Variation-Aware Joint Optimization of the Supply Voltage and Sleep Transistor Size for 7nm FinFET Technology

Qing Xie, Yanzhi Wang, Shuang Chen, and Massoud Pedram
Department of Electrical Engineering
University of Southern California
Los Angeles, California, United States
{xqing, yanzhiwa, shuangc, pedram}@usc.edu

***Abstract*—Power gating is a very effective method in reducing the leakage energy during the standby mode in VLSI circuits at the cost of increased circuit delay. This method has been well studied and widely used for circuits fabricated by using traditional CMOS technology nodes operating at super-threshold supply voltage regime. However, for advanced technology nodes with small feature sizes and low supply voltages, the propagation delay becomes very sensitive to the high process-induced variations. Therefore, this paper first analyzes how the circuit delay depends on the size of the sleep transistor under the process-induced variation for the 7nm gate length FinFET technology. Then a joint optimization problem is formulated to minimize the total energy consumption, while both supply voltage and sleep transistor size are considered as optimization variables. A near-optimal heuristic is presented to solve the optimization problem and determine the energy-optimal supply voltage and sleep transistor size. Experimental results based on HSPICE simulations show that more than 98% energy reduction for applications with relaxed deadline constraints after applying the joint optimization technique, compared to FinFET circuits without using the power gating method.**

I. INTRODUCTION

Power gating (PG) method has been proved to be a very effective method in reducing the leakage energy when circuits are in the standby mode. *Multi-threshold CMOS* (MTCMOS) technology is typically used to implement VLSI circuits with PG structures [1]. The MTCMOS technology provides low leakage and high performance operation by utilizing high speed, low threshold voltage transistors for logic cells and low leakage, high threshold voltage devices for sleep transistors. In the sleep mode, sleep transistors disconnect logic cells from the power supply and/or ground to reduce the leakage. Meanwhile, inserting sleep transistors also results in performance degradation due to the IR drop across sleep transistors in the active mode. Therefore, sizing sleep transistors requires consideration of both energy reduction as well as the performance degradation [1][2].

Previous research revealed that applying an ultra-low supply voltage reduces the total energy consumption per operation at the cost of sacrificing the VLSI circuits' timing performance [3][4]. In particular, for applications with relaxed timing requirements, *near-threshold* (NT) operation is quite effective in minimizing the energy consumption of a design by reducing its supply voltage to a level close to the threshold voltage of the transistors. Indeed, previous work on NT operation proved the existence of and analytically derived the *minimum energy (operation) point* (MEP), which is the optimal supply voltage level that minimizes the energy consumption [5][6].

One important drawback of operating digital circuits in the NT regime is the large impact of the process-induced variation. With the downscaling of transistor dimensions, the process-induced variation increases and, hence, has a greater impact on the circuit performance and yield. FinFET devices have been reported to offer superior properties such as lower gate leakage current [7], excellent control of short-channel effects [8], and relative immunization to gate line-edge roughness [9]. Future sub-20nm FinFET technology nodes are robust to t_{ox} variations and RDF due to its thin body [10], but sensitive to other sources of variability such as *Line-Edge Roughness* (LER) [10][11] and *metal-gate work function variations* [12]. Therefore, we consider process-induced variations in this work.

In this work, we analyze the dependency of propagation delay on the size of sleep transistor under process-induced variations and present a joint optimization method to determine the optimal supply voltage and sleep transistor size, in order to minimize the total energy consumption per operation. We first analyze the impact of the LER phenomenon for one of the most advanced FinFET technology nodes – 7nm gate length FinFET device model. Distributions of standard cell delay with respect to the channel length are obtained by performing Monte Carlo simulations. Then this dependency is fitted by using a lognormal distribution. We estimate propagation delays of complicated VLSI circuits under LER variations by first identifying critical timing paths of synthesized circuits and then statistically combining delays of all standard cells along the critical timing paths by using the Fenton-Wilkinson method. The energy consumption parameters related to dynamic switching and leakage are extracted through curving fitting with reasonable equations. In addition, we formulate a joint optimization problem, considering both supply voltage and sleep transistor size as optimization variables, to minimize the total energy consumption, including both dynamic and leakage energy. The optimization problem is subject to a deadline constraint, in which we account for the 0.1% *value at risk* point of delay distribution under the LER variation. A near-optimal heuristic is also presented to solve the joint optimization problem. The experimental results, obtained after jointly optimizing the supply voltage and sleep transistor size, show more than 98% total energy reduction in benchmark circuits for applications with relaxed deadline constraints.

The rest of paper is organized as follows. Section II reviews related work. Section III briefly introduces the LER phenomenon and Section IV analyzes the performance of standard cells under the LER variation. The optimization problem is formulated and solved in Section V. We present the simulation results in Section VI.

