity factor for a micro-processor), the minimum energy point is lower than $V_{th} = 0.235$V and lies in the subthreshold regime. When α is lower than 0.2, the minimum energy point lies in the near-threshold regime. Similarly, we derive the V_{min} values for the other four types of FinFET devices using the same method. Details are omitted due to space limitation.

IV.   Technology Selection for Energy Minimization

A. Logic Circuits and Cache Memory Modeling for Energy Comparison

Logic Circuits: For energy analysis and minimization, we model generic FinFET logic circuits by a 40-stage FO4 inverter

chain using a specific type of FinFET devices. Similar to [10], we simulate the inverter chain in SPICE to determine propagation delay and energy consumption. We only use clock gating during the standby mode in order to reduce the energy consumption. More efficient leakage saving techniques, such as power gating, are out of the scope of the current paper.

The nominal supply voltage of the FinFET circuits is 0.45V, but based on the performance requirements of each application space, an appropriate supply voltage is derived. The derived supply voltage should be larger than or equal to V_{min}, because as the supply voltage is reduced below V_{min}, we start losing both energy saving and performance. At the selected supply voltage level, the total energy consumption is comprised of three parts: (i) E_{switch} which is the switching energy consumption, (ii) E_{leak} which is the leakage power consumption within active cycles, and (iii) $E_{standby}$ which is the standby power consumption during idle time.

Cache Memory: In order to analyze and model the energy consumption of FinFET-based cache memories, we have modified the CACTI tool [13], which is a widely utilized architecture-level simulation tool for cache memory design and characterization. We have incorporated 7nm FinFET support into CACTI. More specifically, we (i) extracted process- and device-level parameters from Sentaurus Device, (ii) derived SRAM cell-level parameters (e.g. leakage current) from SPICE simulations using the Verilog-A models, and (iii) used most

recent ITRS predictions for interconnect scaling [14].

The nominal supply voltage used for FinFET-based cache memories is 0.45V, and the supply voltage will not scale down even if there is slack time in each clock cycle (i.e., when required performance is low) due to process variation and robustness considerations of SRAM cells. For each pair of required performance level and duty ratio, the total energy consumption calculation is similar to that of logic circuits, and is calculated for a 16KB, 2-way set-associative, 64B line, L1 cache memory.

B. Technology Selection for Energy Minimization in Logic Circuits

Figure 4 shows the optimal FinFET device that leads to the minimum total energy consumption for different application spaces. The application space covers a wide range of clock frequencies, from 500KHz to 5GHz, and duty ratios, from 0.001 to 1. We observe that the STD device is the optimal technology selection for high frequency and high duty ratio applications. The reason is that nominal devices for each technology node are typically designed and optimized for high performance applications in order to satisfy the increasing demand for faster digital computation. On the other hand, by moving towards lower clock frequencies or duty ratios, the standby energy consumption becomes the dominant component of the total energy consumption. Hence, FinFET devices optimized for leakage saving are becoming more favored in these applications. More precisely, from the top-right corner of Figure 4 where high performance applications stand, by lowering clock frequency or duty ratio, low leakage FinFET devices appear as the choice of technology in the same order of their associated OFF current, i.e., LC1, LC2, HVT1, and HVT2 (cf. Table II).

In order to evaluate the effectiveness of the optimal technology and V_{DD} selection procedure on reducing the total energy consumption, we consider the STD device operating at 0.45V as the baseline. Choosing the optimal FinFET device and V_{DD} level for different application spaces then results in $6\times$ on average energy reduction. Specifically, for very low performance applications, up to $11\times$ energy reduction is observed. In such low performance applications, $E_{standby}$ dominates the total energy consumption, and hence, using low leakage FinFET devices can significantly enhance the energy efficiency of circuits.

