A Thermally-Aware Methodology for Design-Specific Optimization of Supply and Threshold Voltages in Nanometer Scale ICs

Sheng-Chih Lin, Navin Srivastava and Kaustav Banerjee
Department of Electrical and Computer Engineering, University of California, Santa Barbara, CA 93106
E-mail: {sclin, navins, kaustav}@ece.ucsb.edu

ABSTRACT

As CMOS technology scales deeper into the nanometer regime, factors such as leakage power and chip temperature emerge as critically important concerns for VLSI design. This paper, for the first time, proposes a systematic methodology to determine a generalized design metric for simultaneously optimizing power and performance in nanometer-scale integrated circuits to achieve design-specific targets while incorporating electrothermal effects. This methodology is shown to provide a more meaningful basis to compare different design choices. The implications of technology scaling and parameter variations on this thermally-aware methodology are also presented.

1. INTRODUCTION

In the past two decades, the steady downscaling of transistor dimensions has ensured higher packing density, higher performance, and lower cost of integrated circuits [1]. The efforts of technology scaling have been focused on achieving highest performance. In recent years, power constraint has become an important issue for circuit designers. Many hand-held devices including wireless applications require low power design due to a limited battery budget. Also, the power dissipation and associated thermal effects have strong impact on the packaging, cooling costs, and reliability for deep submicron technologies [2-5].

For power-constrained applications, lowering supply voltage (V_{dd}) offers the biggest potential to decrease the active 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_{th}) at the same time, but then the subthreshold leakage power increases exponentially. Consequently, the need for low power and high performance circuit applications motivates the search for an optimal set of supply and threshold voltages to tradeoff 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, they have a direct impact on gate-oxide and hot carrier reliability [6-8] and an indirect impact on electromigration reliability through the junction temperature [9].

Several methodologies have been proposed in the literature to 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 [10][11] for evaluating tradeoffs of power and performance. 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 [10-12]. General metrics for improving the energy-delay efficiency have also been explored. In [13], Pénzes and Martin showed that the Et^n metric characterizes any feasible trade-off. Hofstee [14] conclude that optimal metric is not unique for all designs but depends on the desired level of performance. 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 and threshold voltages and gate sizing to achieve an energy-efficient design. Zyuban and Strenski [15][16] use "hardware intensity" to quantify the relative cost of enhancing performance and resultant power dissipation at the circuit and micro-architecture levels. Markovic′ et al. [17] analyze the ratio of sensitivity of energy to the sensitivity of delay in order to achieve energy-performance optimization. However, these works do not comprehend the interdependence of thermal and power dissipation issues which become critical in nanometer scale designs, as discussed below.

Due to technology scaling and parameter variations [18], leakage power dissipation, which is dominated by subthreshold leakage for high-performance ICs, becomes a significant component of total chip power consumption [2][19]. The subthreshold leakage is exponentially dependent on temperature and the dependence gets stronger with 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 [5]. 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.

Contribution of This Work:

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 toward the deep nanometer scale. Although there is evidence of the increasing use of different optimization metrics [20-22] in the existing literature, 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 that captures the relative importance of power dissipation and performance to achieve design-specific targets as they change from one technology generation to the next. The advantage of the proposed thermally-aware methodology as compared to the traditionally used optimization metrics is discussed and it is shown to provide a more meaningful basis to optimize supply and threshold voltages.

The paper is organized as follows. In Section 2, we begin with a review of design parameters and metrics including power, energy, and delay using both traditional and a thermally-aware EDP metric as an example. In Section 3, we present a comparative analysis of three commonly used optimization metrics, using the electrothermally coupled methodology [5] that takes temperature dependence into account. In Section 4, we present the methodology for selecting a design-specific optimization metric. The impact of this methodology on the optimization is shown through circuit and system level examples of design optimization. Scaling and parameter variations are known to significantly impact on leakage power dissipation. In Section 5, we show the implications of this methodology for CMOS technology scaling as well as for parameter variations. Finally, concluding remarks are made in Section 6.

2. DESIGN PARAMETERS AND METRICS REVISITED

The critical path of a chip normally goes through a variety of gates each with a different value of delay. However, changes in supply voltage, temperature and threshold voltage affect all gates in the same way so that delay of any gate remains roughly proportional to the delay of an inverter [11]. The average delay of an inverter (T_g) can be estimated by

the Alpha-Power model [23] as shown in (1). The parameter α accounts for velocity saturation condition of the transistors and is between one (complete velocity saturation) and two (no velocity saturation) while K is a proportionality constant specific to a given technology. The maximum operating frequency (f) of the chip is given by (2) where the parameter L_d is the logic depth. For most of the modern microprocessors, L_d is usually around 20 [24].