II. RELATED WORK

Tons of research efforts have been conducted in improving sleep transistor designs [1][13], exploring novel gating structures [14][15][16], improving power gating algorithms [17][18], and investigating the effect of power gating method for new VLSI

This research is supported by grants from the PERFECT program of the Defense Advanced Research Projects Agency and the Software and Hardware Foundations of the National Science Foundation.

technologies [19]. A co-optimization framework of supply voltage and sleep transistor size was presented in [2] for planar CMOS circuits in very ultra-low voltage regime. However, the model of delay penalty is over-simplified in [2], i.e., lacking of a quantitative analysis on the delay penalty. All these works mentioned above were focusing on conventional planar CMOS technologies. The authors in [20] presented simulation results showing that power gating structures in FinFET circuits offer robust circuit operations and reduced standby leakage, without significant performance and area penalties. However, how to choose the operating points (e.g., supply voltage) and sleep transistor size is yet to be studied for FinFET circuits. Therefore, in this work, we present a joint optimization method to determine the energy-optimal supply voltage and sleep transistor size, subjecting to certain deadline constraints. Compared to [2], we account for the process-induced variation and analyze the impact of LER effect on FinFET circuits. We consider variations in both circuits delay and virtual ground voltage, and quantitatively calculate the joint circuit delay distribution in critical timing paths by using pre-characterized information and statistical methods.

III. BACKGROUND

A. Line-Edge Roughness in FinFET Devices

LER refers to the fluctuation of a given line edge around its mean value. Due to continued downscaling of feature sizes, roughness in printed transistor features is no longer negligible compared to the geometric dimension of a device. The LER in FinFET device is more complicated because it not only affects the gate length, but also the fin height. Figure 1(a) illustrates the LER effect in FinFET devices. *Fin LER* results in a rough channel surface in the front and back gates, while the *gate LER* has a smooth channel surface but suffers from a non-uniform channel length. Note that both of fin LER and gate LER occur simultaneously in actual devices; they are separated in Figure 1(b) and (c) for clarity. Considerable research efforts have been conducted on modeling the impact of LER effect on FinFET devices [10][21]. According to [22][23], fortunately, the sidewall surfaces and fin corners can be smoothened in advance by using thermal annealing method. Therefore, in this work, we mainly consider the gate LER, which gives rise to the variation of the effective channel length, denoted by L_g.

B. 7nm FinFET Standard Cell Library

We build compact models of FinFET devices, in which each N-type or P-type fin is modeled as a set of current sources and parasitic capacitances. The model parameters are extracted by simulating 7nm FinFET devices using Sentaurus TCAD tools and stored in look-up tables. The look-up tables are made compatible to HSPICE through a Verilog-A format interface. A standard cell library is built using the 7nm FinFET technology node.

IV. MODELING LER VARIATION

The circuit performance is affected by the LER variation as it results in variations of the effective channel length as well as the threshold voltage. To estimate the impact of the LER variation on the circuit performance, we perform Monte Carlo (MC) simulations on all types of logic cells in the previously presented standard cell library. In our problem setup, like authors in [22], we consider a Gaussian inter-die LER variation with 0.8nm standard deviation around the mean value of effective channel length (7nm).

A. Characterizing Delay Distributions for Standard Cells

We assume that the effective channel length L_g follows a Gaussian distribution with μ_L of 7nm and $\pm 3\sigma_L$ point at 4.6 and 9.4nm, respectively. For each standard cell, we carry out a Monte Carlo simulation, in which we randomly generate channel length variation ΔL_g, such that $L_g + \Delta L_g \sim N(\mu_L, \sigma_L)$. We simulate each cell using HSPICE and measure propagation delays for 6K ~ 8K MC samples.

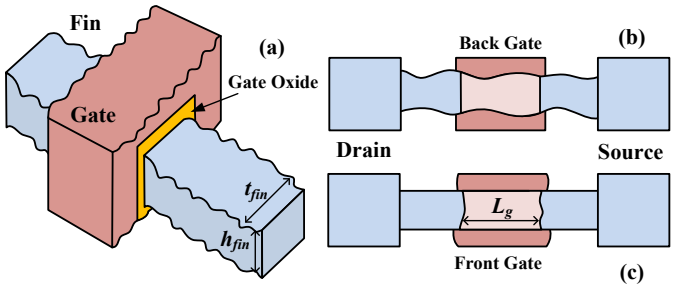

Figure 1. 3D illustration of LER in FinFET devices (a), cross-section view of fin LER (b) and gate LER (c).

Previous researches that studied the impact of LER variation revealed that the threshold voltage of FinFET devices changes due to the short channel effect [10][21]. In addition, authors in [5][6] had proved that the circuit delay follows a lognormal distribution with respect to the threshold voltage V_{th} when circuits are operating in the sub/near-threshold regime. Note that the LER variation also affects the FinFET device performance in another way by changing L_g. However, it affects the performance in a polynomial manner, whereas the V_{th} does in an exponential manner. Therefore, we approximately consider that V_{th} is the dominant way that the LER variation affects circuit delay. Thus, we consider that delay of a standard cell is a random variable with a lognormal distribution as,

$$D_{cell,pin,r/f}(V_{sw}) \sim e^{l+sZ} \quad (1)$$

where l and s are the *location parameter* and *scale parameter*, respectively, and Z is a *standard Gaussian variable*, i.e., $N(0,1)$. Note that at the super-threshold regime, a Gaussian distribution is applied to relate the delay variation to L_g. V_{sw} in (1) denotes the *voltage swing* across the standard cell, which is defined as the voltage difference between the supply voltage rail and the ground rail.