C. Technology Selection for Energy Minimization in Cache Memories

We also derived the optimal FinFET device that leads to the minimum energy cache memory for different application spaces, and the results are shown in Figure 5. The highest L1 cache clock frequency obtained by our FinFET devices is 2.9GHz (for STD device), so the 5GHz column is omitted in the memory results. For applications which access the memory very frequently, where E_{switch} is the dominant element, STD
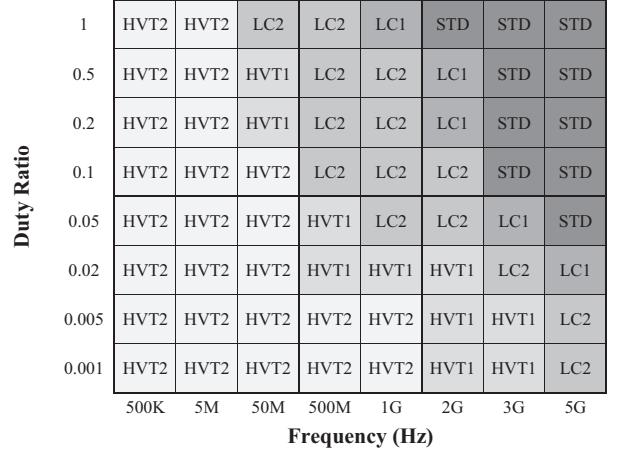
Duty Ratio	500K	5M	50M	500M	1G	2G	3G	5G
1	HVT2	HVT2	LC2	LC2	LC1	STD	STD	STD
0.5	HVT2	HVT2	HVT1	LC2	LC2	LC1	STD	STD
0.2	HVT2	HVT2	HVT1	LC2	LC2	LC1	STD	STD
0.1	HVT2	HVT2	HVT2	LC2	LC2	LC2	STD	STD
0.05	HVT2	HVT2	HVT2	HVT1	LC2	LC2	LC1	STD
0.02	HVT2	HVT2	HVT2	HVT1	HVT1	HVT1	LC2	LC1
0.005	HVT2	HVT2	HVT2	HVT2	HVT2	HVT1	HVT1	LC2
0.001	HVT2	HVT2	HVT2	HVT2	HVT2	HVT1	HVT1	LC2

Frequency (Hz)

Fig. 4. Optimal FinFET device for logic.

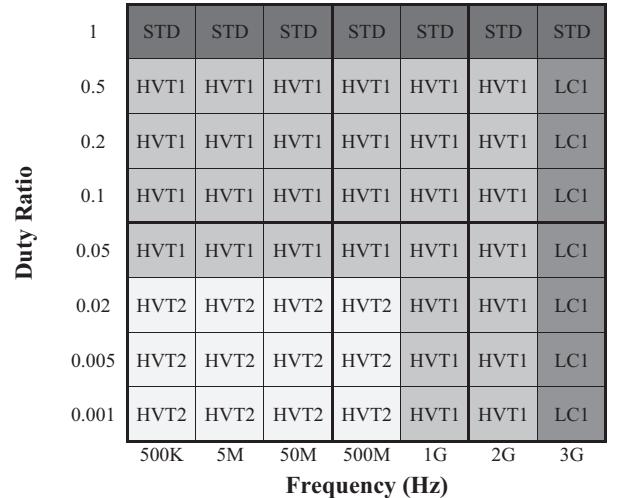| Duty Ratio	500K	5M	50M	500M	1G	2G	3G
1	STD	STD	STD	STD	STD	STD	STD
0.5	HVT1	HVT1	HVT1	HVT1	HVT1	HVT1	LC1
0.2	HVT1	HVT1	HVT1	HVT1	HVT1	HVT1	LC1
0.1	HVT1	HVT1	HVT1	HVT1	HVT1	HVT1	LC1
0.05	HVT1	HVT1	HVT1	HVT1	HVT1	HVT1	LC1
0.02	HVT2	HVT2	HVT2	HVT2	HVT1	HVT1	LC1
0.005	HVT2	HVT2	HVT2	HVT2	HVT1	HVT1	LC1
0.001	HVT2	HVT2	HVT2	HVT2	HVT1	HVT1	LC1

Frequency (Hz)

Fig. 5. Optimal FinFET device for cache memory.

device is the optimal choice. However, such applications are very rare, and if we ignore those results, the rest of optimal choices are among low leakage FinFETs. This is because of the large number of SRAM cells that are used in cache memories, which produce significant leakage current paths.

