

Minimizing Total Power by Simultaneous Vdd/Vth Assignment

Ashish Srivastava

Dennis Sylvester

University of Michigan, EECS Department, Ann Arbor, MI 48109, {ansrivastava,dennis}@eecs.umich.edu

Abstract - We investigate the effectiveness of *simultaneous* multi-supply and threshold voltage assignment in minimizing the total power (static + dynamic) for the first time. Achievable power reductions under varying conditions are investigated, including static-power limited designs and sub-1V processes. Rules of thumb are developed for optimal Vdd's and Vth's to be used in future designs. These models show the optimal second Vdd to be approximately half the nominal Vdd while the total power savings is significantly greater than previously anticipated. We describe the impact of level conversion delays and highlight the tradeoff between power savings and critical path count.

I. INTRODUCTION

Power consumption is a top priority in high-performance circuit design today. From a dynamic power perspective, supply voltage reduction is the most effective technique to limiting power. However, the delay increase with reducing Vdd degrades the throughput of the circuit. Similarly, to reduce static power an increase in Vth provides exponential improvements, again at the expense of speed.

To counter the loss in performance, dual Vdd [1] and dual Vth [2] techniques have been proposed. These approaches assign gates on critical paths to operate at a higher Vdd or lower Vth and non-critical portions of the circuit are made to operate at lower Vdd or higher Vth respectively, reducing the total power consumption without degrading performance. These techniques have been successfully implemented but most of the existing work focuses on one of these techniques in isolation as opposed to jointly. Also, as the contribution of static power to the total power grows, a need to minimize the total power as opposed to either dynamic or static power alone becomes evident. For example, leakage power makes up approximately 15% of total power consumption for functional units of the Pentium 4 [3]. Power reduction techniques must target both static and dynamic components to be most effective. In [4] the authors show that intelligently reducing Vth in multi Vdd systems offsets the traditional delay penalties at low-Vdd with lessened static power consequences (due to reduced Vdd and Ioff levels). Taking this approach, *total* power minimization becomes the only practical goal since dynamic power can be continually reduced through lowered Vth values. The rise in static power under this circumstance will eventually outweigh the smaller dynamic power.

Additionally, in dual or multi-Vdd designs the effect of drain-induced barrier-lowering (DIBL) causes the effective Vth of lower Vdd gates to increase. This results in larger delay penalties, reducing the number of gates that can be set to low Vdd and limiting the achievable improvement in dynamic power. This further points to the use of several thresholds in conjunction with a multi-Vdd design. We refer to the use of dual Vdd and Vth values together in the same design as dual Vdd/Vth in this paper (and multi Vdd/Vth for cases with more than two supply or threshold voltages as in Section VI) while the term dual Vdd implies a single Vth is used for all gates.

Previous work [5] estimates the optimal Vdd and Vth values to be used in multi-voltage systems to maximize either dynamic or static power savings respectively. The paper does not acknowledge the advantages of combining multi-Vdd and multi-Vth to reduce the total power of the design. They confirm earlier work [6] claiming that, in a dual Vdd system, the optimal lower Vdd is 0.6-0.7 times the original Vdd. In general, [5,6] found optimized multi Vdd systems to provide power reductions of approximately 40%. The application of multiple supply and threshold voltages and gate oxide thickness for SOCs was explored in [7]. The authors assume the same transistor parameters within an entire circuit block and optimize an energy-delay based metric (ED²). In our approach, path delays are tailored to exploit slack available on paths (given a timing constraint). In this way, optimal Vdd and Vth values are found.

In this paper we make several contributions: 1) We minimize total power consumption, defined as the sum of static and dynamic components, 2) we simultaneously optimize Vdd and Vth to achieve this goal, and 3) we consider DIBL, which limits the achievable power reduction in a multi Vdd, single Vth design environment. We also develop rules of thumb to estimate the optimal Vdd and Vth to be used in future designs. We extend our analysis to explore level conversion delay penalties and also introduce the key tradeoff between power reduction and critical path proliferation.

