Simultaneous Optimization of Supply and Threshold Voltages for Low-Power and High-Performance Circuits in the Leakage Dominant Era

Anirban Basu, Sheng-Chih Lin, Vineet Wason, Amit Mehrotra†, and Kaustav Banerjee
Dept. of Electrical and Computer Engineering, University of California, Santa Barbara, CA 93117
†Berkeley Design Automation Inc., Santa Clara, CA 95054
{anirban, sclin, vwason, kaustav}@ece.ucsb.edu, amit.mehrotra@berkeley-da.com

ABSTRACT

Electrothermal couplings between supply voltage, operating frequency, power dissipation and die temperature have been shown to significantly impact the energy-delay-product (EDP) based simultaneous optimization of supply (V_{dd}) and threshold (V_{th}) voltages. We present for the first time, the implications of an electrothermally aware EDP optimization on circuit operation in leakage dominant nanometer scale CMOS technologies. It is demonstrated that electrothermal EDP (EEDP) optimization restricts the operation of the circuit to a certain region in the V_{dd}-V_{th} plane. Also, the significance of EEDP optimization has been shown to increase with increase in leakage power and/or process variations.

Categories and Subject Descriptors

B.7.1 [**Hardware**]: Integrated circuits – *VLSI.*

General Terms

Performance, Design.

Keywords

Electrothermal couplings, energy delay product, subthreshold leakage, temperature aware design.

1. INTRODUCTION

Simultaneous Optimization of V_{dd} and V_{th}:

Low-power consumption in high performance circuits is highly desirable as it directly relates to battery life, reliability, packaging, and heat removal costs [1]. Scaling of V_{dd} reduces dynamic power consumption but degrades the performance of the circuit as well. This can be partially compensated by lowering V_{th} but at the cost of increased leakage power. Thus, the need for low power and high performance circuit design motivates the finding of an optimal set of V_{dd} and V_{th} that ensures the required performance of the circuit with lowest power consumption [2]-[4]. For these kind of applications, where both performance and amount of computations that can be done for a given energy budget are of importance, energy-delay product (EDP) is an appropriate metric to optimize and compare different designs [5], [6].

***Electrothermal EDP Optimization*:**

It has been recently reported that in the domain of increasing subthreshold leakage; the supply voltage, operating frequency, power dissipation and die temperature of a chip are electrothermally coupled to each other, rather than being related by simple independent analytical equations [7]. The various electrothermally coupled equations can be solved *self-consistently* and can be employed to study power-performance-reliability-cooling cost trade offs, which can in turn, be used to improve the performance of nanometer scale ICs [7, 8]. In this paper we develop an EDP based V_{dd}-V_{th} optimization technique that takes these electrothermal couplings into account by solving them iteratively in a self-consistent manner. By applying this electrothermal-energy-delay-product (EEDP) optimization, we provide more accurate guidelines for power-performance tradeoffs with the help of energy-delay contours, and iso-performance and iso-leakage power curves. It is shown that simple numerical optimization [5, 6] of the EDP does not generate true optimal values of V_{dd} and V_{th}. In fact, such optimization techniques become increasingly ineffective in subthreshold leakage-dominant technologies. Moreover, it is illustrated that the electrothermal couplings forbid the operation of circuits in certain regions of the V_{dd}-V_{th} plane. Most significantly, the importance of the EEDP optimization method is shown to increase with increase in subthreshold leakage and process variations. The EEDP technique can be employed to study various electrical-thermal tradeoffs as well as circuit and device level optimization in deeply scaled CMOS circuits.