Figure 2 shows MC results (red curves) of the propagation delay of 1X inverter, 2-input NAND, and 2-input NOR in the standard cell library at different voltage swings. The fitted lognormal distributions are plotted as blue curves. Although lognormal distributions do not fit the MC results well on the left side, they accurately capture the tail of the distribution, which is more important in evaluating circuits delays at the 0.1% value at risk point (calculated by assigning $Z = 3$ in (1)). One can observe from Figure 2 that value at 0.1% risk point points of fitted distributions cover 99.9% of corresponding MC results. In addition, the main body of the distribution is also well captured by the lognormal distribution fittings so that the power consumption can also be evaluated based on the fitted distribution. Similar to these three cells, we obtain the fitted lognormal distribution for all input pins and rise/fall transitions of all logic cells in our standard cell library, at a series of typical voltage swings.

B. Obtaining Delay Distributions of Critical Timing Paths

The delay of a VLSI circuit is mainly determined by its critical timing path. We utilize static timing analysis tools to identify the critical timing path of the benchmark circuits, which saves us from extremely computational expensive circuit-level simulations. However, the potential weakness of this approach is that the critical timing path reported at nominal process corners is not necessarily the critical one considering the LER variation. To overcome this issue, we obtain the critical timing path by synthesizing the benchmark circuits using Synopsys Design Compiler with our characterized 7nm FinFET standard cell library, and letting the Design Compiler report K number of critical timing paths. In practice, we set $K = 20$ in this work.

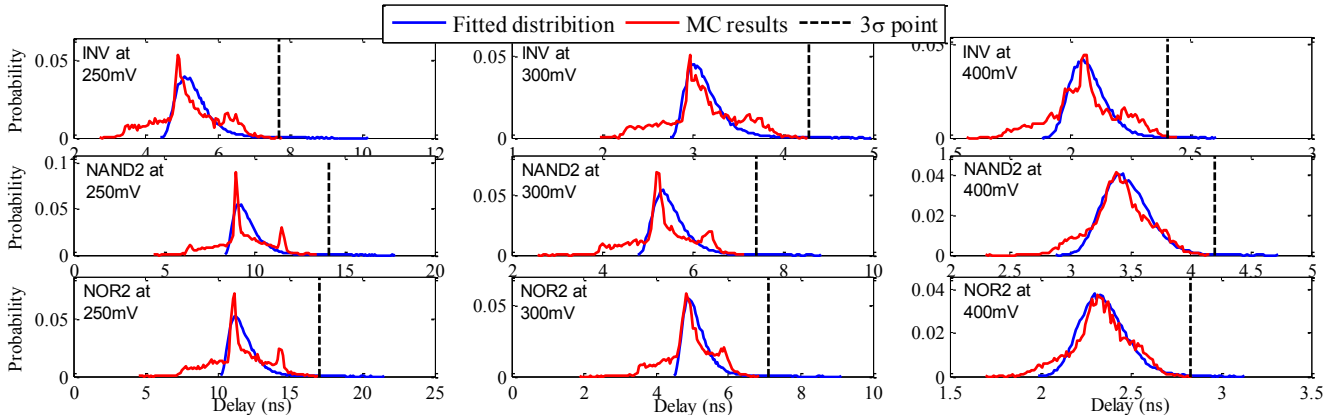

Figure 2. MC results and lognormal distribution fittings for INV, NAND2, and NOR2 at different voltage swings.

For an identified critical timing path p_j, the path delay under the LER variation is given by summarizing the propagation delays of all cells along this path,

$$D_{p_j}(V_{sw}) = \sum_{i \in p_j} D_{i,pin,r/f}(V_{sw}) \quad (2)$$

Since the propagation delay of a cell depends on its input slew and output capacitance, the delay distribution of a path may also change at different input slews and output capacitances. Therefore, to avoid performing MC simulations and distribution fittings at all combinations of input slew and output capacitance, we adopt an assumption made by authors in [5] such that the ratio of the variance and mean of a lognormal distribution is fixed for different input slews and output capacitances. In particular, we perform MC simulation to obtain FO4 delay distributions for all standard cells and extract the corresponding lognormal distribution parameters. For each cell along the critical timing path, we scale the fitted FO4 delay distribution so that the mean delay of the distribution matches the value reported by Design Compiler. The variance is scaled accordingly to ensure a constant variance/mean ratio. Having scaled the delay distribution of each cell along the critical timing path, the overall delay of the critical timing path becomes a summation of a series of random variables with a given distribution. We apply the Fenton-Wilkinson Approximation to obtain the summation of multiple lognormal random variables [25].

C. *Considering the Sleep Transistor*

Without loss of generality, we consider footer sleep transistors in this work. The presented analysis and optimization method in this work can also be applied to header sleep transistors. Adding a sleep transistor can significantly reduce the leakage power consumption. However, it also brings side effects. For example, the sleep transistor acts like a resistor when the circuits are in the active mode. The driving current flowing through the sleep transistor creates a small voltage drop across the virtual ground rail and the actual ground rail, which degrades the voltage swing of the logic circuit and results in longer propagation delay. It is known that increasing the size of the sleep transistor is helpful in alleviating this voltage swing degradation, at the cost of larger circuit area and lower power gating efficiency.