$$T_g = K \frac{V_{dd}}{(V_{dd} - V_{th})^{\alpha}} \quad (1)$$

$$f = \frac{1}{T_g L_d} \quad (2)$$

There are two main sources of power dissipation in CMOS circuits: dynamic (switching) and static (leakage). Dynamic power results from the charging and discharging 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; therefore we could ignore it throughout this paper. The total dynamic ($P_{dynamic}$) and static (P_{static}) power consumption per operation of a chip thus can be written as (3) and (4) respectively.

$$P_{dynamic} = aC_{eff}V_{dd}^2 f \quad (3)$$

$$P_{static} = I_s e^{-V_{th}/\gamma V_o}(1 - e^{-V_{ds}/\gamma V_o})W_{eff}V_{dd} \quad (4)$$

where a is the activity factor of the output node, and C_{eff} accounts for the total capacitance of the output node. I_s is the zero-threshold leakage current, γ is subthreshold slope factor, V_0 is the subthreshold slope, and W_{eff} is the effective transistor width (transistor width that contributes to the leakage current) of the gate cell.

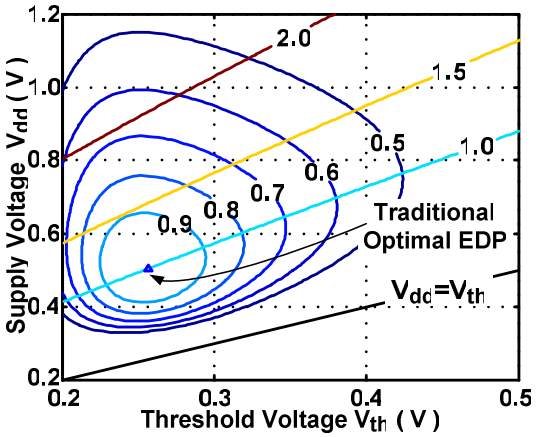

Fig. 1 Traditional optimization uses EDP as a design metric. Here, the EDP contours and performance curves are obtained by simple numerical solution without considering electrothermal couplings between temperature and static power dissipation for 100 nm technology node.

Traditionally, the design metric used to minimize both power and delay of a circuit is the energy-delay product (EDP) [10]. **Fig. 1** has been generated simply by direct numerical evaluation of energy and delay for a specific design. The EDP contours can be found by normalizing with respect to the value of the EDP at the optimal point (V_{dd} = *0.504 V* and V_{th} = *0.257 V*). For instance, any point on the curve labeled *0.5* has an EDP value twice that of optimal (EDP = 2·EDP_{opt}), i.e., minimum value. The 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. Note that the traditional EDP evaluation does not consider the region where circuits operate in subthreshold mode. Besides energy-delay product (EDP), two other design metrics are also used for different applications: Power-delay product (PDP) and power-energy product (PEP). The PDP gives identical weightage to power and delay while the PEP prioritizes power above delay. In all of these metrics, power and delay are the two fundamental parameters and the metric to be chosen depends on the design optimization goal. The relationships between power (P), delay (T) and these three metrics are shown in (5).

$$\begin{aligned} \text{EDP} &= \text{Energy} \cdot \text{Delay} = \text{PT}^2 \\ \text{PDP} &= \text{Power} \cdot \text{Delay} = \text{PT} \\ \text{PEP} &= \text{Power} \cdot \text{Energy} = \text{P}^2\text{T} \end{aligned} \quad (5)$$

Fig. 2 shows the scaling trend of supply voltage, threshold voltage, and subthreshold leakage current. It can be seen that the leakage power increases substantially as technology scales. Also, the leakage power, which is becoming a major source of total power dissipation [2][4], is exponentially dependent on temperature and the dependence gets stronger with scaling (**Fig. 3**). Moreover, V_{th} is a function of temperature, which in turn, depends on total power dissipation. Hence, it is crucial to incorporate electrothermal couplings when evaluating the power and delay [5]. The traditional way to evaluate $P_{dynamic}$ by (3) and P_{static} by (4) neglects these electrothermal couplings.

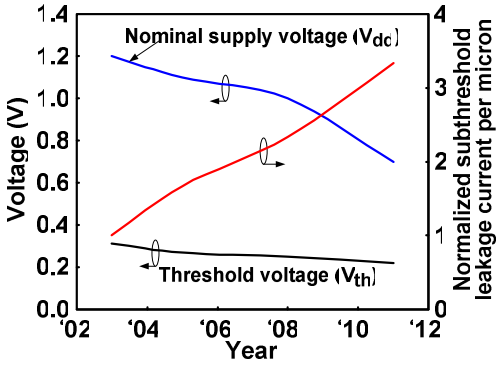

Fig. 2 Trend of nominal supply voltage, threshold voltage and leakage current per micron based *ITRS 2004* [1].