2. ENERGY-DELAY PRODUCT AS AN OPTIMIZATION FUNCTION

The two main sources of power dissipation in CMOS circuits are leakage power, which is mainly due to subthreshold leakage, and dynamic power, which results from switching capacitive loads between different voltage levels. The short-circuit component is usually small; therefore we ignore it throughout this paper. By writing the total power consumption of an average gate in the circuit (equivalent to considering a homogenous circuit model) as the sum of the switching power and subthreshold leakage power, and delay according to the α-power law model, the energy-delay product can be expressed as [5, 9]:

$$EDP = \frac{K^2 \cdot I_s \cdot L_d \cdot V_{dd}^3}{(V_{dd} - V_{th})^{\alpha}} \left[\frac{a \cdot C_{eff}}{I_s \cdot K \cdot L_d} + \frac{e^{-V_{th}/\gamma V_0}(1 - e^{-V_{ds}/\gamma V_0})}{(V_{dd} - V_{th})^{\alpha}} \right] \quad (1)$$

where K is a proportionality constant specific to a given technology, I_s is the zero-threshold leakage current, L_d is the logic depth of the microprocessor, γ is the body effect coefficient, and V_0 denotes the subthreshold slope. Index α accounts for velocity saturation condition of the transistors (α=1 when transistors are under complete velocity saturation and α=2 when no velocity saturation).

DAC 2004, June 7–11, 2004, San Diego, California, USA.

C_{eff} is the average capacitance and aC_{eff} is the average capacitance switched every cycle per micron of transistor width. The gate delay (T_g) of the chip can be modeled as that of an inverter using the alpha-power model [9]:

$$T_g = \frac{K \cdot V_{dd}}{(V_{dd} - V_{th})^{\alpha}} \tag{2}$$

The maximum operating frequency of the chip is given by

$$f = \frac{1}{T_g L_d} \tag{3}$$

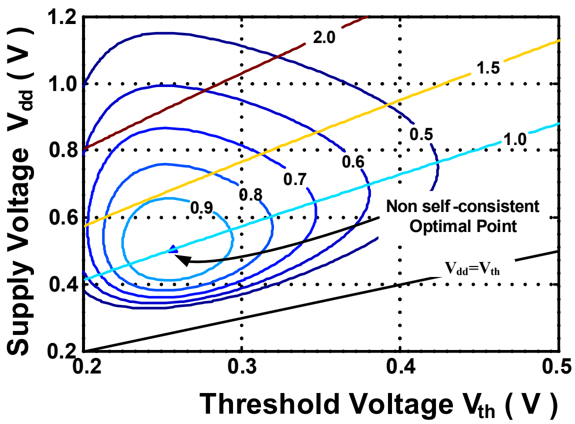

Fig. 1 EDP contours and performance curves computed non-self-consistently for transistors with $\alpha = 1.3$.

Fig. 1 shows contours of the inverse of the relative EDP. The relative EDP 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, *i.e.* minimum value. The diagonal lines in Fig. 1 are curves for constant performance. The numbers on the performance curves indicate the normalized value of the frequency where normalization is done with respect to the frequency of operation at the optimal point. The $V_{dd} = V_{th}$ line represents a boundary below which we do not consider operating our circuit.

The optimal point and the curves in Fig. 1 are called *non-self-consistent* as they are obtained by direct numerical solution of equation (1) without considering electrothermal couplings among junction temperature, frequency and power. Therefore, the solution set of V_{dd} and V_{th} is not truly optimal. In the next section, we present a fully coupled EDP evaluation method that incorporates electrothermal inter-dependencies while solving for the true optimal values of V_{dd} and V_{th} that yields minimum EDP for the circuit. We term these solutions as *self-consistent* solutions.

3. SELF-CONSISTENT METHODOLOGY FOR SIMULTANEOUS OPTIMIZATION OF V_{dd} AND V_{th}

With scaling of CMOS technology beyond 100 nm, circuit performance and junction temperature are strongly affected by subthreshold dominated leakage power ($P_{leakage}$), which constitutes a significant part of total chip power (P_{chip}) [3],[7]. However, subthreshold leakage ($P_{leakage}$) is exponentially dependent on junction temperature (T_j) and the dependence becomes stronger with scaling. Also, T_j increases nonlinearly with junction-to-ambient thermal impedance (θ_j) due to coupling between P_{chip} and T_j, arising primarily due to the strong dependence of $P_{leakage}$ on T_j [7]. Furthermore, the total power dissipation and T_j increase as the chip frequency increases with an increase in V_{dd}. Also, frequency itself is dependent on temperature due to the dependence of the transistor on-current (I_{on}) on T_j. Moreover, T_j has two counteracting effects on I_{on}: a) increase in I_{on} due to lowering in V_{th} at increased T_j, and b) decrease in I_{on} due to reduction in mobility at higher T_j [6]. The details of the various couplings are summarized in **Fig. 2** in functional forms that represent our electrothermal model. It can be observed that supply voltage, power, frequency and temperature are all intricately coupled. Hence, a self-consistent electrothermal analysis method is imperative for accurate estimation of T_j for any value of V_{dd} (or frequency) so that energy-delay can be evaluated correctly. Thus, any optimization in V_{dd} and V_{th} should also incorporate this notion of self-consistent evaluation in its methodology, and thus, a straightforward optimization of equation (1), which does not take into account the effect of various couplings indicated in Fig. 2, is not accurate.