The dependency of propagation delay effect on the sleep transistor becomes more involved when the process-induced variation is accounted. In particular, the sleep transistor also suffers from the LER variation so that its driving strength also varies in a certain range. To relate the virtual ground voltage, sleep transistor size, and the driving strength, we perform a 2D sweep with different drain voltages (which is the virtual ground voltage V_{VGND}) and channel lengths of the sleep transistor L_s, while the width of sleep transistor is set to be a fixed value W_{char}. We record the characterization results in a 2D array A_{2D}, in which each row corresponds to one data point (V_{VGND}, L_s, I_{char}/W_{char}). Note that we normalize the driving current to a single fin by dividing the characterized driving currents I_{char} by W_{char}. This 2D array is sorted based on values of its third row I_{char}/W_{char}.

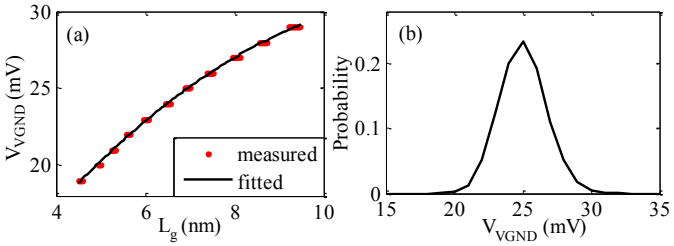

Figure 3. (a) relation between virtual ground voltage and the channel length for a particular driving current, and (b) distribution of V_{VGND} due to the LER variation.

In practice, for benchmark FinFET circuits, we approximately determine the desired driving strength I_D of the sleep transistor by measuring the maximum on-current of the circuits without power gating structures. Then the normalized single-fin driving current is calculated as I_D/W_s, where W_s is the width of the sleep transistor used in circuits. We access the sorted 2D array A_{2D} with index key of I_D/W_s and locate those V_{VGND}'s and L_s's that produce the I_D/W_s amount of driving current. A curve fitting is performed to obtain the relation F_{VL} between V_{VGND} and L_s, and this relation is further used to estimate V_{VGND} based on the actual channel length of the sleep transistor used in circuits. More precisely, V_{VGND} is determined as follow,

$$\begin{aligned} &[V_{VGND}\text{'s}, L_g\text{'s}] \leftarrow FindDatainArray\left(A_{2D}, \left(\frac{I_D}{W_s} - \epsilon, \frac{I_D}{W_s} + \epsilon\right)\right) \\ &F_{VL} \leftarrow CurveFitting(V_{VGND}\text{'s}, L_s\text{'s}) \\ &V_{VGND} = F_{VL}(L_s + \Delta L_s) \end{aligned} \quad (3)$$

where ϵ is a small value that all rows in A_{2D} with current values between $\left(\frac{I_D}{W_s} - \epsilon, \frac{I_D}{W_s} + \epsilon\right)$ are considered in the curving fitting. Figure 3(a) shows the relationship between V_{VGND} and the channel length L_s for a specific driving current level I_D and sleep transistor width W_s. A clear polynomial relationship exists between these two variables, and thereby we perform a second-order curve fitting to relate them. Moreover, since the channel length follows a Gaussian distribution due to the LER effect, we also obtain V_{VGND} distribution by considering the variation of L_s in (3). Figure 3(b) shows the probability of observing a particular virtual ground voltage in the gated circuits in the presence of the LER variation.

To account for the delay variation caused by the sleep transistor, we jointly consider the propagation delay distribution of the critical

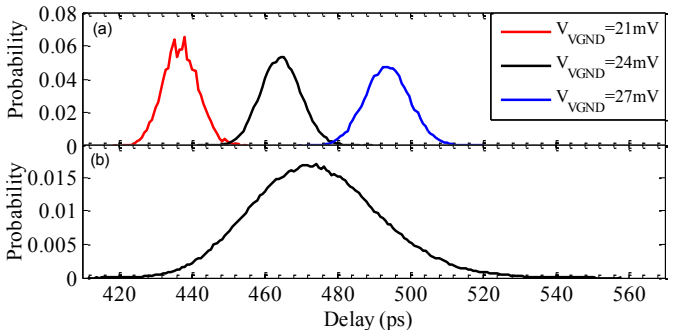

Figure 4. (a) delay distribution of the critical timing path at some particular V_{VGND} points, and (b) joint delay distribution considering both critical timing path delay variation and V_{VGND} variation. Delay values are calculated for C432 circuits operating at 0.23V and having sleep transistor width of 16 fins.

timing paths at a particular voltage swing and the V_{VGND} distribution. The voltage swing for circuits with footer sleep transistors is,

$$V_{sw} = V_{dd} - V_{VGND} \tag{4}$$

Therefore, we obtain the propagation delay distribution of critical timing paths in gated circuits by calculating the probability of observing a particular delay value t_{delay} as,

$$p(t_{delay}) = \sum_{V_{VGND}} p(t_{delay}|V_{dd} - V_{VGND}) \times p(V_{VGND}) \tag{5}$$

where the $p(t_{delay}|V_{dd} - V_{VGND})$ and $p(V_{VGND})$ are probabilities obtained from (4) and (5), respectively.