As a result, $E_{standby}$ becomes significantly important for cache memories. In order to minimize the cache energy consumption, we adopt hybrid cache designs where peripheral circuits and SRAM cells can take different device types. Generally, the cache access latency is mainly dependent on peripheral circuits, such as row decoder and wordline drivers, whereas cache standby energy significantly depends on the leakage current of SRAM cells. Hence, high speed devices for peripheral circuits, but low leakage devices for SRAM cells, are preferred. Results of the optimal FinFET selection for minimum energy hybrid cache design are shown in Figure 6, which confirm the effectiveness of hybrid cache designs.

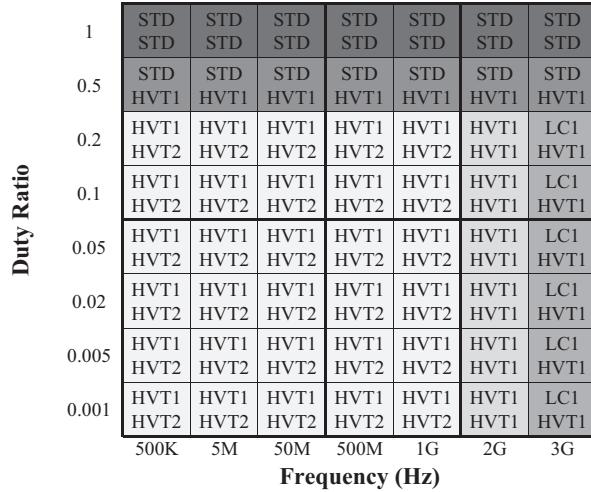| Duty Ratio	500K	5M	50M	500M	1G	2G	3G
1	STD STD	STD STD	STD STD	STD STD	STD STD	STD STD	STD STD
0.5	STD HVT1	STD HVT1	STD HVT1	STD HVT1	STD HVT1	STD HVT1	STD HVT1
0.2	HVT1 HVT2	HVT1 HVT2	HVT1 HVT2	HVT1 HVT2	HVT1 HVT2	HVT1 HVT1	LC1 HVT1
0.1	HVT1 HVT2	HVT1 HVT2	HVT1 HVT2	HVT1 HVT2	HVT1 HVT2	HVT1 HVT1	LC1 HVT1
0.05	HVT1 HVT2	HVT1 HVT2	HVT1 HVT2	HVT1 HVT2	HVT1 HVT2	HVT1 HVT1	LC1 HVT1
0.02	HVT1 HVT2	HVT1 HVT2	HVT1 HVT2	HVT1 HVT2	HVT1 HVT2	HVT1 HVT1	LC1 HVT1
0.005	HVT1 HVT2	HVT1 HVT2	HVT1 HVT2	HVT1 HVT2	HVT1 HVT2	HVT1 HVT1	LC1 HVT1
0.001	HVT1 HVT2	HVT1 HVT2	HVT1 HVT2	HVT1 HVT2	HVT1 HVT2	HVT1 HVT1	LC1 HVT1

Frequency (Hz)

Fig. 6. Optimal FinFET device for hybrid cache memory, where device selection of peripheral circuits and SRAM cells could be different. Device names on top and bottom of each cell denote the device optimal device selection for peripheral circuits and SRAM cells, respectively.

V. CONCLUSION

We analyzed the optimal selection of deeply-scaled FinFET technology to minimize energy consumption for different applications, which differ from each other in terms of the clock frequency and duty ratio. For each application type, we compared different FinFET devices for energy minimization for both logic circuits and cache memories. We developed a device-circuit-architecture cross-layer framework to facilitate the optimal technology selection, and demonstrated significant energy saving (up to $11\times$) through this optimal technology selection procedure.

VI. ACKNOWLEDGMENTS

REFERENCES

[1] P. Mishra, A. Bhoj, and N. Jha, "Die-Level Leakage Power Analysis of FinFET Circuits Considering Process Variations," in *International Symposium on Quality Electronic Design (ISQED)*, 2011.