$$frequency = f_1(V_{dd}, V_{th}, L_d, \mu)$$
$$P^{total}_{switching} = \alpha C^{total}_{eff} \cdot V^2_{dd} \cdot frequency$$
$$P^{total}_{leakage} = f_2(V_{dd}, V_{th}, T_j)$$
$$T_j = T_{amb} + \theta_j \cdot (P^{total}_{switching} + P^{total}_{leakage})$$
$$\Delta V_{th} = f_3(T_j)$$

Fig. 2 Models for various metrics are expressed in functional format. Electrothermal couplings are indicated using broken lines.

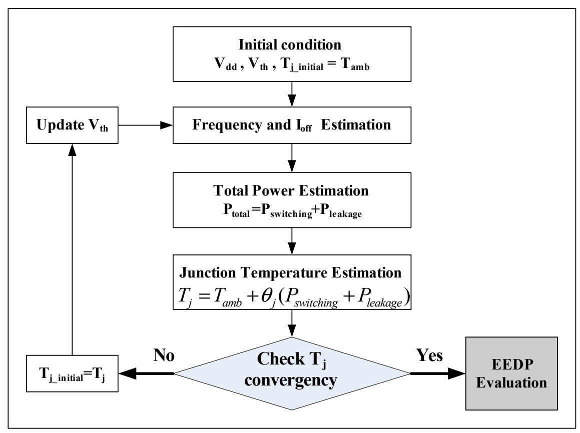

Fig. 3 An overview of the self-consistent optimal EDP estimation methodology.

We propose our EEDP methodology that is based on an integrated device, circuit, and system level modeling approach and has been summarized in **Fig. 3**. For a given V_{dd}, V_{th} and initial T_j (we use T_{amb} as a initial value), the operating frequency and the total leakage current of the chip are first estimated. The estimated frequency is then used in the calculation of the switching (active) power. Also, the leakage power can be estimated using I_{off}. For our analysis, nominal value of I_{off} was calibrated against measured data at ambient temperature. The total chip power (equation (4)) is then used to calculate the new junction temperature using compact thermal models for the IC packaging and cooling technology.

$$P = aC_{eff}V^2_{dd}fW_{eff} + I_s e^{-V_{th}/\gamma V_0}(1 - e^{-V_{ds}/\gamma V_0})V_{dd}L_dT_gfW_{eff} \tag{4}$$

where W_{eff} is the effective width contributing to power dissipation and frequency is given by (3).

In order to accurately estimate the junction temperature, we use the thermal model from Fig. 2 in each iteration, where, the estimated junction temperature is then compared with the initial value of T_j to check for convergence. The process continues till a convergence in the value of T_j is achieved. The new junction temperature is used to calculate the new threshold voltage (equation (5)).

$$V_{th} = V_{th0} - k(T_j - T_{amb}) \tag{5}$$

where k is threshold voltage temperature coefficient whose typical value for 130 nm is 0.7 mV/K [10].

Following this methodology, and by choosing different sets of V_{dd} and V_{th} as starting points we can obtain energy delay product value (in other words, EEDP) for each point in the sample plane. Therefore, using this approach, each value of EEDP calculated in the V_{dd} - V_{th} plane is evaluated using self-consistent T_j. Hence, the point corresponding to minimum EEDP value is the *true* optimal point and the corresponding V_{dd} and V_{th} are *true* optimal voltages. Frequency of the circuit at these voltages is the optimum frequency that yields minimum energy-delay product. These optimum values can now be used to normalize EDP and performance values at other points in the plane and obtain constant EDP and performance curves of the chip under self-consistent condition, as we have shown and discussed in the next section. Although, we illustrate our results for a 130 nm technology based 32-bit microprocessor chip, the methodology is not specific to the technology node, and also can be applied to any chip without any loss of generality, provided the chip data are available.