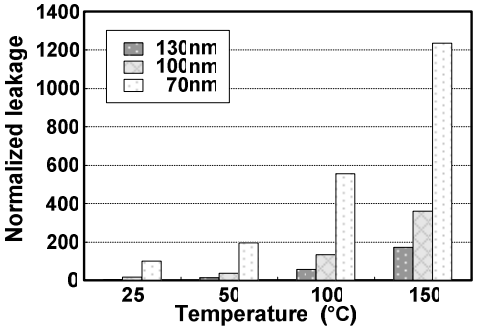

Fig. 3 Leakage power dissipation of an NMOS device for different technology nodes based on SPICE simulations using BSIM3 models showing the impact of temperature. The leakage power dissipation is normalized w.r.t I_{off} at 130 nm node at 25 °C.

Recently, a methodology has been developed that takes these electrothermal couplings into consideration to evaluate an electrothermally coupled EDP [25] (**Fig. 4**). This methodology incorporates both analytical models and results from the circuit simulator based on an integrated device, circuit, and system level modeling approach [5]. In **Fig. 4**, the line (V_{dd} = V_{th}) represents a boundary below which we do not consider operating our circuit, while the region (thermal runaway) is determined by a passive cooling model [5], assuming junction-to-ambient thermal resistance θ_{ja} = 0.85 °C/W.

In comparison with **Fig. 1** that is 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. 1**. The optimal point (marked by 'o') shifts to (V_{dd} = *0.481 V* and V_{th} = *0.279 V*). The iso-leakage curve in **Fig. 4(a)** shows the ratio of leakage power to total power consumption. It essentially provides the limit of supply and threshold voltage scaling when the ratio of active to idle power is constrained. Moreover, as shown in **Fig. 4(b)**, the iso-temperature curve can be simultaneously obtained by applying the electrothermally coupled methodology. It shows the average junction temperature estimation for various designs (different V_{dd}-V_{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. Consequently, if electrothermal couplings are not considered, power dissipation and delay evaluations will be inaccurate and mislead the design optimization process.

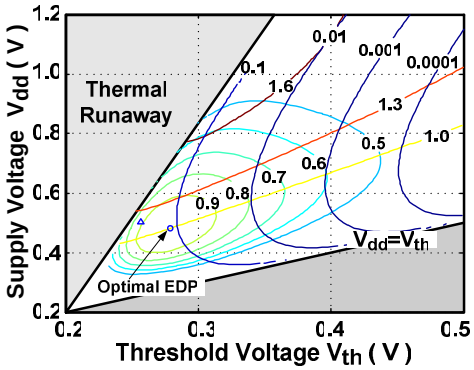

Fig. 4(a) Energy-delay product evaluation by electrothermally coupled analysis. EDP contours along with iso-performance and iso-leakage curves provide a basis for power-performance tradeoffs in circuit design. "Δ" indicates the traditional optimal point for comparison and it is evaluated without considering electrothermal couplings. Note that the design space gets restricted by thermal constraint (thermal runaway) when various electrothermal couplings are taken into account.

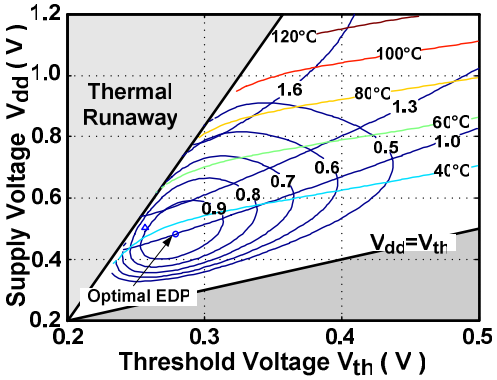

Fig. 4(b) Energy-delay product evaluation by electrothermally coupled analysis. EDP contours along with iso-performance and iso-temperature curves are also shown for power and performance tradeoffs. The iso-temperature curves can be used to provide an additional thermal (or reliability) constraint.

3. Design-Specific Optimization Metrics

In this section, first the logic behind the use of different design metrics is explained through comparison between three general design metrics (EDP, PDP, and PEP). In practice, the optimal point, for example the lowest EDP point, is seldom used due to the need to satisfy other requirements like very high performance or very low power which cannot be captured by that particular evaluation. Hence, we propose a new optimization methodology that allows designers to choose a correct design metric that directly satisfies their design-specific needs. Comparison based on the proposed metric is more meaningful than the use of a single design metric, for example EDP, which does not comprehend design-specific requirements.

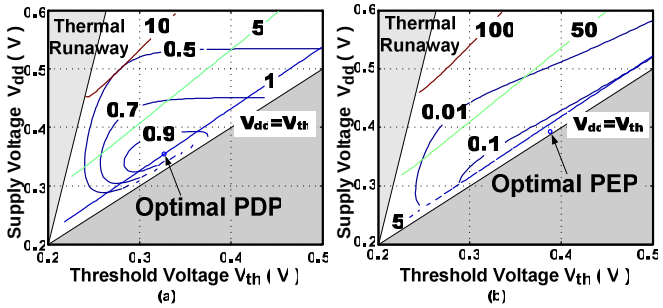

Fig. 5 Design optimization **(a)** using Power-Delay Product (PDP) evaluation **(b)** using Power-Energy Product (PEP) evaluation.