[2] A. Bhoj and N. Jha, "Design of Ultra-Low-Leakage Logic Gates and Flip-Flops in High-Performance FinFET Technology," in *International Symposium on Quality Electronic Design (ISQED)*, 2012.

[3] S. Tang, L. Chang, N. Lindert, Y.-K. Choi, W.-C. Lee, X. Huang, V. Subramanian, J. Bokor, T.-J. King, and C. Hu, "FinFET - A Quasi-Planar Double-Gate MOSFET," in *IEEE International Solid-State Circuits Conference (ISSCC)*, 2001, pp. 118–119.

[4] E. Nowak, I. Aller, T. Ludwig, K. Kim, R. Joshi, C.-T. Chuang, K. Bernstein, and R. Puri, "Turning Silicon on its Edge [Double Gate CMOS/FinFET Technology]," *IEEE Circuits and Devices Magazine*, vol. 20, no. 1, pp. 20–31, 2004.

[5] J. Kedzierski, D. Fried, E. Nowak, T. Kanarsky, J. Rankin, H. Hanafi, W. Natzle, D. Boyd, Y. Zhang, R. Roy, J. Newbury, C. Yu, Q. Yang, P. Saunders, C. Willets, A. Johnson, S. P. Cole, H. E. Young, N. Carpenter, D. Rakowski, B. Rainey, P. Cottrell, M. Ieong, and H. S. P. Wong, "High-Performance Symmetric-Gate and CMOS-Compatible V_t Asymmetric-Gate FinFET Devices," in *Electron Devices Meeting, 2001. IEDM '01. Technical Digest. International*, 2001, pp. 19.5.1–19.5.4.

[6] Synopsys Insight Newsletter. [Online]. Available: http://www.synopsys.com/Company/Publications/SynopsysInsight/Pages/Art2-finfet-challenges-ip-IssQ3-12.aspx

[7] L. Chang and W. Haensch, "Near-Threshold Operation for Power-Efficient Computing? It Depends..." in *Design Automation Conference (DAC)*, June 2012.

[8] Y.-K. Choi, L. Chang, P. Ranade, J.-S. Lee, D. Ha, S. Balasubramanian, A. Agarwal, M. Ameen, T.-J. King, and J. Bokor, "FinFET Process Refinements for Improved Mobility and Gate Work Function Engineering," in *International Electron Devices Meeting (IEDM)*, Dec 2002, pp. 259–262.

[9] P. Gupta, A. Kahng, P. Sharma, and D. Sylvester, "Selective Gate-Length Biasing for Cost-Effective Runtime Leakage Control," in *Design Automation Conference (DAC)*, 2004.

[10] M. Seok, D. Sylvester, and D. Blaauw, "Optimal Technology Selection for Minimizing Energy and Variability in Low Voltage Applications," in *ACM/IEEE International Symposium on Low Power Electronics and Design (ISLPED)*, Aug 2008, pp. 9–14.

[11] S. Chen, Y. Wang, X. Lin, Q. Xie, and M. Pedram, "Performance Prediction for Multiple-Threshold 7nm-FinFET-based Circuits Operating in Multiple Voltage Regimes using a Cross-Layer Simulation Framework," in *IEEE SOI-3D-Subthreshold Microelectronics Technology Unified Conference (S3S)*, Oct. 2014.

[12] Synopsys Technology Computer-Aided Design (TCAD). [Online]. Available: http://www.synopsys.com/tools/tcad

[13] N. Muralimanohar, R. Balasubramonian, and N. Jouppi, "Optimizing NUCA Organizations and Wiring Alternatives for Large Caches with CACTI 6.0," in *40th Annual IEEE/ACM International Symposium on Microarchitecture (MICRO)*, Dec 2007, pp. 3–14.

[14] A. Shafaei, Y. Wang, X. Lin, and M. Pedram, "FinCACTI: Architectural Analysis and Modeling of Caches with Deeply-Scaled FinFET Devices," in *IEEE Computer Society Annual Symposium on VLSI (ISVLSI)*, July 2014, pp. 290–295.