A Design-Specific and Thermally-Aware Methodology for Trading-Off Power and Performance in Leakage-Dominant CMOS Technologies

Sheng-Chih Lin, *Member, IEEE*, and Kaustav Banerjee, *Senior Member, IEEE*

***Abstract*—As CMOS technology scales deeper into the nanometer regime, factors such as leakage power and chip temperature emerge as critically important concerns for high-performance VLSI design. Consequently, enhancing processing performance is no longer the most important factor that dominates future circuit design considerations. This paper, for the first time, proposes a systematic methodology to determine a generalized design optimization metric for simultaneously trading-off power and performance in nanometer scale integrated circuits to achieve design-specific targets. The methodology incorporates interconnect effects as well as electrothermal couplings between substrate temperature, power, and performance for nanometer scale design optimization. Implications of choosing a specific design optimization metric on power, performance, and operating temperature are illustrated and discussed. The proposed methodology is shown to provide a more meaningful optimization metric (for power-performance tradeoff analysis) and basis, with considerations of chip-level thermal management including maximum allowable operating temperature and packaging/cooling solutions. Furthermore, implications of CMOS technology scaling and parameter variations on the proposed methodology are discussed.**

***Index Terms*—Chip-package co-design, integrated circuit (IC), leakage, performance, power, thermal-aware design, thermal management.**

I. Introduction

THE STEADY downscaling of transistor dimensions has ensured higher packing density, higher performance, and lower cost of integrated circuits in the past few decades [1]. Main efforts of technology scaling have been focused on achieving highest processing performance. In recent years, however, the awareness of low-power has become a critical issue for circuit designers especially for all cost-performance, portable, and battery-constrained electronic products [2]–[4]. For instance, many handheld devices including wireless applications prefer low-power over high-performance design due to limited battery budget. Also, with the minimum feature size of the transistor entering the nanometer regime ($<$ 100 nm), the leakage power has become a significant fraction of the overall chip power dissipation and severely impacts the packaging, cooling costs, and reliability for leakage dominant CMOS technologies [5]–[8].

Manuscript received February 12, 2007; revised July 31, 2007. Current version published October 22, 2008. This work was supported in part by Intel Corporation, by the University of California-MICRO Program (03-004), and by the National Science Foundation under Award CCF-0541465.

S-C. Lin was with the Department of Electrical and Computer Engineering, University of California, Santa Barbara, CA 93106 USA. He is now with Intel Corporation, Chandler, AZ 85226 USA (e-mail: sclin@ece.ucsb.edu).

K. Banerjee is with the Department of Electrical and Computer Engineering, University of California, Santa Barbara, CA 93106 USA (e-mail: kaustav@ece.ucsb.edu).

Digital Object Identifier 10.1109/TVLSI.2008.2001060

A. Simultaneous Power-Performance Optimization: Motivation, Design Metrics, and Prior Work

For power-constrained applications, lowering supply voltage ($V_{\rm dd}$) offers the most obvious option to decrease the total power consumption, since CMOS switching power has a quadratic dependence on supply voltage. On the other hand, lowering supply voltage degrades the performance of circuits. It is, however, possible to maintain the performance by decreasing the threshold voltage ($V_{\rm th}$) simultaneously, but then the subthreshold leakage power increases exponentially. Consequently, the need for low-power as well as high-performance circuit applications motivates the search for an optimal set of supply and threshold voltages to tradeoff processing performance and power consumption. The choice of supply and threshold voltages is critical not only from power and performance aspects, but also because of reliability issues. For example, the supply voltage has a direct impact on gate-oxide and hot carrier reliability [9]–[11] and an indirect impact on electromigration (EM) reliability through the junction temperature [12].

Several design metrics and methodologies have been proposed in the literature to evaluate and simultaneously meet the targets of low-power and high-performance in modern VLSI designs. Design metrics such as power-per-operation and energy-per-operation have been shown to be inadequate for evaluating tradeoffs of power and performance [13], [14] because these two metrics monotonically depend solely on the supply voltage, and hence, the optimization using these metrics will lead to lower performance, which is not practical. Energy-delay product (EDP) is widely used as an appropriate metric to optimize and compare different designs where both performance and amount of computational energy are of importance [13]–[16]. In [17], Martin showed that the $\rm Et^2$ metric (where E is energy and t is delay) is a better measure of computational efficiency due to its voltage independency. Furthermore, general metrics have also been explored for improving the energy-delay efficiency. In [18], Pénzes and Martin showed that the $\rm Et^n$ metric (where n is the energy-delay efficiency index) characterizes any feasible tradeoffs between the energy and the delay of the computation. In [19], Hofstee proposed the energy-performance ratio (EPR) for analyzing power efficiency. An EPR of m suggests that an energy increment of $m\%$ corresponds to a 1% improvement in performance

(e.g., similar to the Et^n metric, Et^2 corresponds to an EPR of 2). Moreover, it was concluded that optimal metric is not unique for all designs but depends on the desired level of performance [19]. Although the idea of the generalized optimal metric has been proposed, there is no systematic methodology for choosing an appropriate design metric, which captures design-specific requirements.

Some recently proposed approaches employ tuning of variables such as supply voltage, threshold voltage, and gate size to achieve an energy-efficient design. In [20] and [21], Zyuban and Strenski use "hardware intensity" to quantify the relative cost of enhancing performance and resultant power dissipation at the circuit and micro-architecture levels. Similarly, in [22], Marković *et al.*, considered gate size, supply voltage, and threshold voltage as tuning knobs to trade off energy for performance. In their method, by analyzing and balancing the ratio of sensitivity of energy to the sensitivity of delay, energy-performance optimization can be achieved. In [23], Pant *et al.* proposed a heuristic technique for minimizing the total power consumption under a given delay constraint. The approach simultaneously determines transistor power supply, threshold voltage, and device width by two distinct phases. However, these approaches model and tradeoff energy and delay invariably by tuning variables (supply voltage, threshold voltage, transistor size, etc.) and do not comprehend the interdependence of thermal and power dissipation issues, which become critical in nanometer scale designs, as discussed in the following sections. As a consequence, all these prior works cannot comprehend the implications of thermal management issues including packaging/cooling solutions on design optimization and vice versa.

B. Significance of Thermal Effects in Nanometer Scale Designs and Scope of This Work

Due to technology scaling and parameter variations [24], leakage power dissipation, which is dominated by subthreshold leakage for high-performance ICs, becomes a significant component of total chip power consumption [5], [6], [25]. The subthreshold leakage is exponentially dependent on temperature and exacerbates with technology scaling. Also, increase in total chip power consumption causes higher junction temperatures (T_j), which further increases the subthreshold leakage power, thereby creating a strong feedback loop leading to various electrothermal couplings [8]. Hence, for nanometer scale technologies where power and associated thermal issues are the primary concerns, it is critical to consider the impact of thermal effects on design optimization and on the choice of design metrics.

This paper is motivated by the search for an appropriate design metric for optimizing power and performance that can comprehend circuit specific requirements as well as the thermal and power dissipation issues that are becoming increasingly significant as CMOS technology migrates towards the deep nanometer scale. Although there is evidence of the increasing use of different optimization metrics in the existing literature [26]–[28], there is no clear explanation of why one particular optimization metric is more suitable than another and whether one metric can universally be applied to all designs at all technology nodes. This paper proposes a systematic methodology for choosing an appropriate design metric in the $V_{\mathrm{dd}} - V_{\mathrm{th}}$ design space that simultaneously captures: 1) the relative importance of power dissipation and performance and 2) the interdependence of thermal and power dissipation, to achieve design-specific targets as they change from one technology generation to the next. The advantage of the proposed electrothermally-aware method as compared to the traditionally used optimization metrics is discussed and the proposed technique is shown to provide a more meaningful basis to optimize supply and threshold voltages in nanometer scale designs. Moreover, the proposed method allows circuit designers to comprehend the implications of design choices on packaging/cooling solutions and vice versa.

This paper is organized as follows. In Section II, a review of design parameters and metrics including power and delay is presented. Power and performance optimization by the traditional EDP methodology is illustrated and discussed. In Section III, the electrothermal couplings between various design parameters for power and delay evaluation are explained. This is followed by a methodology that allows incorporation of the electrothermal couplings in the EDP-based optimization process. The electrothermally coupled energy-delay product (EEDP) methodology [29] is demonstrated and compared with the traditional EDP evaluation. Effects of activity factor on EDP estimation and implications for circuit operation and design rules by using the EEDP methodology are also shown. In Section IV, first, the logic behind the choice of different design metrics is explained through comparisons between three commonly used optimization metrics while taking electrothermal couplings into account. Since the selection of the optimization metric severely impacts the design choices, a methodology for selecting a design-specific optimization metric [30] is presented. In Section V, the proposed methodology is illustrated by an example for design optimization. In Section VI, the implications of CMOS technology scaling as well as parameter variations on the proposed methodology are demonstrated. Finally, concluding remarks are made in Section VII.