Fig. 5(a) and **Fig. 5(b)** show the PDP and PEP contours respectively. The optimal operating points for three general design metrics (EDP, PDP, and PEP) are shown in **Table 1**.

Table 1 Optimal operating points of different design metrics.

Optimization	Energy-Delay	Power-Delay	Power-Energy
V_{dd} (V)	0.481	0.354	0.393
V_{th} (V)	0.279	0.327	0.388

By definition, EDP prioritizes delay over power because it is proportional to $(\text{delay})^2$. When EDP is the design metric, the optimal operating point will have higher supply voltage and lower threshold voltage, as seen in **Table 1**, in order to have relatively higher performance. However, since PEP prioritizes power over delay, the threshold voltage should increase to reduce the leakage power dissipation.

Fig. 6 compares the result of using these three common optimization metrics on a given design from the perspectives of delay, temperature, and power dissipation. It can be observed from **Fig. 6(a)** that power-energy product (PEP) leads to the highest delay as compared to other metrics. However, the power dissipation of PEP as shown in **Fig. 6(b)** is the lowest. Moreover, PEP will have the highest ratio of $P_{dynamic}$ to P_{static} that gives the highest power efficiency of 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 we use the parameter "μ" that represents the ratio of exponent of delay to that of power. The generalized metric thus is represented as PT^{μ}. For instance, μ of power-energy product (P^2T) is *0.5* and energy-delay product (PT^2) is *2*. When performance is the primary concern, μ is larger than *1*. On the contrary, when the power dissipation is the primary concern, μ is less than *1*.

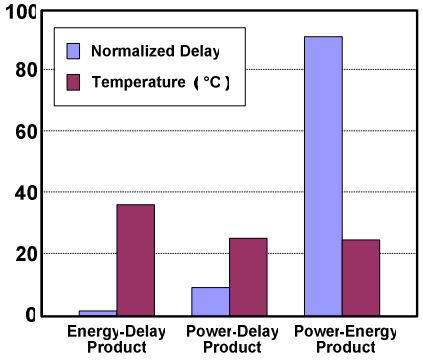

Fig. 6(a) Normalized delay and die temperature corresponding to optimal operating point obtained by three optimization metrics (EDP, PDP, and PEP).

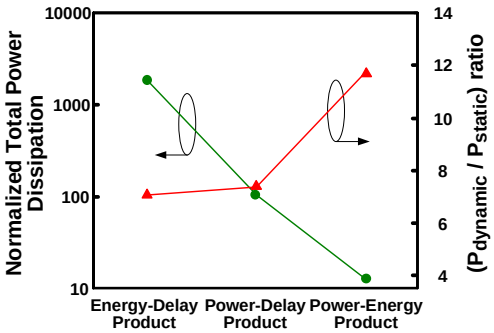

Fig. 6(b) Normalized total power dissipation and ($P_{dynamic}$ / P_{static}) ratio corresponding to optimal operating point obtained by different optimization metrics.

Fig. 7 shows the optimal operating points obtained for different μ. These are the points where PT^{μ} has the lowest value in the design space. An optimization metric with a higher μ will lead to higher performance than lower μ. Traditionally, designers choose EDP ($\mu = 2$) as an optimization metric for trading-off performance and power dissipation. As seen in this figure, EDP provides medium performance and medium power dissipation as compared to other values of μ. When the designer wants to lay higher emphasis on performance, μ can be chosen to be higher than *2*. On the other hand, when the emphasis is on low power, the μ chosen should be less than *2*. It must be noted that for low power applications, the optimal point shifts by a larger amount for a certain change of 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 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. It is important to mention that the optimal operation point is only considered in the region where supply voltage is larger than threshold voltage because of the validity of the Alpha-Power model [23] i.e., we do not consider the sub-threshold operations in this analysis.

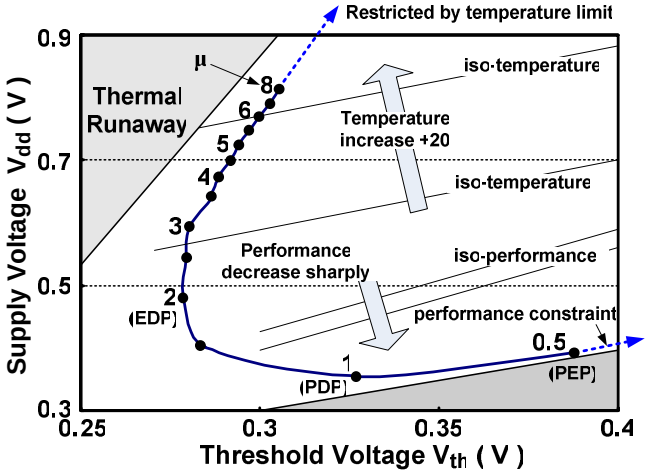

Fig. 7 Optimal operation curve for different μ with the iso-temperature and iso-performance curves superimposed. The shaded regions in the two corners correspond to thermal runaway region and the region where the supply voltage is less than threshold voltage. The optimal operating points are obtained by using different values of μ for 100 nm technology node.