Figure 4(a) shows the delay distribution of the critical timing path in c432 circuits at different V_{VGND} points, which corresponds to the first term in RHS in (5). One can see that the delay distribution disperses as the V_{VGND} increases. This is because higher V_{VGND} results in smaller V_{sw}, and the impact of process variation at low V_{sw} level is greater. Figure 4(b) shows the simulated joint delay distribution considering both circuit delay variation and the V_{VGND} variation. The distribution curve in Figure 4(b) is obtained by combining the V_{VGND} variation in Figure 3(b) with circuit delay variations, like those shown in Figure 4(a).

V. Energy Minimization

We first analyze and characterize the important energy consumption terms in the FinFET circuits. Then relations of these important energy consumption terms versus supply voltage are presented and corresponding curve fittings are applied accordingly to extract the parameters of interest. After that, we formulate an optimization problem targeting at minimizing total energy consumption of FinFET circuits. An energy-optimal solution is found by jointly optimizing both the supply voltage and sleep transistor size.

A. Energy Consumption Characterization in FinFET Circuits

We separate the energy consumption into two parts: a *dynamic* part and a *static* part. The dynamic energy consumption mainly contains the switching energy at the output capacitance and the short-circuit energy consumption, while the static part, also known as *leakage* part, accounts for the energy consumption due to the leakage. For circuits without power gating, assuming a deadline time $T_{deadline}$ is given to a circuit operation, the total energy consumption E_{tot} is calculated as,

$$E_{tot} = a_f E_{dyn} + E_{leak} = a_f E_{dyn} + P_{leak} T_{deadline} \tag{6}$$

where E_{dyn} stands for the dynamic energy consumption, a_f is the activity factor, and P_{leak} stands for the leakage power consumption while circuits are in the idle mode. However, for many applications, the circuit operation can finish earlier than the deadline time. Assuming the circuits operation takes t_{delay} time to finish, circuits can be put to the sleep state after t_{delay} time. Thus, with power gating, E_{tot} is calculated as,

$$\begin{aligned} E_{tot} &= a_f E_{dyn} + E_{leak} + E_{sleep} \\ &= a_f E_{dyn} + P_{leak} t_{delay} + P_{sleep}(T_{deadline} - t_{delay}) \end{aligned} \tag{7}$$

where P_{sleep} is the remaining leakage power consumption after sleep transistors cut off the circuits from power supply rails. Note that though P_{sleep} is typically two orders of magnitude smaller than P_{leak}, they shall be accounted correctly as omitting the sleep energy consumption results in highly suboptimal energy consumption [2].

E_{dyn} of combinational logics depends on the supply voltage V_{dd}. We approximately capture this relation using a power-law equation,

$$E_{dyn} = a \cdot (V_{dd})^b \tag{8}$$

where a and b in (8) are fitting parameters. On the other hand, because the subthreshold leakage current has an exponential relationship with respect to V_{dd} [6][26], the leakage and sleep power consumption are captured by using a similar equation,

$$\begin{aligned} P_{leak} &= cV_{dd} \cdot exp(V_{dd} - d) \\ P_{sleep} &= eV_{dd} \cdot exp(V_{dd} - f) \end{aligned} \tag{9}$$

where c, d, e, and f in (9) are fitting parameters. Note that in (9), V_{dd} also affects P_{leak} and P_{sleep} linearly. However, the exponential dependency of p_{leak} on V_{dd} dominates in (9), therefore, we approximately consider the exponential part only and treat the linear part as a constant.

B. Joint Optimization of V_{dd} and the Sleep Transistor Size

It is known that applying different V_{dd}'s results in significant changes of circuit delay, E_{dyn}, as well as P_{leak}. Thus, V_{dd} has always been a crucial optimization variable for VLSI circuits. In addition to that, the size of the sleep transistor W_s also plays an important role in determining the delay and energy consumption of the circuits. On the one hand, small W_s effectively reduces energy consumption in the sleep mode at the cost of higher delay penalty, as well as high leakage energy consumption in the active mode, according to (7). On the other hand, a large sleep transistor suffers from less performance degradation in the active mode, however, is less effective in saving the leakage energy. Therefore, we include W_s as another optimization variable. The joint optimization problem is formulated as follows.

Given: i) FinFET benchmark circuits;

Heuristic: Min Energy of circuits w/ PG under LER (MEPL)

Inputs: circuits of interest in verilog format; l's, s's for all standard cells; a's, b's, c's, d's, e's, f's for benchmark circuits; A_{2D}; $T_{deadline}$; available V_{dd}'s; available W_s's.

Synthesize the circuits of interest;

Extract several critical paths from the synthesized circuits;

Pick a starting point $(V_{dd,min}, W_{s,min}) = (V_{dd,i}, W_{s,j})$;

While $V_{dd,min}$ and $W_{s,min}$ are not converged

  For $(V_{dd,i}, W_{s,j})$ and its four neighbor points $(V_{dd,i+1}, W_{s,j})$, $(V_{dd,i}, W_{s,j+1})$, $(V_{dd,i-1}, W_{s,j})$, $(V_{dd,i}, W_{s,j-1})$, do

    Generate V_{VGND} from L_g distribution and A_{2D} from (3);

    Calculate t_{delay} distribution from (4) and (5);

    Calculate E_{dyn}, P_{leak}, and P_{sleep} from (8) and (9);

    Calculate E_{tot} from (7);

  Endfor

  Pick the pair that gives the minimal E_{tot} among the five pairs and update $(V_{dd,min}, W_{s,min})$;

Endwhile

Return: converged $(V_{dd,min}, W_{s,min})$.