II. Design Parameters and Design Optimization

A. Traditional EDP Optimization

The critical path of a chip normally goes through a variety of gates, each with a different value of delay. However, it has been shown that changes in supply voltage, temperature, and threshold voltage affect all gates in the same way [14]. Hence, the delay of any gate remains roughly proportional to the delay of an equivalent inverter [14].

Traditionally, considering an inverter gate with a capacitive load (see Fig. 1), the average gate delay (T_g) can be estimated as (1), where C denotes the effective load capacitance $(C_p + C_L), \Delta V$ denotes the voltage swing, and I_{avg} is the average drive current. According to the Alpha-Power model [31], T_g can be simply modeled as (2) without considering the effect of interconnects. The parameter α accounts for velocity saturation condition of the transistors and is between one (complete velocity saturation) and two (no velocity saturation). In this analysis, α is chosen to be 1.3. K is a proportionality constant specific to

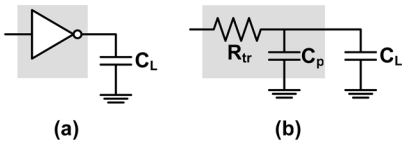


Fig. 1. (a) Schematic diagram of an inverter gate with a capacitive load (C_L). (b) Equivalent *RC* model of an inverter gate with a capacitive load (C_L). The inverter gate [shaded region in (a)] is represented by an effective on-resistance ($R_{\rm tr}$) and an output parasitic capacitance (C_p).

a given technology. Note that K lumps the capacitance and the transistor process parameters [16]

$$T_g = \frac{C\Delta V}{I_{\rm avg}} \tag{1}$$

$$T_g = \frac{KV_{\rm dd}}{(V_{\rm dd} - V_{\rm th})^{\alpha}}. \tag{2}$$

The maximum operating frequency (F) of the chip can be modeled as (3), where the parameter L_d is the logic depth (for most of the modern microprocessors, L_d is around 20 [32])

$$F = \frac{1}{T_g L_d}. \tag{3}$$

There are two main sources of power dissipation in a CMOS chip: dynamic (switching) and static (leakage). Dynamic power results from the charging and discharging of circuit capacitances between different voltage levels. Static power, on the other hand, results from the resistive paths between power supply and ground. The short-circuit component is relatively small and temperature independent. Thus, it can be considered as a constant factor of total power [33], [34]; and hence, it has been neglected throughout this paper. Note that impact of considering other power dissipation sources (e.g., short-circuit power and gate leakage) on the proposed methodology is discussed in Section III-B.

The total chip dynamic ($P_{\rm dynamic}$) and static ($P_{\rm static}$) power consumption thus can be modeled as (4) and (5), respectively, by employing effective parameters

$$P_{\rm dynamic} = aC_{\rm eff}V_{\rm dd}^2F \tag{4}$$

$$P_{\rm static} = I_s e^{-V_{\rm th}/\gamma V_o}(1 - e^{-V_{\rm th}/\gamma V_o})W_{\rm eff}V_{\rm dd} \tag{5}$$

where a is the activity (switching) factor and is taken as 0.15 [35]. $C_{\rm eff}$ accounts for the total effective output-load capacitance of the entire chip. I_s is the nominal zero-threshold leakage current, γ is a device factor, V_0 is the subthreshold slope, $V_{\rm dd}$ is the supply voltage, and $W_{\rm eff}$ is the effective transistor width (transistor width that contributes to the leakage current) of the entire chip. Hence, total power is given by the sum of $P_{\rm dynamic}$ and $P_{\rm static}$ and the energy can be calculated by (6)

$$\text{Energy} = (P_{\rm dynamic} + P_{\rm static}) \bullet \text{Delay}. \tag{6}$$

Traditionally, the design metric used to minimize both power and delay of a circuit is the EDP [13], i.e., the product of energy from (6) and delay from (2). Fig. 2 has been generated simply by direct numerical evaluation of energy and delay for a specific design. The optimal $V_{\rm dd}$-$V_{\rm th}$ set ($V_{\rm dd} = 0.498$ V and $V_{\rm th} =$ 0.252 V), denoted by "▲," corresponds to the minimum EDP value ($\text{EDP}_{\rm opt}$). The EDP contours can be found by connecting all $V_{\rm dd}$-$V_{\rm th}$ sets with the same EDP values, i.e., any point x on the contour labeled 0.5 has an EDP value twice that of the optimal value ($\text{EDP}_x = 2 \cdot \text{EDP}_{\rm opt}$). Similarly, numbers on the iso-performance curves indicate the normalized value of the frequency where normalization is done with respect to the frequency of operation at the optimal point. For instance, any point y on the curve labeled 2.0 has $2\times$ performance than that of the optimal $V_{\rm dd}$-$V_{\rm th}$ set. Note that the traditional EDP evaluation does not consider the region below the $V_{\rm dd} = V_{\rm th}$ line where circuits operate in subthreshold mode (i.e., $V_{\rm dd} < V_{\rm th}$).

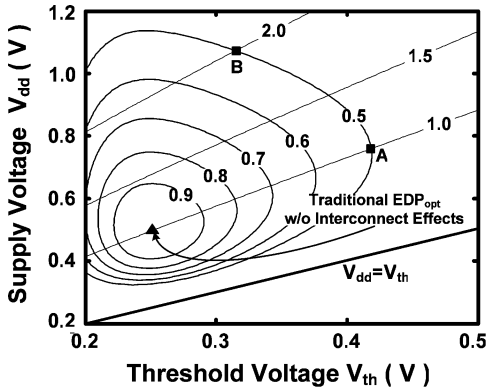


Fig. 2. Traditional $V_{\rm dd}/V_{\rm th}$ optimization uses EDP as a design metric (without considering the effect of interconnects). The EDP contours and iso-performance curves are obtained by simple numerical calculation without considering electrothermal couplings between temperature and static power dissipation at 100-nm technology node. Note that although five EDP contours (0.9, 0.8, 0.7, 0.6, 0.5) and three iso-performance curves (1.0, 1.5, 2.0) are drawn in this figure, other values of EDP contour and iso-performance curve can be easily employed.

The usage of the EDP evaluation is to optimize and compare design choices. From Fig. 2, although different design choices can have the same EDP, the performance can vary from 1 to $2\times$ (for designs A and B). Similarly, different design choices with the same performance can result in different EDP values (e.g., design A and the optimal point). Thus, the EDP-based $V_{\rm dd}$-$V_{\rm th}$ design space provides guidelines for design optimization. Also, design choices can be made based on the EDP-based evaluation.

Besides EDP, two other design metrics are often used for different applications: Power-delay product (PDP) and power-energy product (PEP). Since the logic depth (L_d) is independent of power (P) and gate delay (T_g), the parameter L_d is lumped into T_g and only T_g is used to represent the critical path delay [refers to the term "Delay" in (6) and (7)] throughout this paper for simplicity. Hence, as shown in (7), the PDP gives identical weightage to power and delay (PDP is simply the energy) while the PEP prioritizes power over delay (the exponent of power is larger than that of delay). Among these metrics, as well as the idea of the generalized optimal metric [18], power and delay are the two fundamental parameters and the metric to be chosen depends on the design optimization goal [19]

$$\begin{aligned} \text{EDP} &= \text{Energy} \cdot \text{Delay} = \text{PT}_g^2 \\ \text{PDP} &= \text{Power} \cdot \text{Delay} = \text{Energy} = \text{PT}_g \\ \text{PEP} &= \text{Power} \cdot \text{Energy} = \text{P}^2\text{T}_g. \end{aligned} \tag{7}$$

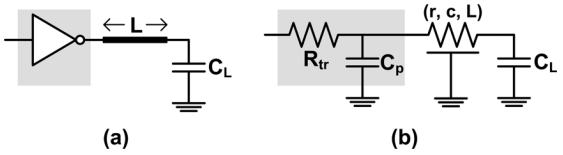


Fig. 3. (a) Schematic diagram of an inverter gate driving a capacitive load (C_L) through a wire of length L. (b) Equivalent *RC* model for (a). The inverter gate [shaded region in (a)] is represented by an effective on-resistance $(R_{\rm tr})$ and an output parasitic capacitance (C_p). Note that C_L is the input capacitance of the gate in the next stage.