In nanometer technologies, increased power dissipation makes temperature rise a very important concern as described before. **Fig. 7** superimposes the iso-temperature curves, thermal constraint (thermal runaway), operational constraint ($V_{dd} > V_{th}$) and iso-performance curves onto the trend of optimal operating points. Depending on the packaging and cooling technologies available for a particular design, the iso-temperature curve corresponding to any maximum allowable operating temperature provides the upper bound of μ. On the other hand, the iso-performance curve corresponding to minimum allowable performance provides the lower bound of μ. It can be observed in **Fig. 7** that the optimal point for $\mu = 0.5$ (PEP) is very close to the point where the supply voltage is equal to threshold voltage. Since the supply voltage cannot be lower than threshold voltage for normal circuit operation, decreasing the value of μ below *0.5* results in the same optimal operating point as for PEP.

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 example of design optimization.

4. Examples of Design Optimization

i. Unsigned Bit Array Multiplier

In order to demonstrate the utility of the proposed electrothermally aware methodology for design optimization, examples of design optimization are provided in this section. From a microprocessor point of view, the datapath which includes all computational blocks (logic and arithmetic operations) determines the performance and contributes a significant amount to the total power consumption. At circuit level, we consider a multiplier which is an important arithmetic unit of microprocessors. **Fig. 8** shows the schematic diagram of a typical 4X4 unsigned bit array multiplier. The longest ripple path which determines the maximum propagation delay i.e., a signal from Y_0 propagates the entire circuit until it reaches Z_7. An example of this is when all the Y inputs are fixed to high (Y=1111) and when X changes from X=1000 to X=1001. The output Z thus changes from 01111000 to 10000111 respectively. For the sake of simplicity and to isolate the impacts of sizing and input vectors on power and performance analysis, here we only consider that the multiplier is designed using complementary static CMOS with minimum transistor size at 100 nm technology node and operates under the scenario with maximum propagation delay.

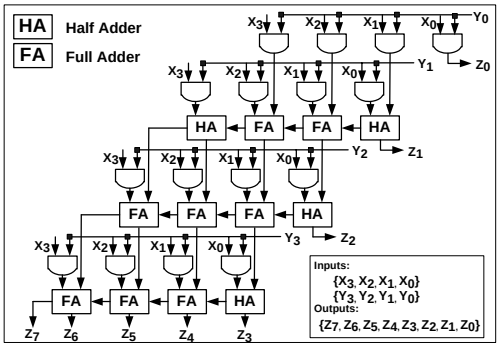

Fig. 8 The schematic diagram illustrates an unsigned bit array multiplier with 4 bits input {X, Y} and 8 bits output {Z}. The typical multiplier uses 16 2-bit AND gates, 4 half-adders, and 8 full-adders implemented by complementary static CMOS at 100 nm technology node.

Fig. 9 shows the iso-temperature and the optimal operation curve for different μ for this case. The performance curves in this figure show the maximum propagation delay of the multiplier. If the design objective is to maximize the performance, a desirable metric would have the highest possible μ under the maximum temperature constraint. For example, if the maximum temperature constraint is 60°C, the highest possible μ should be 4. On the other hand, if the design objective is to achieve the minimum power dissipation without allowing performance to fall below a certain level, a desirable metric would have the lowest possible μ under the minimum performance constraint (maximum propagation delay). For instance, if the maximum allowable propagation delay is 2.6 ns, the parameter μ should be chosen around 1.5 to meet the performance requirement. Hence, the appropriate μ will be given by the intersection of the minimum performance curve and the optimal operation curve.

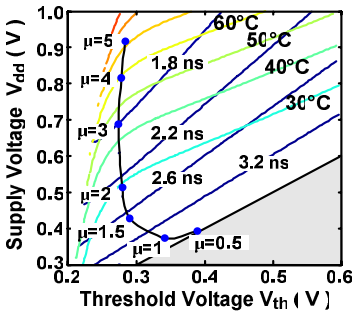

Fig. 9 Optimal operation curve for different μ with the iso-temperature and iso-performance curves superimposed. Iso-performance curve shows the maximum propagation delay i.e., a signal from Y_0 propagates the entire circuit until it reaches Z_7 in **Fig. 8** above.

ii. High-Performance Logic Block

Furthermore, at the system level, we consider a logic block of a high performance integrated circuit at 100 nm technology node. As described in Section 2, a logic block can be represented by an equivalent inverter by using effective transistor width, load capacitance, and activity factor [11] as shown in **Fig. 10**. We consider a uniform (average) temperature over this logic block for simplicity. As this integrated circuit does not employ active cooling like a modern microprocessor, the maximum operating temperature is only *40°C*. The target of the design is to achieve the maximum possible performance under this maximum operating temperature constraint due to packaging and cooling limitations.