4. IMPLICATIONS ON CIRCUIT OPERATION AND DESIGN RULES

Following the EEDP methodology as explained in the previous section, self-consistent curves for energy-delay and performance are obtained (**Fig. 4**). Additionally, contours are shown for different ratios of the leakage power to the total power dissipation; the ratio being varied as multiples of 10. It can be observed that the self-consistent optimal point (marked by 'o' at V_{dd} = 0.481 V and V_{th} = 0.279 V) is different from the non-self-consistent one (marked by 'Δ'). By comparing Fig. 1 and Fig. 4, we can see that self-consistent calculation of electrothermally coupled quantities results in an overall shift of the energy-delay and performance contours. Therefore, it is imperative to compare the implications of these shifts on circuit operation. For instance, operating the circuit at V_{dd}= 0.6 V and V_{th} = 0.3 V results in only about 10% worse EDP considering electrothermal couplings (Fig. 4) as opposed to about 20 % as obtained from Fig. 1 [*Note*: EDP values in Fig. 1 and Fig. 4 are compared with their respective optimal EDP; hence, the two figures should be read independent of each other]. Furthermore, if high performance is desired one may boost the supply to 0.8 V at V_{th}= 0.3 V, which means, from Fig. 4, $(0.6)^{-1}$ i.e. 66% worse EDP but more than 60% higher performance than at optimal point. However, same set of voltages indicates about 43% worse EDP with slightly higher than 50 % performance without consideration of electrothermal couplings. Moreover, it is possible to operate at the same EDP and performance, but at different leakage percentage. For instance, point A and B in Fig. 4 have same EDP and performance, but B has a lower leakage power percentage. For leakage dominant applications, it is beneficial to operate at point B as compared to point A. Thus, the EEDP contours along with performance and iso-leakage curves provide an accurate basis for power-performance tradeoffs in circuit design.

As can be seen from Fig. 4, the non-self-consistent point is located on the 1.2x self-consistent performance (frequency) curve. This means that the performance of the circuit at the non-self-consistent optimal point is 20% higher than the optimum frequency. However, as this point lies near the 0.9 curve in Fig.4, the overall energy-delay product is actually $(0.9)^{-1}$ *i.e.* 11.1 % worse than the true optimal operation point for this particular circuit. Therefore, unless the electrothermal couplings are taken into consideration as we proposed in our EEDP optimization methodology, a true minimum energy-delay product is not obtained.

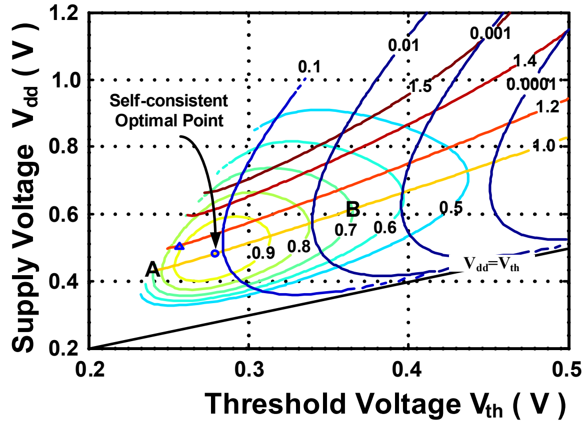

Fig. 4 Inverse of relative EDP contours and performance curves drawn from self-consistent electrothermal considerations. Δ indicates the non-self-consistent optimal point for comparison.

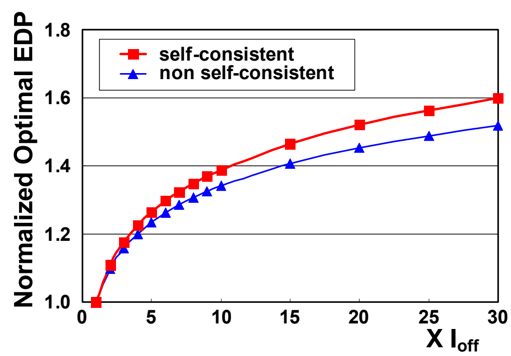

Fig. 5 Normalized optimal EDP for self-consistent and non-self-consistent methodologies as a function of increasing leakage current.