B. Incorporating Interconnect Effects in Traditional EDP Optimization

As shown in (1) and (2), the gate delay (T_g) is simply modeled by an inverter gate with a capacitive load. However, as CMOS technology scales into the nanometer regime, the parasitic effects introduced by long (global and intermediate level) interconnects become more and more critical on integrated circuit design issues such as delay, power, and reliability. Typically, long interconnects in high-performance ICs are divided into a number of segments by inserting buffers or repeaters to reduce the delay. Fig. 3 shows an inverter gate driving a capacitive load (C_L) through a wire of length (L). Assuming the interconnect (L) to be uniform and homogeneous, r and c represent the resistance and capacitance per unit length of the wire. The time constant (τ) of the equivalent resistance–capacitance (*RC*) circuit in Fig. 3(b) can be modeled as (8), while the delay (T_g) is shown in (9) (time difference between the input and the output waveforms crossing 50% of the full-swing values) [36], [37]

$$\tau = R_{\rm tr}(C_P + C_L + cL) + rLC_L + (1/2)rcL^2 \quad (8)$$

$$T_g \cong 0.69R_{\rm tr}(C_p + C_L + cL) + 0.69rLC_L + 0.38rcL^2. \quad (9)$$

Assuming that the interconnect length (L) has been optimized for minimal delay and chosen to be 2.22 mm at 100-nm technology node [34]. Moreover, in (9), the capacitances $(C_p, C_L,$ and $c)$ of the inverter gates and the wire are technology specific parameters and independent of the operating condition ($V_{\rm dd}$ and $V_{\rm th}$). Thus, the delay (T_g) of the *RC* model in Fig. 3(b) can be estimated by (10) using the same proportionality constant (K) shown in (2)

$$T_g \cong \frac{KV_{\rm dd}}{(V_{\rm dd}-V_{\rm th})^{\alpha}}\left[1+\frac{cL}{C_p+C_L}\right]+0.69rLC_L+0.38rcL^2. \quad (10)$$

Similar to Fig. 2, the EDP evaluation is performed using (10) instead of (2). In Fig. 4, with the effect of interconnects, the optimal $V_{\rm dd}$-$V_{\rm th}$ set ($V_{\rm dd}$ = 0.520 V and $V_{\rm th}$ = 0.267 V) is denoted by "■" (corresponding to $\rm EDP_{opt}$). The trend of the $\rm EDP_{opt}$ (dotted arrow in Fig. 4) is shown to be monotonously increasing toward higher $V_{\rm dd}$ and $V_{\rm th}$ values along the 1.0 iso-performance curve with increasing delay (due to the effect of interconnects). Table I lists related interconnect parameters used for generating Figs. 2 and 4.

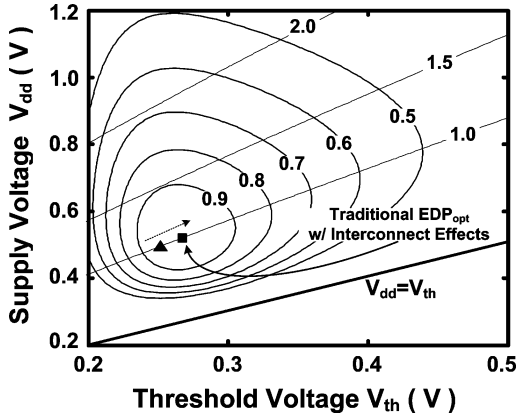


Fig. 4. $V_{\rm dd}/V_{\rm th}$ optimization using EDP as a design metric (with the effect of interconnects included). The EDP contours and iso-performance curves are obtained by simple numerical calculation without considering electrothermal couplings between temperature and static power dissipation at 100-nm technology node. $\rm EDP_{opt}$ without considering the effect of interconnects is also shown ("▲") for comparison.

TABLE I
TECHNOLOGY AND EQUIVALENT CIRCUIT MODEL PARAMETERS [34]

Tech. node (nm)	100	70
L (mm)	2.22	1.32
c (pF/m)	154	125
r (kΩ/m)	103.9	206.6
s	110	82
c_x (fF)	1.5	1.3
c_y (fF)	2.5	1.5

Note that s is the ratio of the inverter's size with respect to that of the minimal sized inverter at a given technology node. For an inverter size of s, $C_L = c_x s$ and $C_p = c_y s$.

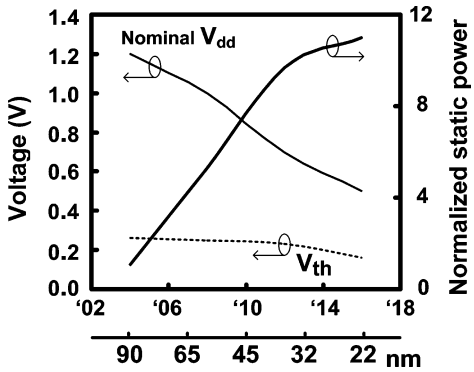


Fig. 5. Trends of nominal supply voltage, threshold voltage, and normalized leakage power based on ITRS [2].

III. ELECTROTHERMAL EDP OPTIMIZATION

A. Impact of Electrothermal Couplings on EDP Optimization

Fig. 5 shows the scaling trend of supply voltage, threshold voltage, and static power consumption projected by ITRS. Although the supply voltage decreases with scaling, the dynamic power consumption increases from generation to generation because of the increasing transistor density and switching speed.

The leakage power, which is becoming a major source of total power dissipation, is exponentially dependent on temperature

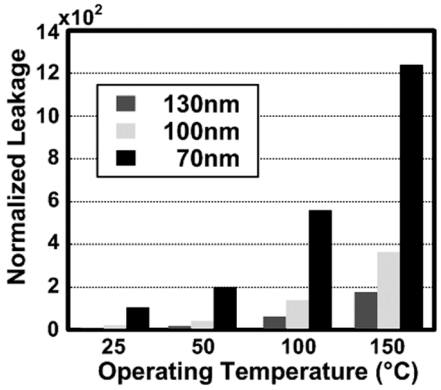


Fig. 6. Leakage power dissipation of an nMOS device for different technology nodes based on BSIM3 models showing the impact of temperature. The leakage power dissipation is normalized w.r.t. the value at 130-nm node at 25 °C.

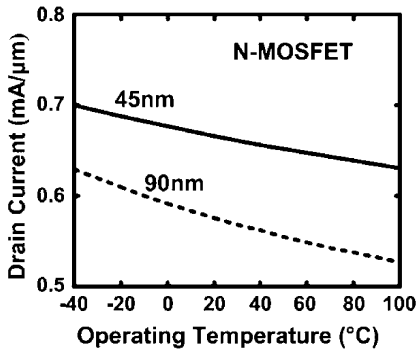


Fig. 7. Transistor drive (drain) current for N-MOSFET (45- and 90-nm effective channel length) based on BSIM3 models as a function of operating temperature.

and exacerbates with technology scaling [5], [7] (see Figs. 5 and 6). Moreover, transistor threshold voltage is a linear function of temperature, which in turn, depends on total power dissipation [38]. Dynamic power dissipation also depends on operating temperature due to the dependence of the transistor on-current on temperature. Although transistor threshold voltage decreases at higher operating temperature and partially offsets the performance degradation resulting from the lower carrier mobility, the transistor on-current still decreases at higher operating temperature (see Fig. 7).

Since power consumed by the ICs is converted into heat, the increase of operating temperature has significant impact on omnifarious design parameters and leads to various electrothermal couplings [8], which had been inconspicuous for earlier generations of ICs. Fig. 8 illustrates the details of various electrothermal couplings between performance, power dissipation, supply voltage, threshold voltage, and die temperature. Hence, it is crucial to incorporate electrothermal couplings when evaluating the power and delay [8]. The traditional ways of evaluating $P_{\rm dynamic}$ by (4) and $P_{\rm static}$ by (5) neglects these electrothermal couplings.