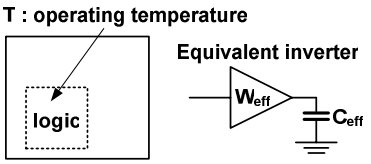

Fig. 10 The schematic diagram illustrates an example of design optimization. A high-performance logic block of an integrated circuit is represented by an equivalent inverter with effective transistor width and load capacitance.

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. 11** that the appropriate μ is at the intersection of the *40°C* iso-temperature curve and the optimal operation curve. For the case shown in **Fig. 11**, the intersection occurs at $\mu = 3.7$. Once the operating temperature value is set to be *40°C* as a constraint, the value of parameter "μ" can be directly obtained by the electrothermally coupled analysis [25] as described in Section 2 (hence this evaluation of the parameter μ does not need any additional computation).

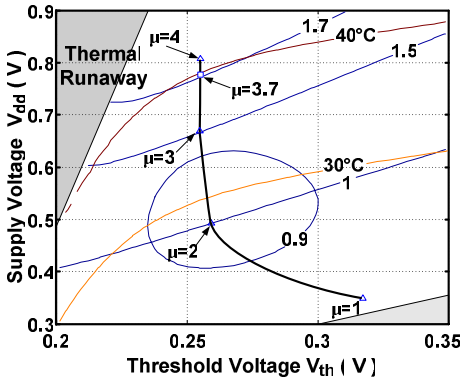

Fig. 11 Illustration of the methodology for finding a suitable optimization metric to meet design-specific requirements. EDP (μ=*2*) contour for EDP=(1/0.9)EDP_{opt}, iso-performance and iso-temperature at 100 nm technology node curves are also shown. "□" indicates the optimal point that meets all design-specific requirements. Note that this figure is evaluated by incorporating an active cooling model.

Given the same constraint as mentioned before, we now consider two possible design choices depicted by points A and B in **Fig. 12** 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 ($PT^{3.7}$) is compared to that based on EDP, which is the most widely used design metric. The optimal PT^2 point (EDP_{opt}) and a corresponding sub-optimal contour of all points where the ratio EDP_{opt} / EDP = *0.9,* is shown. All points outside this contour shown have EDP higher than the points that lie inside this contour. Hence a traditional comparison based on the energy-delay-product would lead to the decision that A is a better choice than B. On the other hand, the optimal point corresponding to the metric $PT^{3.7}$ (which captures the design-specific needs) and the sub-optimal *0.9* contour surrounding this point are also shown. It can be seen that the value of the metric $PT^{3.7}$ 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. However, point B has a delay of *7.82 ns*, whereas point A has a delay of *8.44 ns*. 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 above example, once the parameter "μ" is determined by the proposed methodology, the appropriate metric (PT^{μ}) can capture all design-specific requirements. A procedure similar to EDP evaluation (replacing the quantity PT^2 with PT^{μ}) 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.

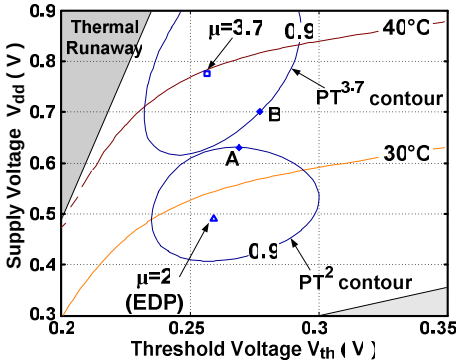

Fig. 12 Example comparing the use of the proposed metric ($PT^{3.7}$) in choosing between the two design options (A and B) to the use of conventional EDP evaluation.

Modern nanometer scale designs often have multi-threshold voltages for improving performance as well as reducing power dissipation. Such designs can be easily handled in the proposed methodology by using multiple equivalent inverters corresponding to the different threshold voltages instead of the single equivalent inverter, which is shown in **Fig. 10**. In the next section, the impact of technology scaling and process variations on the proposed optimization methodology is discussed.