II. POWER OPTIMIZATION FRAMEWORK

To estimate the power improvement obtained by applying multiple Vdd's and Vth's we perform a path-based analysis of the logic network. To simplify the problem we assume non-crossing parallel paths. It is also assumed that it is possible to apply a combination of Vdd and Vth to any fraction of the total path capacitance. This is equivalent to stating that extended clustered voltage scaling (ECVS) is used, allowing for level conversion anywhere along a path [6]. While we do not explicitly consider overhead due to level conversion in most of this work, we describe later the impact of level conversion delay penalties. If C_{1,1} is the total path capacitance, then the total dynamic power dissipation in a n-Vdd/m-Vth logic network path can be expressed as

$$P = f \left[\left[C_{1,1} - \sum_{i=2}^n \left(\sum_j C_{i,j} \right) \right] V_1^2 + \sum_{i=2}^n \left(\sum_j C_{i,j} \right) V_i^2 \right] \quad (1)$$

where C_{i,j} is the capacitance operating at the voltage V_i and threshold voltage V_{THj}. Hence the gain in dynamic power can be expressed as

$$Gain_{D_{dyn}} = 1 - \frac{1}{C_{1,1}} \sum_{i=1}^n \left[\sum_j C_{i,j} \left[1 - \left(\frac{V_i}{V_1} \right)^2 \right] \right] \quad (2)$$

The static power can be expressed similarly. If W_{1,1} is the total width of PMOS and NMOS and W_{i,j} is the width of PMOS and NMOS at power supply V_i and threshold voltage V_{THj} then the gain in static power is given by

$$Gain_{static} = 1 - \sum_{i=1}^n \left[\sum_{j=1}^m \left[\frac{W_{i,j}}{W_{1,1}} \left(1 - 10^{-\frac{V_{THj} - V_{TH1}}{S}} \right) \left(\frac{V_i}{V_1} \right)^2 \right] \right] \quad (3)$$

where S is the subthreshold swing (typically given in units of mV/decade). The reduction in static power in low-Vdd devices is due to DIBL, the lower Vdd itself, and other complex device-related phenomena such as the relationship among doping, Vth, and S.¹ While our results use (3) to reflect the relationship between Ioff and Vdd, experiments using a linear (V_i/V₁) term showed only minor changes in the overall power reductions and optimal Vdd/Vth values. The degradation in speed when the power supply or Vth is changed can be estimated using the alpha-power law model [9]:

$$D_{i,j} = \left(\frac{V_j}{V_1} \right) \left(\frac{V_1 - V_{TH1}}{V_i - V_{THj}} \right)^\alpha \quad (4)$$

As shown in [5], the capacitance and transistor width along a path are largely proportional to the path's delay. Hence the ratios of widths in (3) can be replaced by ratios of capacitance. At this point the problem of power minimization for given voltages and thresholds can be formulated as a linear programming problem with the ratios of capacitances as the variables. We define a weight factor K as the ratio of

¹ For example, a long-channel device in a modern technology (large L to suppress DIBL) demonstrates a linear reduction in Ioff with Vds, contributing to the quadratic term in (3). This effect is not properly captured in traditional Ioff expressions [8].

dynamic to static power in the original single Vdd/Vth design (e.g. $K = 10$ implies that 10/11 of the total initial power was dynamic). Total power minimization is achieved by minimizing a weighted sum of the static and dynamic power. Hence the goal of total power reduction can now be expressed as

$$\begin{aligned} & \text{Maximize: } K \cdot \text{Gain}_{\text{dyn}} + \text{Gain}_{\text{static}} \\ & \text{s.t.: } \left(1 + \sum_{i,j} \frac{C_{i,j}}{C_{1,1}} (D_{i,j} - 1) \right) t \leq 1 \end{aligned}$$

where t is the original path delay. The constraint ensures the delay of each path is less than the critical delay of the network, which is normalized to one. Since paths are independent of each other, minimizing the power dissipation on each of the paths will lead to the minimum power of the complete logic network. Given the initial path delay distribution $p(t)$ of the network, the total power improvement can be found by summing over paths with different path delays.

The ratio K distinguishes between static power limited designs (portable) and dynamic (high-performance and non-mobile). For example, choosing $K = 1$ implies that reductions in static and dynamic power take on equal importance in an effort to minimize total power. In this work, we look at a range of K values from 1 to 50 with particular focus on $1 < K < 10$. Designs with $K < 1$ are likely to make heavy use of standby modes and other techniques to suppress leakage power which is beyond the scope of this work. A lambda-shaped $p(t)$, peaking at half of the critical delay, is assumed for all further analysis based on static timing analysis results shown in [5,10]. We also performed experiments on a flat path distribution and a sloped distribution where the maximum number of paths occur at the critical delay. Trends were consistent with expected results – the lambda shaped $p(t)$ gives the largest power savings while the sloped $p(t)$ enables about 2/3 the savings from the lambda case.²

III. COMPARISON OF DUAL VDD WITH DUAL VDD/VTH

The use of a second threshold voltage to both 1) reduce static power, and 2) provide speed improvements in logic gates that run at lowered supply voltages, is leveraged in a dual Vdd/Vth approach. In this section we contrast this with the use of two supply voltages and a single Vth. Both approaches to reducing power dissipation were applied to a design using a Vdd₁ of 0.9V and Vth₁ = 0.225V.