Recently, a methodology has been proposed to evaluate design parameters (mentioned in Fig. 8) in a self-consistent manner and then calculated an EEDP [29]. This methodology incorporates both analytical models and results from circuit simulation, based on an integrated device, circuit, and system

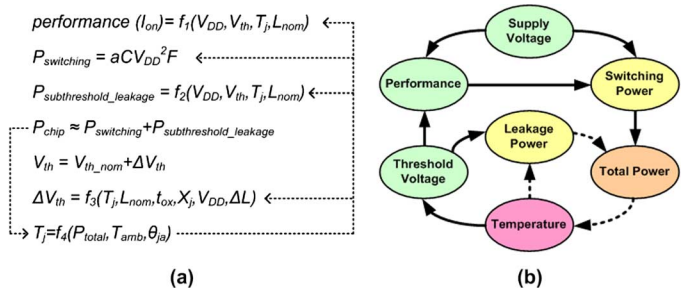


Fig. 8. (a) Models for various metrics are expressed in functional format. Couplings are indicated using broken lines. T_j represents the operating temperature, $L_{\rm nom}$ is the nominal gate length, a is the switching activity, C is the total load capacitance, F is the operating frequency, $V_{\rm th\_nom}$ denotes the nominal threshold voltage, $\triangle V_{\rm th}$ is the change of threshold voltage, $t_{\rm ox}$ is the gate oxide thickness, X_j is the junction depth, $\triangle L$ is the change of gate length due to variation, $T_{\rm amb}$ denotes the ambient temperature, and θ_{ja} denotes the equivalent junction-to-ambient thermal resistance. (b) Schematic view of electrothermal couplings between different design parameters. As technology scales, the couplings between power, leakage, and temperature (shown by dotted arrows) become increasingly prominent.

level modeling approach [8]. Incorporating the effect of interconnects, Fig. 9 is generated by the EEDP methodology using the same parameters in Table I. In Fig. 9(a), the line $(V_{\rm dd} = V_{\rm th})$ represents a boundary below which the operation is not considered. In practice, a maximum allowable operating temperature $(T_{\rm max})$ is determined mainly based on packaging/cooling and reliability requirements. Depending on the chip packaging and cooling solutions, maximum allowable power dissipation or thermal design power $(P_{\rm max})$ and the maximum temperature limit can be related as (11) by an equivalent junction-to-ambient thermal resistance (θ_{ja}). Here, a packaging and cooling solution with an equivalent junction-to-ambient thermal resistance $(\theta_{ja} = 0.7\,^\circ$C/W) is considered in the evaluation. The thermal runaway region shown in Fig. 9 (shaded region) indicates that the operation (set of $V_{\rm dd}$-$V_{\rm th}$ values) exceeds a temperature criterion $T_{\rm max}$ (set to be 200 °C in this analysis) due to the strong couplings between leakage and temperature

$$T_{\rm max} - T_{\rm amb} = \theta_{ja} \cdot P_{\rm max}. \tag{11}$$

In comparison with Fig. 4 (with interconnect effect) generated by the traditional method (without considering electrothermal couplings), it can be observed that not only the EDP contours and iso-performance curves shift but also the design space gets restricted by thermal constraint that cannot be known from Fig. 4. The traditional optimal point in Fig. 4 (marked by "■") is also shown in Fig. 9 for comparison. It can be observed that the optimal operation ($V_{\rm dd}$-$V_{\rm th}$ set) shifts to a new value ($V_{\rm dd} = 0.470$ V and $V_{\rm th} = 0.307$ V) marked by "●." Moreover, part of the optimal operation region suggested by the traditional method is not practical and results in high temperatures that exceed the maximum temperature criterion.

In Fig. 9(a), the iso-leakage curves (dotted lines) are superimposed and correspond to the ratios of leakage power to total power consumption. They essentially provide the limit of supply and threshold voltage scaling when the ratio of active to idle power is constrained. Moreover, as shown in Fig. 9(b), the iso-temperature curves (dotted lines) can be simultaneously obtained. These curves show the average junction temperature

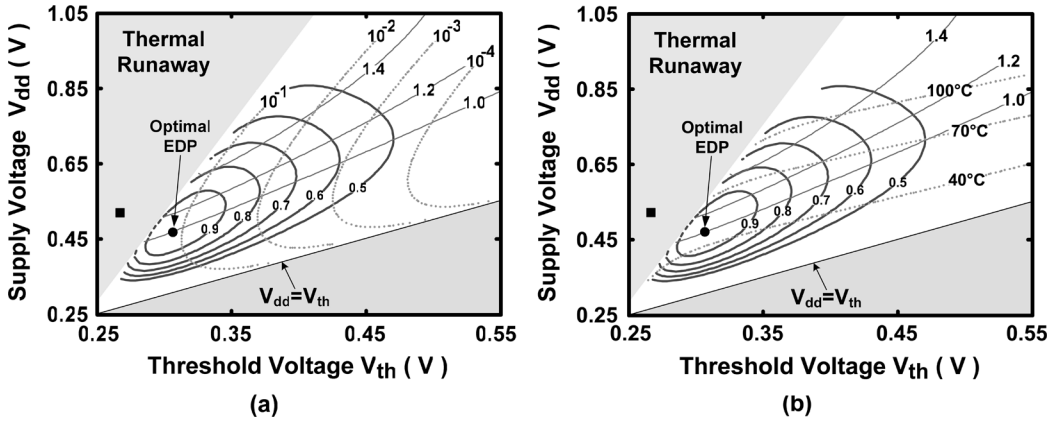


Fig. 9. $V_{\rm dd}$-$V_{\rm th}$ design space generated by the EEDP evaluation at 100-nm technology node. Note that the effect of interconnects has been incorporated. Five EDP contours (0.9, 0.8, 0.7, 0.6, 0.5) and three iso-performance curves (1.0, 1.2, 1.4) are shown for power-performance tradeoff analysis. "•" represents the optimal point based on EEDP evaluation. "■" indicates the operation set suggested by the traditional evaluation, but with interconnect effects (as shown in Fig. 4), for comparison. Note that the design space gets restricted by thermal constraint (thermal runaway) when electrothermal couplings are taken into account. (a) The iso-leakage curves $(10^{-1}, 10^{-2}, 10^{-3}, 10^{-4})$ have also been superimposed (dotted curve). The iso-leakage curves represent the ratio of leakage power to total power dissipation. (b) The iso-temperature curves (40 °C, 70 °C, 100 °C) have also been superimposed (dotted curve). The iso-temperature curves essentially provide an additional thermal (or reliability) constraint on the $V_{\rm dd}$-$V_{\rm th}$ design space.

estimation [by employing the relation in (11)] for various designs (different $V_{\rm dd}$-$V_{\rm th}$). The temperature information can be used as a thermal constraint because not only the power dissipation but many important reliability mechanisms are highly temperature sensitive. Hence, for EDP-based optimization, if electrothermal couplings are not considered, power dissipation and delay evaluations will be inaccurate and mislead the design optimization process.

B. Impact of Model Parameters on Optimal EDP Operation

As discussed in Section II, power and delay are two fundamental factors in the electrothermal EDP evaluation. In order to comprehend the impact of model parameters on the optimal operation, various parameters used in the power and delay models are discussed.

Power dissipation is modeled by two main factors. In the case of complex circuits, the switching activity [parameter a in (4)] depends on the application and input patterns. Increase in the switching activity directly increases the dynamic (switching) power dissipation. From (5), the leakage power at the nominal condition is modeled by the parameter I_s. As projected by ITRS, subthreshold leakage is expected to increase with technology scaling. Thus, larger I_s will be used for scenarios involving scaled technologies.

Although the effect of interconnects has been incorporated into the delay model in (10), the delay overhead due to the latches used in the pipelined logic is neglected in (3). In order to comprehend the impact of the delay overhead on optimal operation, a parameter (β) is defined as the ratio of the overhead time (due to the latches) to the delay T_g (i.e., $\beta = T_{\rm overhead}/T_g$). Thus, the cycle time can be calculated as $T_g \cdot (L_d + \beta)$.

Table II summarizes the optimal operation set evaluated by the EEDP methodology for different scenarios. When the switching activity is increased, it can be observed that the supply voltage of the optimal operation (with lowest energy-delay product) is decreased, which in turn, degrades the performance. When I_s is increased with technology scaling,

TABLE II
OPTIMAL OPERATION SET

a	I_s	β	Optimal Operation Point			
			V_{dd} (V)	V_{th} (V)	F_n	T_n
0.10	3 mA	0	0.498	0.309	1.157	- 3.1
0.15			0.470	0.307	1.000	0.0
0.30			0.448	0.310	0.806	+ 9.9

a	I_s	β	Optimal Operation Point			
			V_{dd} (V)	V_{th} (V)	F_n	T_n
0.15	3 mA	0	0.470	0.307	1.000	0.0
	5 mA		0.492	0.320	1.010	+ 2.3
	8 mA		0.510	0.335	0.988	+ 3.0

a	I_s	β	Optimal Operation Point			
			V_{dd} (V)	V_{th} (V)	F_n	T_n
0.15	3 mA	0	0.470	0.307	1.000	0.0
		5	0.487	0.308	0.876	- 1.4
		10	0.498	0.309	0.771	- 3.1

Note that F_n represents the normalized performance and T_n represents the temperature difference with respect to the case with a = 0.15, I_s = 3mA, and β = 0 (shaded case) in the table.

the suggested optimal operation has higher threshold voltage which compensates the performance enhancement from higher supply voltage. Severe performance degradation can be seen in the last scenario when the overhead factor (β) increases. It can be observed that the suggested optimal operation has lower performance even with higher supply voltages.