5. 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 will directly impact the junction temperature that determines the limit of μ. It can be observed from **Fig. 13** that the optimal curve shifts when technology scales from *100nm* to *70nm* nodes. Given the same criteria for two circuits, the design employing an advanced technology (*70nm* technology node) will have higher optimal values for threshold voltages due to the increase of leakage power dissipation (**Fig. 3**). Moreover, due to technology scaling and the resultant increasing leakage, it can be clearly seen that the design space gets increasingly restricted by thermal constraint.

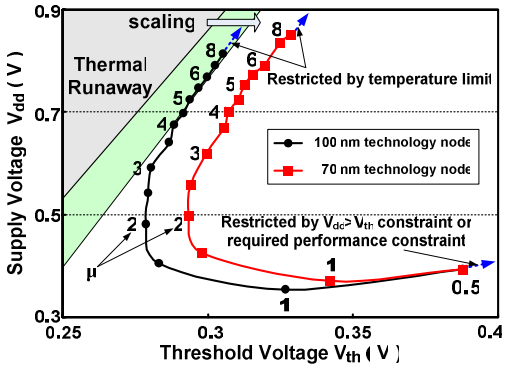

Fig. 13 Scaling analysis of optimal operating points by applying different optimization metrics (shown for 100 nm and 70 nm technology nodes). Note that the region (thermal runaway) expands due to technology scaling.

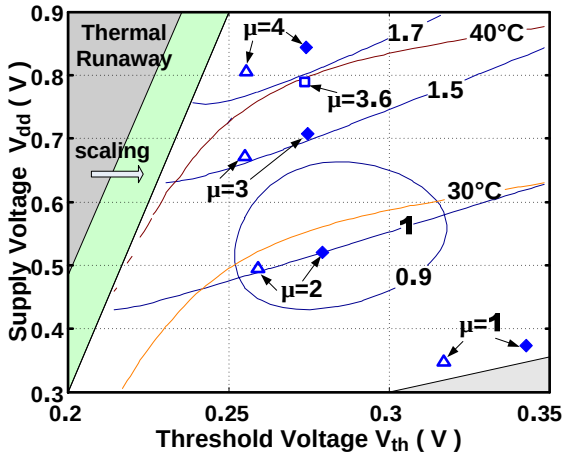

Fig. 14 Effect of technology scaling from *100nm* to *70nm* on operating point selection methodology based on EDP evaluation versus the proposed methodology. EDP (μ = *2*) contour for EDP=(1/0.9)EDP_{opt}, iso-performance and iso-temperature curves for *70nm* technology node evaluation are also shown. Optimal operating points based on different optimization metrics for *100nm* technology node are indicated by "Δ". "♦" indicates the corresponding optimal points for *70nm* technology node. "□" indicates the optimal point that meets all design-specific requirements at *70nm* technology node. Note that this figure is evaluated under the same conditions as those in **Fig. 11**.

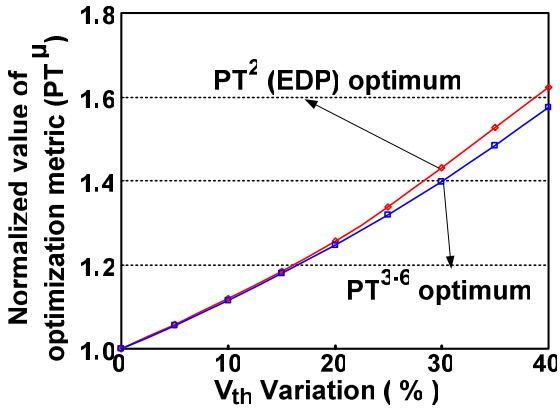

Fig. 15 Impact of threshold voltage variation on the optimal value of different optimization metrics (PT^2 versus $PT^{3.6}$) for *70nm* technology node (with active cooling). The values shown are normalized to the corresponding optimal values without threshold voltage variations.

Fig. 14 shows the impact of technology scaling on selecting μ for design-specific optimization. Under the same constraints as used in the example in Section 4, it is observed that if the same optimization metric ($PT^{3.7}$) is chosen for *70nm* technology node, the optimal operating point exceeds the maximum allowed temperature. For the *70nm* technology node, the correct optimization metric that meets the design specific requirements is found to be $PT^{3.6}$. Thus, the design optimization metric needs to be sensitive to technology scaling.

In nanometer technologies, parameter variations are known to have increasing impact on all aspects of circuit design [18]. For the same example discussed in the previous section, **Fig. 15** shows the impact of threshold voltage variations on the optimal values of the optimization metrics obtained by using the proposed methodology and by conventional EDP evaluation. Note that this evaluation is done at *70nm* technology node where μ is found to be *3.6* (refer to **Fig. 14**). It can be observed that for the specific requirements of this design, the optimal point of the proposed metric shifts by a smaller amount than the optimal point of EDP and this difference between the two increases as variations become larger. Hence the proposed metric is less sensitive to threshold voltage variation than EDP-based optimization in this case.