Table 1. Joint optimization results for different benchmark circuits.

Benchmark circuits	$T_{deadline} = 3D_{030}$				$T_{deadline} = 10D_{030}$				$T_{deadline} = 100D_{030}$			
	V_{dd} (V)	W_s	E_{tot} (fJ)	E_{tot}^{bl} (fJ)	V_{dd} (V)	W_s	E_{tot} (fJ)	E_{tot}^{bl} (fJ)	V_{dd} (V)	W_s	E_{tot} (fJ)	E_{tot}^{bl} (fJ)
20-stage FO4 inverter chain	0.22	2	0.139	0.274	0.21	1	0.157	0.473	0.20	1	0.372	3.157
16-bit carry ripple adder	0.36	1	0.156	0.872	0.33	1	0.206	2.67	0.27	1	0.581	21.90
C432	0.26	40	4.42	7.02	0.27	24	4.49	21.41	0.28	14	5.07	215.5
C499	0.28	58	6.20	16.44	0.29	30	6.37	55.83	0.31	14	7.34	659.0
C880a	0.28	42	4.35	9.95	0.29	24	4.42	33.51	0.28	12	4.91	342.4
16-bit binary multiplier	0.32	23	12.30	33.78	0.33	14	13.04	106.2	0.33	12	20.08	969.7
C1355	0.24	105	11.34	11.88	0.21	145	10.89	15.33	0.23	80	13.17	102.44
C1908	0.21	250	25.42	26.01	0.21	145	26.09	41.6	0.24	85	30.88	268.7

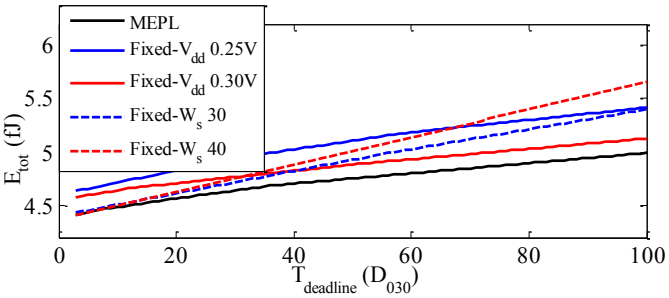

Figure 5. Comparison of energy results obtained by MEPL and other optimization heuristics for benchmark circuits C432 at different deadline constraints.

ii) delay distribution parameters l's, s's of all standard cells;
iii) energy consumption parameters (a's, b's, c's, d's, e's, f's) of benchmark circuits;
iv) characterized relations between drain voltage, driving current, and size of sleep transistor;
v) deadline of circuit operation $T_{deadline}$.
Find: supply voltage V_{dd} and sleep transistor size W_s.
Minimize: total energy consumption per operation in (7) under the LER variation.
Subject to: delay of the circuits at 0.1% value at risk point is within the deadline time, i.e., t_{delay}(0.1% value at risk)< $T_{deadline}$.

We present *MEPL,* a heuristic to minimize the energy consumption per operation for FinFET circuits with power gating structure under the LER variation as follow. The optimality of MEPL heuristic is analyzed in the Section VI.

VI. Experimental Results

We adopt a very advanced FinFET technology – 7nm FinFET technology. We perform the MEPL heuristic for a number of ISCAS'85 benchmarks, as well as an inverter chain, 16-bit carry ripple adder, and 16-bit binary multiplier. All benchmarks circuits are synthesized using Synopsys Design Compiler. We extract up to 20 critical timing paths for each synthesized benchmark circuits. A pre-characterization process is carried out to extract delay distribution parameters for all standard cells and energy consumption parameters for each benchmark, respectively. We measure delays and energy consumptions in HSPICE.

Table 1 shows joint optimization results of various benchmark circuits at different deadline constraints. For each benchmark, we test the presented heuristic at 3, 10, 100 times of critical timing path delay when circuits are operating at 0.30V, which is denoted by D_{030}. We consider discrete available supply voltage levels with step size of 0.01V. The available width of sleep transistors are also discretized with different step sizes, depending on their peak on-currents. The activity factors of all benchmark circuits are assumed to be 0.2. We also calculate the energy consumption of the same benchmark circuits operating at the same supply voltage without using power gating structures, and denote them as E_{tot}^{bl} in Table 1. Compared to E_{tot}^{bl}, the FinFET circuits with properly sized sleep transistors can achieve up to 98% total energy reduction for applications with relaxed deadline constraints.

One can observe from Table 1 that as the benchmark circuits become larger (i.e., number of gates increases), our heuristic assigns larger sleep transistor. This is because that generally the peak on-current increases with the number of gates in the circuits. Assigning large sleep transistor results in smaller voltage drop as well as less performance penalty in the active mode. Moreover, according to (7), the leakage energy E_{leak} is proportional to the circuit delay. Therefore, assigning large sleep transistor is helpful to reduce the total energy consumption per operation.