As mentioned in Section II-A, the short-circuit power has been ignored. Also, gate leakage (tunneling based) is neglected because it is temperature independent and can be mitigated by gate engineering [39]. However, if temperature independent power dissipation sources have to be considered, one can incorporate a constant amount into the total power evaluation in the proposed method. Impact of these additional power components on the proposed method will be similar to the case with higher switching activity (a) or I_s shown in Table II.

In Table II, only optimal operation point suggested by the electrothermal EDP evaluation is illustrated and compared. However, it is important to mention that if different design

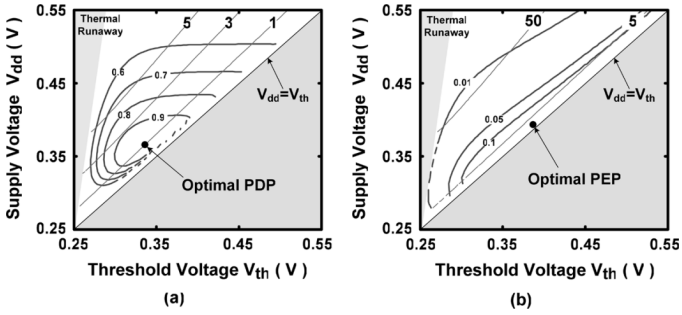


Fig. 10. $V_{\rm dd} - V_{\rm th}$ design space generated by electrothermally coupled evaluation. (a) Using PDP as the optimization metric. Four PDP contours (0.9, 0.8, 0.7, 0.5) and three iso-performance curves (1, 3, 5) are drawn for power-performance tradeoff analysis. (b) Using PEP as the optimization metric. Three PEP contours (0.1, 0.05, 0.01) and two iso-performance curves (5, 50) are drawn for power-performance tradeoff analysis.

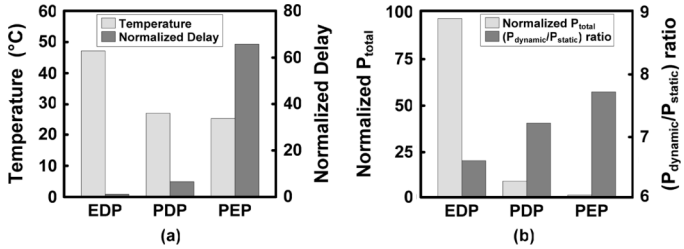


Fig. 11. Comparison of delay, temperature, and power between different optimization metrics. (a) Normalized delay (w.r.t. the EDP case) and average junction temperature corresponding to the optimal operating points obtained using three different optimization metrics (EDP, PDP, and PEP). (b) Normalized total power dissipation (w.r.t. the PEP case) and the $(P_{\rm dynamic}/P_{\rm static})$ ratio corresponding to optimal operating points of different optimization metrics.

TABLE III
OPTIMAL OPERATING POINTS OF DIFFERENT DESIGN METRICS

Design Parameters	Optimization Metric		
	EDP	PDP	PEP
$V_{\rm dd}$ (V)	0.470	0.365	0.393
$V_{\rm th}$ (V)	0.307	0.336	0.388

constraints (such as temperature or performance) have to be satisfied, the optimal operation based on EDP optimization will not be practical.

IV. METRICS FOR DESIGN-SPECIFIC OPTIMIZATION

In Section III, the impact of incorporating electrothermal couplings was clearly shown by using EDP as the design optimization metric. However, the selection of an appropriate metric determines the basis of optimization, i.e., the optimal point of operation will change if another optimization metric is chosen.

In this section, first the logic behind the use of different design metrics is explained through comparison between three general metrics [as shown in (7)] by the electrothermally coupled analysis. In practice, the optimal point (e.g., the lowest EDP point) is seldom used due to the need to satisfy other requirements like performance or power consumption which cannot be captured by that particular evaluation. Hence, a new optimization method is proposed and allows designers to choose a correct design metric that directly satisfies their design-specific needs. Comparisons based on the proposed appropriately chosen metric are more meaningful than those using a single design metric, for example EDP, which does not comprehend design-specific requirements.

Fig. 10(a) and (b) show the $V_{\rm dd}$-$V_{\rm th}$ design spaces of the PDP-based and PEP-based electrothermally coupled analyses, respectively, for the same design and technology node used in Fig. 9. As expected, the contours, iso-performance curves, and the optimal operating points based on these two design metrics (PDP and PEP) are different. Table III summarizes the optimal points of different design metrics. Note that the optimal point suggested from the evaluation corresponds to a set $(V_{\rm dd}$-$V_{\rm th})$ which has a minimum product (e.g., when PEP is the design metric, the optimal operating point results in a minimum power-energy product).

From (7), it is clear to see that the EDP metric prioritizes delay (proportional to T_g^2) over power (proportional to P^1). When EDP is the design metric, the optimal operating point will have higher supply voltage and lower threshold voltage, as seen in Table III, in order to have relatively higher performance. On the other hand, the PEP metric prioritizes power (proportional to P^2) over delay (proportional to T_g^1). Thus, the optimal operating point has larger threshold voltage to reduce the leakage power dissipation.

Impact of optimization metric is illustrated in Fig. 11 from the perspectives of delay, temperature, and power dissipation. Note that the optimal $V_{\rm dd}$-$V_{\rm th}$ sets for these three optimization metrics are compared. It can be observed from Fig. 11(a) that PEP leads to the highest delay as compared to other metrics. However, the total power dissipation for PEP as shown in Fig. 11(b) is the lowest. Moreover, PEP will have the highest ratio of $P_{\rm dynamic}$ to $P_{\rm static}$ that indicates the highest power efficiency for a design.

As shown in the preceding discussion, the relative emphasis on power dissipation and performance, and thus the optimization metric, need to be changed depending on design-specific requirements. A change in the optimization metric has a significant impact on design choices. However, there is no systematic methodology existing in the literature to guide the designer to intelligently choose an appropriate optimization metric that satisfies all the design requirements.

In order to comprehend the varying requirements of different designs, a generalized optimization metric based on power and delay is needed. Here, the parameter "μ" is used to represent the ratio of exponent of delay to that of power. The generalized metric thus is represented as $\mathrm{P}(\mathrm{T}_g)^{\mu}$. Existing metrics such as power-energy product and energy-delay product can also be represented by choosing μ as 0.5 and 2, respectively. Since design optimization is carried out by finding the minimum value of the product $\mathrm{P}(\mathrm{T}_g)^{\mu}$, naturally, μ should be larger than 1 if performance is the primary concern, i.e., an optimization metric with a higher μ will lead to higher performance than that with a lower μ. On the contrary, when power dissipation is the primary concern, μ should be less than 1.

Fig. 12 shows the locus formed by the optimal operating points obtained for different μ with the same design and technology node used in Fig. 9. Each point on the optimal operation locus represents the $V_{\rm dd}$-$V_{\rm th}$ set which has the lowest value of

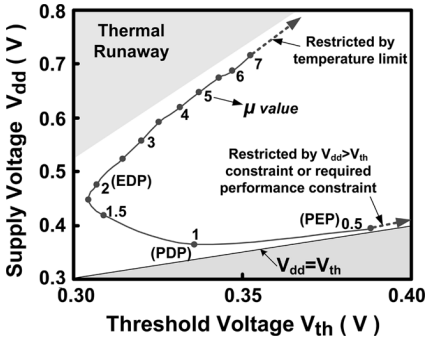


Fig. 12. Optimal operation locus for 100-nm technology node. The optimal locus is formed by the optimal points (denoted by "•") with respect to different values of μ, i.e., the minimal values of the product $\mathrm{P}(\mathrm{T}_g)^{\mu}$. Although values of μ between 0.5 and 7 are shown in the plot, μ can be chosen to be any positive rational number depending on design space and constraints. The shaded regions in the two corners correspond to thermal runaway region and the region where the supply voltage is less than the threshold voltage.