6. CONCLUSION

In this work, a systematic methodology for choosing design-specific optimization metrics for simultaneous optimization of power and performance has been proposed. The methodology incorporates electrothermal couplings between temperature, power dissipation, and performance. The design metric evaluated using this methodology provides a more meaningful basis to optimize supply and threshold voltages under design-specific constraints as compared to traditional methodologies that do not comprehend design specifics and electrothermal effects. Using the proposed methodology, an appropriate optimization metric that is sensitive to CMOS scaling and parameter variations can be obtained.

ACKNOWLEDGEMENTS

This work was supported by a grant from Intel Corporation and the University of California-MICRO Program.

REFERENCES

[1] *International Technology Roadmap for Semiconductors* (ITRS), 2004 edition, http://public.itrs.net/
[2] V. De and S. Borkar, "Technology and Design Challenges for Low Power and High Performance," in *Proc. ISLPED*, 1999, pp. 163-168.
[3] P. P. Gelsinger, "Microprocessors for the New Millennium: Challenges, Opportunities, and New Frontiers," in *Proc. ISSCC*, 2001, pp. 22-25.
[4] P. Gelsinger, *41st DAC Keynote, DAC*, 2004. (www.dac.com)
[5] K. Banerjee et al., "A Self-Consistent Junction Temperature Estimation Methodology for Nanometer Scale ICs with Implications for Performance and Thermal Management," in *IEDM Tech. Dig.*, 2003, pp. 887-890.
[6] C-K. Hu et al., "Scaling Effect on Electromigration in On-Chip Cu Wiring," in *Proc. IITC*, 1999, pp. 267-269.
[7] R. Blish et al., "Critical Reliability Challenges for The Internatinonal Technology Roadmap for Semiconductors," *International Sematech Technology Transfer Document 03024377A-TR*, 2003.
[8] A. M. Yassine et al., "Time Dependent Breakdown of Ultra-Thin Gate Oxide," *IEEE Trans. Electron Devices*, Vol. 47, pp. 1416–1420, 2000.
[9] S-C. Lin et al., "Impact of Off-state Leakage Current on Electromigration Design Rules for Nanometer Scale CMOS Technologies," in *Proc. IRPS*, 2004, pp. 74-78.
[10] M. Horowitz et al., "Low Power Digital Design," in *Proc. ISLPED*, 1994, pp. 8-11.
[11] R. Gonzalez, et al., "Supply and Threshold Voltage Scaling for Low Power CMOS," *IEEE J. Solid-State Circuits*, Vol. 32, pp. 1210–1216, 1997.
[12] K. Nose, and T. Sakurai, "Optimization of Vdd and Vth for Low Power and High Speed Applications," in *Proc. ASP-DAC*, 2000, pp. 469-474.
[13] P. I. Pénzes and A. J. Martin, "Energy-Delay Efficiency of VLSI Computations," *in Proc. GLSVLSI*, 2002, pp. 104–111.
[14] H. P. Hofstee, "Power-Constrained Microprocessor Design," *in Proc. ICCD*, 2002, pp. 14–16.
[15] V. Zyuban and P. N. Strenski, "Balancing Hardware Intensity in Microprocessor Pipelines," *IBM J. RES. & DEV.*, Vol. 47, pp. 585-598, 2003.
[16] V. Zyuban and P. N. Strenski, "Unified Methodology for Resolving Power-Performance Tradeoffs at the Microarchitectural and Circuit Levels," *in Proc. ISLPED*, 2002, pp. 166–171.
[17] D. Markovic′ et al., "Methods for True Energy-Performance Optimization," *IEEE J. Solid-State Circuits*, Vol. 39, pp. 1282–1293, 2004.
[18] S. Borkar et al., "Parameter Variations and Impact on Circuits and Microarchitecture," in *Proc. DAC*, 2003, pp. 338-342.
[19] Y-S. Lin et al., "Leakage Scaling in Deep Submicron CMOS for SoC," *IEEE Trans. Electron Devices*, Vol. 49, pp. 1034-1041, 2002.
[20] H. Soeleman et al., "Robust Subthreshold Logic for Ultra-Low Power Operation," *IEEE Trans. VLSI Systems*, Vol. 9, pp. 90-99, 2001.
[21] A. Wang et al., "Optimal Supply and Threshold Scaling for Subthreshold CMOS Circuits," *in Proc. ISVLSI*, 2002, pp. 5-9.
[22] D. Sengupta and R. Saleh, "Power-Delay Metrics Revisited for 90 nm CMOS Technology," *in Proc. ISQED*, 2005, pp. 291-296.
[23] 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, pp. 584–593, 1990.
[24] www.intel.com
[25] A. Basu et al., "Simultaneous Optimization of Supply and Threshold Voltages for Low-Power and High-Performance Circuits in the Leakage Dominant Era," in *Proc. DAC*, 2004, pp. 884-887.