Furthermore, it is important to notice (Fig. 4) that EEDP optimization methodology limits the operation of the circuit in certain region (high V_{dd}, low V_{th}) of the V_{dd} - V_{th} plane besides the $V_{dd} = V_{th}$ boundary line. The region where the junction temperatures become excessively high is forbidden by the self-consistent methodology. Such regions are not restricted in Fig. 1, because simple numerical solution of EDP equation does neither consider electrothermally coupled quantities nor evaluate junction temperature self-consistently. Also from **Fig. 5**, it can be observed that optimum EDP evaluated by non-self-consistent methodology becomes increasingly misleading as the technology gets increasingly leaky.

The trend for optimum V_{dd} and V_{th} with technology scaling is shown in **Table I**. From the table, it can be observed that as velocity saturation index (α) becomes closer to 1, the optimum V_{dd} and V_{th} scale down. This is because of the increase in leakage power with technology scaling. Thus, in order to compensate for the increasing effect of leakage power, the switching power should be reduced by lowering V_{dd} which yields an optimum EDP. The self-consistent and non-self-consistent optimum V_{dd} and V_{th} for different values of activity factor are tabulated in **Table II**. When the activity factor increases the power consumption of the circuit increases too. To compensate for the increase in power consumption and to meet required performance lower values of V_{dd} and V_{th} are needed.

Table-I

Activity Factor (a)	Logic Depth (L_d)	Velocity Saturation Index (α)	Non-Self-Consistent Optimal Point		Self-Consistent Optimal Point	
			V_{dd} (V)	V_{th} (V)	V_{dd} (V)	V_{th} (V)
0.15	23	1.1	0.43719	0.25152	0.42613	0.27445
		1.3	0.50352	0.25657	0.48141	0.27872
		1.5	0.58090	0.25657	0.53668	0.28286

Table-II

Activity Factor (a)	Logic Depth (L_d)	Velocity Saturation Index (α)	Non-Self-Consistent Optimal Point		Self-Consistent Optimal Point	
			V_{dd} (V)	V_{th} (V)	V_{dd} (V)	V_{th} (V)
0.10	23	1.3	0.52010	0.26667	0.49799	0.28055
0.15			0.50352	0.25657	0.48141	0.27872
0.20			0.49246	0.24646	0.45930	0.27232

5. IMPACT OF PROCESS VARIATIONS

Parameter variations, especially *within-chip variations* pose a major challenge in the design optimization of high performance VLSI circuits, especially for sub-100 nm technologies [11]. These within-chip variations that arise either from environmental variations (temperature (T) and supply voltage (V)) or from physical variations (channel length (L), oxide thickness (T_{OX}) etc.) can result in an uncertainty in the power and frequency values, thus causing a spread in the distribution of *EEDP*. As a result, the *EEDP* based simultaneous optimization problem of supply and threshold voltage presented in this paper above needs to be solved probabilistically. To carry out this probabilistic analysis, the optimization problem can be modeled by taking Gaussian distributions for threshold voltage (V_{th}), supply voltage (V) and temperature (T). These variations result in an increase of subthreshold leakage power, thus increasing the optimum EDP as already observed in Fig.5.

Here, we consider only the effect of V_{th} variations to analyze the increasing significance of applying self-consistent methodology (under parameter variations) for EDP based V_{dd} - V_{th} optimization. Since, total chip power (P_{chip}), junction temperature (T_j) and frequency (f) follow statistical distributions, their mean values were used to carry out the analysis. EDP under variations is then calculated by using mean plus one standard deviation for both energy and delay. **Fig. 6** plots the normalized optimum EDP as a function of percentage V_{th} variations for both self-consistent and non-self-consistent methodologies. It can be observed that self-consistent methodology results in a greater increase of optimum EDP since it takes various electrothermal couplings among power, junction temperature and operating frequency into account. Furthermore, it can be clearly observed from Fig. 6 that as percentage V_{th} variations increase, it becomes increasingly important to apply self-consistent methodology for EDP based V_{dd}- V_{th} based optimization. For instance at 40% V_{th} variations, non-self-consistent methodology predicts the increase in EDP by only 1.11X which is extremely misleading as compared to an increase of 1.84X predicted by the self-consistent methodology. Moreover, the significance of applying the self-consistent methodology is expected to increase when other parameter variations such as supply voltage and temperature variations are also taken into account.