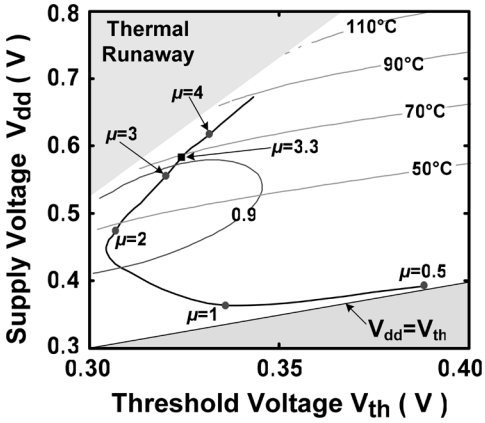


Fig. 13. Illustration of the methodology for finding a suitable optimization metric to meet design-specific requirements. "•" indicates the optimal operating points with different optimization metrics ($\mu = 0.5, 1, 2, 3, 4$). EDP ($\mu = 2$) contour for $\mathrm{EDP} = (1/0.9)\mathrm{EDP}_{\mathrm{opt}}$, and iso-temperature curves are shown at 100-nm technology node. "■" indicates the optimal point that meets all design-specific requirements.

the product $\mathrm{P}(\mathrm{T}_g)^{\mu}$. Moreover, larger power dissipation results in higher temperature, which in turn, leads to many important design concerns especially for nanometer technologies. Fig. 12 superimposes the thermal constraint (shaded region for thermal runaway) and the operational constraint ($V_{\mathrm{dd}} > V_{\mathrm{th}}$) onto the trend of optimal operating locus. Depending on the packaging and cooling technologies available for a particular design, the upper bound of μ is provided by the maximum allowable operating temperature limit. On the other hand, lower bound of μ is defined by the minimum required performance.

Traditionally, designers might choose EDP ($\mu = 2$) as an optimization metric for trading-off performance and power dissipation. As seen in Fig. 12, EDP provides medium performance and medium power dissipation as compared to other values of μ. When designers want to lay higher emphasis on performance, preferably, μ should be chosen to be higher than 2. On the other hand, when more emphasis is placed on low power, μ should be less than 2. Note that for low power applications, the optimal point shifts by a larger amount for a certain change in μ, whereas for high performance applications the corresponding shift is much smaller. This is because leakage power dissipation, which is a major contributor to total power dissipation in nanometer CMOS technologies, exponentially depends on the threshold voltage and temperature. Hence, the choice of operating point becomes very sensitive to threshold voltage when the designer gives more weightage to power. Also, it is important to mention that the optimal operating point is only considered in the region where supply voltage is larger than threshold voltage, i.e., subthreshold operation is not considered in this analysis. This is because of the validity of the Alpha-Power model [31] for $V_{\mathrm{dd}} > V_{\mathrm{th}}$. In this analysis, when μ is less than 0.5 (PEP), the trend of optimal point is very close to the point where the supply voltage is equal to threshold voltage and the optimal locus will follow the $V_{\mathrm{dd}} = V_{\mathrm{th}}$ line.

The question that arises is how does a designer choose to lay a particular emphasis on power vis-à-vis performance? Can there be changing scenarios where the design-specific requirements are beyond those comprehended by traditional metrics such as the most commonly used EDP? Finally, under such requirements, why is it that the proposed metric leads to better design solutions than a traditional metric like EDP? These are the questions addressed and discussed in the subsequent examples of design optimization.

V. Example of Design-Specific Optimization

In this section, design-specific optimization is demonstrated by the proposed methodology using a general homogeneous logic block.

Consider a logic block at 100-nm technology node for optimization. As described in Section II, a logic block can be represented by an equivalent inverter with effective transistor width, load capacitance, and activity factor [14]. The critical path delay of the block is also modeled as (10) with parameters shown in Table I. Note that a uniform (average) temperature over this logic block is evaluated for simplicity. The maximum allowable operating temperature is assumed to be 70 °C, for example, due to packaging and cooling limitations. The target is to achieve the maximum possible performance under this maximum temperature constraint.

Since the design objective is to maximize the performance, a desirable metric would have the highest possible μ under the maximum temperature constraint. It can be observed from Fig. 13 that the appropriate μ is at the intersection of the 70 °C iso-temperature curve and the optimal operation locus. For the case shown in Fig. 13, the intersection occurs at $\mu = 3.3$. Once the operating temperature value is set to be 70 °C as a constraint, the value of parameter "μ" can be directly obtained by the electrothermally coupled analysis (do not require additional computation for obtaining the parameter μ).

Given the same constraint as mentioned before, two possible design choices are considered and depicted by points A and B in Fig. 14 and the designer needs to decide which of these options best fits the design requirements. The result obtained from a comparison of these two design choices based on the proposed new metric $\mathrm{P}(\mathrm{T}_g)^{3.3}$ is compared to that based on energy-delay product (EDP), which is the most widely used design metric. For the EDP metric, the optimal point $(\mathrm{EDP}_{\mathrm{opt}})$ and a corresponding suboptimal contour of all points where the ratio

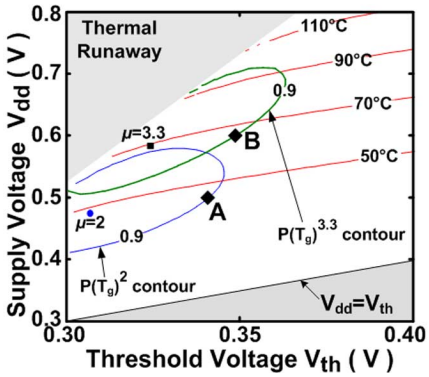


Fig. 14. Example comparing the use of the proposed metric $\mathrm{P(T}_g)^{3.3}$ in choosing between two design options (A and B) against that of conventional EDP evaluation. "•" indicates EDP ($\mu = 2$) optimal operating point and "■" indicates the optimal point that meets all design-specific requirements. Since design optimization is preferably to find an operation set closest to the optimal point, it is clear that the choice between the two points A and B changes depending on the metric of optimization chosen.

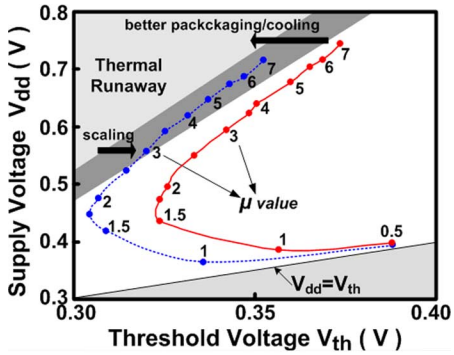


Fig. 15. Scaling analysis of optimal operating points obtained by applying different optimization metrics (shown for 100- and 70-nm technology nodes). Note that the thermal runaway region expands to the right due to technology scaling.

$\mathrm{EDP_{opt}/EDP} = 0.9$, are shown. All points outside this contour have EDP higher than those for the points that lie inside this contour. Hence a traditional comparison based on the EDP metric would lead to the decision that A is a better choice than B. If the optimal point corresponding to the metric $\mathrm{P(T}_g)^{3.3}$ (which captures the design-specific requirements) and the suboptimal 0.9 contour surrounding this point are considered, it can be observed that the value of the metric $\mathrm{P(T}_g)^{3.3}$ at point B is smaller than the value at point A. Hence, based on the new metric, design B should be chosen over design A. Evidently, the choice between the two points A and B changes depending on the metric of optimization chosen. Hence, when the additional requirement of having highest possible performance under the maximum temperature constraint is factored in, option B is obviously the better choice.

As demonstrated by the previous example, once the parameter "μ" is determined by the proposed methodology, the appropriate metric $\mathrm{P(T}_g)^{\mu}$ can capture design-specific requirements. Hence, a procedure similar to EDP evaluation (replacing the quantity $\mathrm{P(T}_g)^2$ with $\mathrm{P(T}_g)^{\mu}$) can be used to compare various designs having the same requirements and belonging to the same design family. The metric selected by this methodology provides a more meaningful basis for making design choices under these particular design-specific requirements.

Typically, the timing-critical path(s) can be determined after transistor level design. It is important to note that the proposed methodology is applicable for any size of homogeneous logic block by modeling the delay using an equivalent inverter (with effective transistor width and activity factor) driving a capacitive load and by modeling the total power dissipation at the nominal temperature. In case of a circuit with multiple timing-critical paths, the proposed methodology can either be applied to the longest path for optimization or one can employ multiple equivalent inverters corresponding to each path and then optimize one at a time.

The proposed method is also applicable to designs with multiple threshold voltages at different circuit blocks. By appropriately dividing the circuit into different blocks (each block can be represented by an equivalent inverter), one can optimize each block and determine the threshold voltage for each block separately.