Another observation one can make from Table 1 is that the optimal width of the sleep transistor decreases for applications with relaxed deadline constraints. For $T_{deadline} = 100D_{030}$, we notice that the MEPL heuristic assigns smaller sleep transistors than those in $T_{deadline} = 3D_{030}$ case. This is because that the sleep energy E_{sleep} plays a much more important role for relaxed deadline constraints ($T_{deadline} \gg t_{delay}$ in (7)), and small sleep transistors significantly reduces the sleep state power consumption P_{sleep}.

Note that according to Table 1, to reduce the total energy consumption per operation, relaxed deadline requirements do not necessarily results in a low supply voltage. The total energy consumption contains not only E_{dyn}, which reduces as V_{dd} decreases, but also the leakage energy consumption E_{leak}. The latter is not a monotonic function of V_{dd} because it is a product of two terms: P_{leak}, which reduces as V_{dd} decreases, and t_{delay}, which increases exponentially as V_{dd} decreases. The energy-optimal configuration of supply voltage and sleep transistor size is determined in our joint optimization heuristic.

Figure 5 compares the solution quality returned by the presented joint optimization heuristic MEPL and other baseline heuristics that only optimize either V_{dd} or W_s, leaving the other one fixed. We apply these baseline heuristics at fixed V_{dd} of 0.25V and 0.30V, and fixed W_s of 30 fins and 40 fins, respectively. In Figure 5, one can observe that the MEPLheuristic consistently outperforms baseline heuristics for all deadline constraints scenarios. The results show that the presented MEPL heuristic, which optimizes both V_{dd} and W_s jointly, brings the most significant reduction of E_{tot}.

The presented MEPL heuristic finds the energy-optimal (V_{dd}, W_s) pair by searching neighbors of the current pair and moving to the one which consumes less total energy. Since the total energy consumption

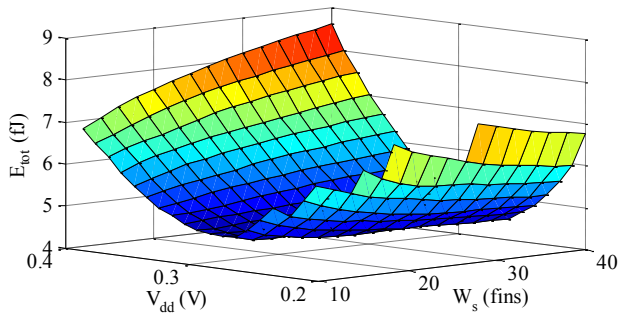

Figure 6. The total energy consumption of all (V_{dd}, W_s) pairs for C432 benchmark circuis with $T_{deadline} = 10D_{030}$.

are continuous with respect to these two parameters, in the MEPL heuristic, we have assumed that there is no other local minima points in the 2D plane of V_{dd} and W_s. Figure 6 validates our assumption by showing E_{tot} of all (V_{dd}, W_s) pairs. One can observe that E_{tot} surface is smooth in the entire region and only one minima point exists. Therefore, the presented MEPL heuristic is able to find the energy-optimal supply voltage and sleep transistor size. In practice, near-optimal solutions are found as MEPL heuristic takes discrete values of supply voltage and sleep transistor size as inputs. Note that there are some points missing in the Figure 6 at very low V_{dd} and very small W_s because deadline constraints cannot be met at those points.

VII. CONCLUSION

Power gating method has been an effective method in reducing the leakage energy consumption for conventional CMOS technologies. This paper investigated the effect of applying power gating method to advanced FinFET technology nodes. Since process-induced variations is an crucial factor that affects the timing and energy for modern technologies, this work accounted for one of the most important variation sources in FinFET circuits – line-edge roughness (LER), and derived the delay distribution under the LER variation. A joint optimization problem was also formulated to minimize the total energy consumption of a given FinFET circuit, while both of the supply voltage and sleep transistor size are considered as optimization variables. We presented an effective heuristic to near-optimally solve the joint optimization problem and determine the supply voltage and sleep transistor size, subjecting to certain deadline constraints. HSPICE simulation results showed that while designed with a proper sleep transistor and operating at a proper supply voltage, power gating method is able to achieve more than 98% energy reduction for applications with relaxed deadline constraints. We also showed that the presented heuristic consistently outperformed other baseline heuristics, which do not consider both optimization variables simultaneously.

REFERENCES

[1] J. Kao, A. Chandrakasan, D. Antoniadis, "Transistor Sizing Issue and Tool for Multi-Threshold CMOS Technology", *Proc. of Design Automation Conf*, pp. 409-414, June, 1997.

[2] M. Seok, S. Hanson, D. Sylvester, and D. Blaauw, "Analysis and optimization of sleep mode in subthreshold circuit design," *ACM/IEEE Design Automation Conference*, 2007.

[3] R. G. Dreslinski et al. "Near-threshold computing: Reclaiming moore's law through energy efficient integrated circuits", *Proceedings of the IEEE*, 98(2).