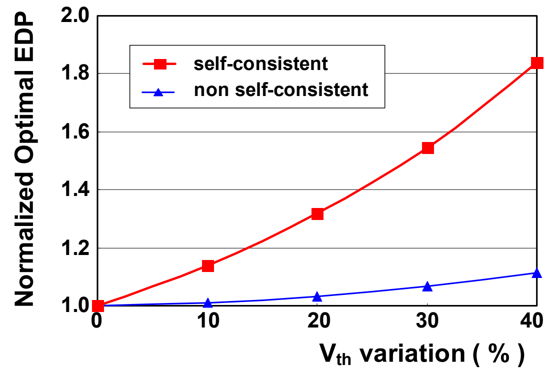

Fig. 6 Normalized optimal EDP as function of percentage V_{th} variation drawn for both self-consistent and non-self-consistent methodologies. Normalized EDP increases with increasing percentage V_{th} variations and the increase is much more for self-consistent methodology. The optimal EDP is normalized to the respective optimum EDP values for 0% variation in V_{th}.

6. CONCLUSION

An electrothermal energy delay product (EEDP) based optimization methodology has been developed for nanometer scale circuits. The optimal circuit operation condition thus obtained is shown to be different from that obtained by optimization of the uncoupled energy delay product. Moreover, it has been shown that EEDP methodology restricts the circuit operation to a certain zone in the V_{dd} - V_{th} plane. Additionally, revised power-performance based tradeoffs and design guidelines have been proposed for leakage dominant technologies. Furthermore, the importance of the EEDP optimization method is shown to increase with increase in subthreshold leakage and process variations.

ACKNOWLEDGMENT

This work was supported by Intel Corp., Fujitsu Labs. of America, and the University of California-MICRO program.

REFERENCES

[1] A.P. Chandrakasan and R.W. Brodersen, "Minimizing power consumption in digital CMOS circuits," *Proc. IEEE*, vol. 83, no. 4, Apr. 1995, pp. 498-523.

[2] D. Liu, et al., "Trading speed for low power by choice of supply and threshold voltage," *IEEE J. Solid-State Circuits*, vol. 28, pp. 10-17, Jan. 1993.

[3] V. De and S. Borkar, "Technology and design challenges for low power and high performance microprocessors," *Proc. ISLPED*, 1999, pp. 163 –168.

[4] P. Pant, V.K. 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 Integration (VLSI) Systems*, vol. 6, no. 4, pp. 538-545, Dec. 1998.

[5] R. Gonzalez, et al., "Supply and threshold voltage scaling for low power CMOS," *IEEE J. Solid-State Circuits*, vol. 32, pp. 1210–1216, Aug. 1997.

[6] K. Nose, T. Sakurai, "Optimization of Vdd and Vth for low power and high speed applications," *Proc. ASP-Design Automation Conf.*, 2000, pp. 469-474.

[7] K. Banerjee et al., "A self-consistent junction temperature estimation methodology for nanometer Scale ICs with implications for performance and thermal management," *IEEE International Electron Devices Meeting*, 2003, pp. 883-887

[8] S-C. Lin et. al., "Impact of off-state leakage current on electromigration design rules for nanometer scale CMOS technologies", *Proc. IRPS 2004*.

[9] 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, Apr. 1990.

[10] Y. Taur and T. Ning, *Fundamentals of Modern VLSI Devices*, Cambridge Univ. Press, 1998.

[11] S. Borkar et al., "Parameter variations and impact on circuits and microarchitecture," in *Proceedings 2003 Design Automation Conference,* 2003, pp. 338-342.