Moreover, for a heterogeneous circuit design, including fully pipelined microprocessors, different functional blocks might have different activities, supply voltages, or threshold voltages for improving performance as well as reducing power dissipation. In this case, although one can still transfer the entire complex circuit into an equivalent inverter and optimize it by the proposed method, it is not practical, and will not benefit as much from the proposed optimization. However, such designs can be easily handled in the proposed methodology by dividing the complex circuit into several homogeneous logic blocks for optimization.

VI. Impact of Technology Scaling and Parameter Variations

Continued scaling of CMOS technologies provides substantial benefits in transistor density and circuit performance. However, the corresponding increase in power consumption directly impacts the junction temperature that determines the upper limit of μ. It can be observed from Fig. 15 that the optimal operation locus shifts to the right when technology scales from 100- to 70-nm nodes. The same design at 70-nm technology node has higher optimal values for threshold voltages due to the increase of leakage power dissipation (see Fig. 6). Moreover, due to technology scaling and the resultant increasing leakage, it can be clearly seen that the design space gets increasingly restricted by the thermal constraint (i.e., the thermal runaway region expands). However, with better packaging/cooling solutions, the design space can be relaxed, i.e., the thermal runaway region shrinks.

Fig. 16 shows the impact of technology scaling on selecting μ for design-specific optimization. Under the same constraints as used in the example in Section V (see Fig. 13), it is observed that if the same optimization metric $\mathrm{P(T}_g)^{3.3}$ is chosen for 70-nm technology node, the optimal operating point exceeds the maximum allowed temperature (70 °C). For the 70-nm technology node, the proper optimization metric that meets the design-specific requirements is found to be $\mathrm{P(T}_g)^3$ by the proposed methodology.

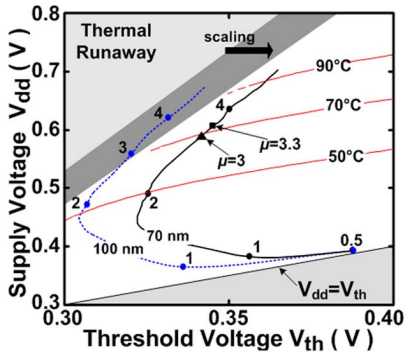


Fig. 16. Impact of technology scaling on the proposed optimization methodology. The optimal operation locus of 100-nm technology node is shown by a dotted curve (the same locus shown in Fig. 13) while the locus of 70-nm technology node with the same design is shown by a solid curve. Three iso-temperature curves (50 °C, 70 °C, 90 °C) are superimposed. "•" indicates the optimal operating points with different optimization metrics. "■" indicates the optimal operating point when $\mathrm{P}(\mathrm{T}_g)^{3.3}$ is chosen for 70-nm technology node. "▲" indicates the optimal operating point when $\mathrm{P}(\mathrm{T}_g)^3$ is chosen. Note that the metric $(\mu = 3)$ meets all design-specific requirements at 70-nm technology node.

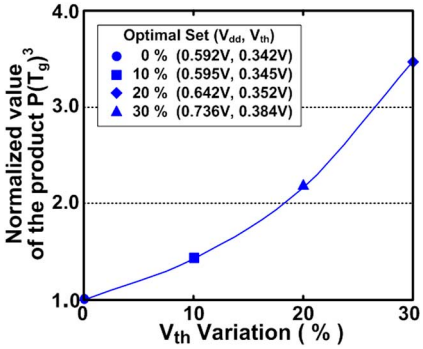


Fig. 17. Impact of threshold voltage variation on the optimization metric $(\mu = 3)$ for 70-nm technology node. The values shown are normalized to the corresponding optimal values without threshold voltage variations (V_{th} variation $= 0\%$). The inset shows the optimal operating set (V_{dd}-V_{th} set) corresponding to the different threshold voltage variations shown in the main figure using identical symbols.

Besides technology scaling, parameter variations, especially within-die variations that arise either from environmental variations (e.g., temperature, supply voltage, etc.) or from physical process variations (channel length, oxide thickness, random dopant fluctuation, etc.) can result in an uncertainty in the power and performance estimation [24] and thereby impact the choice of μ. For the same example discussed above, Fig. 17 shows the impact of threshold voltage variations on the values of the optimization metrics $\mathrm{P}(\mathrm{T}_g)^{\mu}$ obtained using the proposed methodology. Note that this evaluation is carried out at 70-nm technology node where μ is 3 (refer to Fig. 16). The inset of Fig. 17 lists the optimal points corresponding to different amount of V_{th} variation. It can be observed that for the specific requirements of this design, the optimal operating point of the proposed metric shifts with V_{th} variation, and the difference increases as variations become larger. Consequently, it is crucial to consider design-specific requirements as well as parameter variations for appropriate design optimization and comparison between different design choices.

VII. Conclusion

In this paper, a systematic methodology has been proposed to select appropriate design-specific metrics for simultaneous optimization of power and performance in leakage dominant CMOS technologies. The methodology incorporates interconnect effects as well as electrothermal couplings between various design parameters such as power, operating temperature, and performance. It is demonstrated that the metric evaluated by the proposed methodology provides a more meaningful basis to optimize supply and threshold voltages under design-specific constraints as compared to traditional methodologies. While design tradeoffs are traditionally made using electrical parameters, this work introduced a new technique by which circuit designers can comprehend the implications of design choices on chip-scale thermal management issues including maximum allowable operating temperature and packaging/cooling solutions and vice-versa. In addition, it was shown that the design-specific optimization metrics need to be adaptive to increasing leakage and process variations with technology scaling.

References

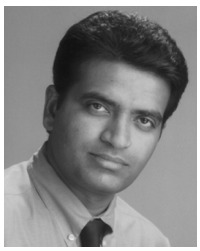
[1] G. E. Moore, "Cramming more components onto integrated circuits," *Electronics*, vol. 38, no. 8, pp. 114–117, 1965.
[2] International Technology Roadmap for Semiconductors (ITRS) 2007 [Online]. Available: www.itrs.net
[3] A. P. Chandrakasan and R. W. Brodersen, "Minimizing power consumption in digital CMOS circuits," *Proc. IEEE*, vol. 83, no. 4, pp. 498–523, Apr. 1995.
[4] K. S. Yeo and K. Roy, *Low Voltage, Low Power VLSI Subsystems*. New York: McGraw-Hill, 2004.
[5] V. De and S. Borkar, "Technology and design challenges for low power and high performance," in *Proc. Int. Symp. Low Power Electron. Des.*, 1999, pp. 163–168.
[6] P. P. Gelsinger, "Microprocessors for the new millennium: Challenges, opportunities, and new frontiers," in *Proc. Int. Solid-State Circuits Conf.*, 2001, pp. 22–25.
[7] P. P. Gelsinger, "Gigascale integration for teraops performance, challenges, opportunities and new frontiers," in *Proc. 41st DAC Keynote, Des. Autom. Conf.*, 2004, p. xxv.
[8] K. Banerjee, S.-C. Lin, A. Keshavarzi, S. Narendra, and V. De, "A self-consistent junction temperature estimation methodology for nanometer scale ICs with implications for performance and thermal management," in *Proc. IEEE Int. Electron Devices Meet.*, 2003, pp. 887–890.
[9] C.-K. Hu, R. Rosenberg, H. S. Rathore, D. B. Nguyen, and B. Agarwala, "Scaling effect on electromigration in on-chip Cu wiring," in *Proc. IEEE Int. Interconnects Technol. Conf.*, 1999, pp. 267–269.
[10] R. Blish, T. Dellin, S. Huber, M. Johnson, J. Maiz, B. Likins, N. Lycoudes, J. McPherson, Y. Peng, C. Peridier, A. Preussger, G. Prokop, and L. Tullos, "Critical reliability challenges for the internatinonal technology roadmap for semiconductors," Int. Sematech Technol. Transfer Doc. 03024377A-TR, 2003.
[11] A. M. Yassine, H. E. Nariman, M. McBride, M. Uzer, and K. R. Olasupo, "Time dependent breakdown of ultra-thin gate oxide," *IEEE Trans. Electron Devices*, vol. 47, no. 7, pp. 1416–1420, Jul. 2000.
[12] S.-C. Lin, A. Basu, A. Keshavarzi, V. De, A. Mehrotra, and K. Banerjee, "Impact of off-state leakage current on electromigration design rules for nanometer scale CMOS technologies," in *Proc. Int. Reliab. Phys. Symp.*, 2004, pp. 74–78.
[13] M. A. Horowitz, T. Indermaur, and R. Gonzalez, "Low power digital design," in *Proc. Int. Symp. Low Power Electron. Des.*, 1994, pp. 8–11.
[14] R. Gonzalez, B. M. Gordon, and M. A. Horowitz, "Supply and threshold voltage scaling for low power CMOS," *IEEE J. Solid-State Circuits*, vol. 32, no. 8, pp. 1210–1216, Aug. 1997.
[15] K. Nose and T. Sakurai, "Optimization of Vdd and Vth for low power and high speed applications," in *Proc. Asia South Pacific Des. Autom. Conf.*, 2000, pp. 469–474.
[16] J. Rabaey, A. Chandrakasan, and B. Nikolić, *Digital Integrated Circuits: A Design Perspective*. Englewood Cliffs, NJ: Prentice Hall, 2003.