[4] D. Markovic, C. C. Wang, L. P. Alarcon, T. T. Liu, and J. M. Rabaey, "Ultralow-power design in near-threshold region", *Proceedings of the IEEE*, 98(2).

[5] B. H. Calhoun, A. Wang, and A. Chandrakasan, "Modeling and sizing for minimum energy operation in subthreshold circuits", *Journal of Solid State Circuits*, 40(9).

[6] B. Zhai, S. Hanson, D. Blaauw, and D. Sylvester, "Analysis and mitigation of variability in subthreshold design", in *ISLPED, 2005.*

[7] L. Chang et. al., "Reduction of direct-tunneling gate leakage current in double-gate and ultra-thin body MOSFETs," *IEDM*, 2001.

[8] B. Yu et. al., "FinFET scaling to 10 nm gate length", *IEDM* 2002.

[9] A.R. Brown, A. Asenov, J. R. Watling, "Intrinsic fluctuations in sub 10-nm double-gate MOSFETs introduced by discreteness of charge and matter," *IEEE Transactions on Nanotechnology*, vol.1, no.4, pp.195-200, Dec 2002.

[10] K. Patel, T. K. Liu, and C. J. Spanos, "Gate Line Edge Roughness Model for Estimation of FinFET Performance Variability," *Electron Devices, IEEE Transactions on*, vol.56, no.12, pp.3055-3063, Dec. 2009

[11] C. Gustin, L. H. A. Leunissen, A. Mercha, S. Decoutere, and G. Lorusso, "Impact of line width roughness on the matching performances of nextgeneration devices," *Thin Solid Films,* vol. 516, no. 11, pp. 3690–3696, Apr. 2008.

[12] A. Asenov, A. R. Brown, J. H. Davies, S. Kaya, and G. Slavcheva, "Simulation of intrinsic parameter fluctuations in decananometer and nanometer-scale MOSFETs," *IEEE Trans. Electron Devices*, vol. 50, no. 9, pp. 1837–1852, Sep. 2003.

[13] C. Long, L. He, "Distributed sleep transistor network for power reduction," *IEEE Trans. on VLSI Systems*, Vol. 12, No. 9, pp. 937-946, Sep. 2004.

[14] S. Kim, S. Kosonocky, D. Knebel, K. Stawiasz, and M. Papaefthymiou, "A multi-mode power gating structure for low-voltage deep-submicron CMOS ICs," *IEEE Trans. Circuits Syst. II, Exp. Briefs*, vol. 54, no. 7, pp. 586–590, Jul. 2007.

[15] E. Pakbaznia and M. Pedram, "Design of a Tri-Modal Multi-Threshold CMOS Switch With Application to Data Retentive Power Gating," *IEEE Trans. on VLSI Systems*, vol.20, no.2, pp.380-385, Feb. 2012.

[16] Z. Zhang, X. Kavousianos, K. Chakrabarty, and Y. Tsiatouhas, "Static Power Reduction Using Variation-Tolerant and Reconfigurable Multi-Mode Power Switches", *IEEE Trans. on VLSI Systems,* vol. 22, 2014.

[17] A. Abdollahi, F. Fallah, and M. Pedram, "A robust power gating structure and power mode transition strategy for MTCMOS design," *IEEE Trans. Very Large Scale Integr. Syst.*, vol. 15, no. 1, Jan. 2007.

[18] B. Yu and M. L. Bushnell, "A Novel Dynamic Power Cutoff Technique (DPCT) for Active Leakage Reduction in Deep Submicron CMOS Circuits," in *ISLPED*, 2006.

[19] K. Kim; Y. Kim, and K. Choi, "Hybrid CMOS and CNFET Power Gating in Ultralow Voltage Design," *IEEE Trans. on Nanotechnology*, vol.10, no.6, pp.1439-1448, Nov. 2011.

[20] K. Kim, R. Kanj, and R.V. Joshi, "Impact of FinFET technology for power gating in nano-scale design," in *ISQED* , pp.543,547, Mar 2014.

[21] E. Baravelli, A. Dixit, R. Rooyackers, M. Jurczak, N. Speciale, K. De Meyer, "Impact of Line-Edge Roughness on FinFET Matching Performance," *Electron Devices, IEEE Transactions on*, vol.54, no.9, pp.2466-2474, Sept. 2007.

[22] Y.-K. Choi, D. Ha, E. Snow, J. Bokor, and T.-J. King, "Reliability study of CMOS FinFETs," in *IEDM Tech. Dig.*, 2003, pp. 177–180.

[23] W. Xiong et. al., "Improvement of FinFET electrical characteristics by hydrogen annealing," *IEEE Electron Device Lett.*, vol. 25, no. 8, pp. 541–543, Aug. 2004.

[24] 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," to appear in *SOI Conference*, Oct. 2014.

[25] P. Pirinen, "Statistical power sum analysis for nonidentically distributed correlated lognormal signals", *Proc. 2003 Finnish Signal Processing Symp.*, pp.254 -258 2003.

[26] Q. Xie, Y. Wang, and M. Pedram, "Designing soft-edge flip-flop-based linear pipelines operating in multiple supply voltage regimes", *Integration, the VLSI Journal,* vol. 47(3), Jun. 2014, pp. 318–328.