[17] A. J. Martin, "Towards an energy complexity of computation," *Inf. Process. Lett.*, pp. 181–187, 2001.
[18] P. I. Pénzes and A. J. Martin, "Energy-delay efficiency of VLSI computations," in *Proc. Great Lakes Symp. VLSI*, 2002, pp. 104–111.
[19] H. P. Hofstee, "Power-constrained microprocessor design," in *Proc. IEEE Int. Conf. Comput. Des.*, 2002, pp. 14–16.
[20] V. Zyuban and P. N. Strenski, "Balancing hardware intensity in microprocessor pipelines," *IBM J. Res. Develop.*, vol. 47, pp. 585–598, 2003.
[21] V. Zyuban and P. N. Strenski, "Unified methodology for resolving power-performance tradeoffs at the microarchitectural and circuit levels," in *Proc. Int. Symp. Low Power Electron. Des.*, 2002, pp. 166–171.
[22] D. Marković, V. Stojanović, B. Nikolić, M. A. Horowitz, and R. W. Brodersen, "Methods for true energy-performance optimization," *IEEE J. Solid-State Circuits*, vol. 39, no. 8, pp. 1282–1293, Aug. 2004.
[23] P. Pant, V. De, and A. Chatterjee, "Simultaneous power supply, threshold voltage, and transistor size optimization for low-power operation of CMOS circuits," *IEEE Trans. Very Large Scale Integr. (VLSI) Syst.*, vol. 6, no. 4, pp. 538–545, Dec. 1998.
[24] S. Borkar, T. Karnik, S. Narendra, J. Tschanz, A. Keshavarzi, and V. De, "Parameter variations and impact on circuits and microarchitecture," in *Proc. Des. Autom. Conf.*, 2003, pp. 338–342.
[25] Y.-S. Lin, C.-C. Wu, C.-S. Chang, R.-P. Yang, W.-M. Chen, J.-J. Liaw, and C. H. Diaz, "Leakage scaling in deep submicron CMOS for SoC," *IEEE Trans. Electron Devices*, vol. 49, no. 6, pp. 1034–1041, Jun. 2002.
[26] H. Soeleman, K. Roy, and B. C. Paul, "Robust subthreshold logic for ultra-low power operation," *IEEE Trans. Very Large Scale Integr. (VLSI) Syst.*, vol. 9, no. 1, pp. 90–99, Feb. 2001.
[27] A. Wang, A. Chandrakasan, and S. Kosonocky, "Optimal supply and threshold scaling for subthreshold CMOS circuits," in *Proc. IEEE Comput. Soc. Annu. Symp. VLSI*, 2002, pp. 5–9.
[28] D. Sengupta and R. Saleh, "Power-delay metrics revisited for 90 nm CMOS technology," in *Proc. Int. Symp. Quality Electron. Des.*, 2005, pp. 291–296.
[29] A. Basu, S.-C. Lin, V. Wason, A. Mehrotra, and K. Banerjee, "Simultaneous optimization of supply and threshold voltages for low-power and high-performance circuits in the leakage dominant era," in *Proc. Des. Autom. Conf.*, 2004, pp. 884–887.
[30] S.-C. Lin, N. Srivastava, and K. Banerjee, "A thermally aware methodology for design-specific optimization of supply and threshold voltages in nanometer scale ICs," in *Proc. IEEE Int. Conf. Comput. Des.*, 2005, pp. 411–416.
[31] T. Sakurai and A. R. Newton, "Alpha-power law MOSFET model and its application to CMOS inverter delay and other formulas," *IEEE J. Solid-State Circuits*, vol. 25, no. 2, pp. 584–593, Apr. 1990.
[32] C. Rice, "Introducing the Intel Pentium 4 Processor," *Intel Developer Update Mag*, 2000 [Online]. Available: www.intel.com
[33] A. Chatterjee, M. Nandakumar, and I. C. Chen, "An investigation of the impact of technology scaling on power wasted as short-circuit current in low voltage static CMOS circuits," in *Proc. IEEE Int. Symp. Low Power Electron. Des.*, 1996, pp. 145–150.
[34] K. Banerjee and A. Mehrotra, "A power-optimal repeater insertion methodology for global interconnects in nanometer designs," *IEEE Trans. Electron Devices*, vol. 49, no. 11, pp. 2001–2007, Nov. 2002.
[35] A. P. Chandrakasan and R. W. Brodersen, "Sources of power consumption," in *Low Power Digital CMOS Design*. Norwell, MA: Kluwer, 1995.
[36] W. C. Elmore, "The transit analysis of damped linear networks with particular regard to wideband amplifiers," *J. Appl. Phys.*, vol. 19, no. 1, p. 55, 1948.
[37] H. B. Bakoglu, *Circuits, Interconnections, and Packaging for VLSI*. Boston, MA: Addison-Wesley, 1990.
[38] Y. Taur and T. H. Ning, *Fundamentals of Modern VLSI Devices*. Cambridge, U.K.: Cambridge Univ. Press, 1998.
[39] P. M. Zeitzoff, "MOSFET scaling trends and challenges through the end of the roadmap," in *Proc. Custom Integr. Circuits Conf.*, 2004, pp. 233–240.

Sheng-Chih Lin (S'03–M'08) received the B.S. degree in electrical engineering from the National Taiwan University, Taipei, Taiwan, in 1996, and the Ph.D. degree in electrical and computer engineering from the University of California, Santa Barbara, in 2007 under the tutelage of Prof. K. Banerjee.

In February 2008, he joined the Assembly and Test Technology Development, Intel, Chandler, AZ. From 1998 to 2002, he was with the Phoenixtec Electronics Company, Ltd., and the CHROMA ATE Inc., respectively, in Taiwan. During the summer of 2005 and 2006, he worked as an intern in the Assembly and Test Technology Development, Intel, Chandler, AZ. His research interests include electrothermal modeling and analysis of integrated circuits, variation-aware circuit design and optimization, and power/thermal management for nanoscale CMOS ICs. He has authored or coauthored over a dozen papers in journals and refereed international conferences.

Mr. Lin is a corecipient of an IEEE Micro Top Picks Award in 2006.



Kaustav Banerjee (S'92–M'99–SM'03) received the Ph.D. degree in electrical engineering and computer sciences from the University of California, Berkeley, in 1999.

In July 2002, he joined the Faculty of the Department of Electrical and Computer Engineering, University of California, Santa Barbara, where he has been a Full Professor since 2007. From 1999 to 2001, he was a Research Associate with the Center for Integrated Systems, Stanford University, Stanford, CA. From February to August 2002, he was a Visiting Faculty with the Circuit Research Laboratories, Intel, Hillsboro, OR. He has also held summer/visiting positions at Texas Instruments Incorporated, Dallas, TX, from 1993 to 1997, and the Swiss Federal Institute of Technology, Lausanne, Switzerland, in 2001. His research has been chronicled in over 150 journal and refereed international conference papers and in two book chapters. He has also coedited the book *Emerging Nanoelectronics: Life with and after CMOS* (Springer, 2004). His current research interests focus on nanometer-scale issues in high-performance/low-power VLSI as well as on circuits and systems issues in emerging nanoelectronics.

Dr. Banerjee was a recipient of a number of awards in recognition of his work, including the ACM SIGDA Outstanding New Faculty Award in 2004, a Best Paper Award at the Design Automation Conference in 2001, an IEEE Micro Top Picks Award in 2006, and an IBM Faculty Award in 2008. He served on the technical program committees of several leading IEEE and ACM conferences, including IEDM, DAC, ICCAD, and IRPS. He has also served on the organizing committee of ISQED, at various positions including Technical Program Chair in 2002 and General Chair in 2005. Currently, he serves as a member of the Nanotechnology Committee of the IEEE Electron Devices Society.