Results in Fig. 1 demonstrate that the power reduction obtained by applying dual Vdd/Vth is consistently much larger than the optimal dual Vdd design. The advantage offered by the second threshold voltage is smallest for lower K values (around 10-20%). This is because the dual Vdd/Vth technique is predicated on using a lower second threshold voltage to allow cells to be run at a lower power supply with good drive capability. At small K values, however, static power is comparable to dynamic power and an increase in static power due to the lower Vth is less acceptable as a trade-off. Modern high-performance designs exhibit K values in the range of 2-20; the dual Vdd/Vth approach delivers 15-30% lower power than dual Vdd alone over this range. This effect is also seen in Fig. 2 which shows the variation of the optimized second power supply voltage, Vdd₂.

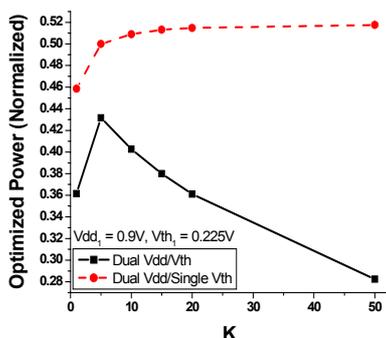


Figure 1. Dual Vdd/Vth shows 15-45% larger total power reduction than dual Vdd/single Vth throughout the range of K values.

² For example, a case demonstrating 60% power reduction for a lambda-shaped $p(t)$ shows about 40% power reduction for a sloped $p(t)$.

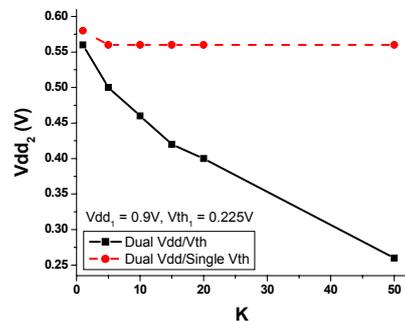


Figure 2. The presence of a second threshold voltage enables significantly more Vdd scaling, especially in dynamic power constrained applications (large K).

At higher K a much lower voltage can be used to achieve considerable dynamic power savings at the cost of static power, which constitutes a small fraction of the total power. Using a second power supply as low as 0.26V (for $K = 50$) provides approximately an 80% reduction in total power by using a very low threshold (in this case it is found to be 0.06V, referenced to Vdd₂). The rise in static power is approximately 5X under these conditions which is greatly outweighed by the dynamic power savings.

IV. SCALABILITY AND OPTIMAL VDD/VTH SELECTION

As shown in Fig. 1, the dual Vdd/Vth technique achieves considerably larger total power savings than dual Vdd alone. For comparison, [5] predicts a maximum (dynamic) power savings of 47% at Vdd₁ = 0.9V with an optimized Vdd₂ of 0.56V. Fig. 3 shows that dual Vdd/Vth designs can achieve power savings of 60% at Vdd₂ = 0.46V and $K = 10$. The value of Vth₁ has a strong impact on the characteristics of the optimized systems; a lower Vth₁ allows for more voltage scaling and power reduction (Fig. 3) but would also lead to a smaller K value. The smaller value of K will lessen the need for lower supply voltages and shift the focus to increasing the second Vth.

Fig. 3 also shows that the power improvements of dual Vdd/Vth designs increase as the initial power supply is scaled down, as opposed to previous results for dual Vdd where the improvements were projected to decrease with process scaling [5]. Results at Vdd₁ = 0.7V, anticipated for 65nm technologies, demonstrate even larger gains – 10% more power savings compared to Vdd₁ = 0.9V. The improvement of multi-voltage systems with scaling is due to the growing importance of Vth in determining delay in sub-1V technologies. Although the sensitivity of delay to Vth is rising at lower supply voltages, the dependency of leakage current with Vth is unchanged (neglecting major shifts in subthreshold swing). Thus, in future technologies Vth presents a more favorable Ion/Ioff tradeoff [4].

In contrast to dual Vdd, a dual Vdd/Vth approach is not strictly limited by the value to which the lower Vdd can be reduced, but by the value to which the second threshold can be decreased since that gives rise to an exponential increase in static power. This important distinction, combined with the above argument, makes dual Vdd/Vth inherently scalable. Fig. 4 shows that lower values of Vdd become optimal as K is increased but the effect saturates; this point is also reflected in the rules of thumb developed later. Fig. 5 supports the same conclusion from the standpoint of power savings, where small K values lead to very large reductions in static power, but for higher K static power is traded off to obtain dramatic savings in dynamic power.

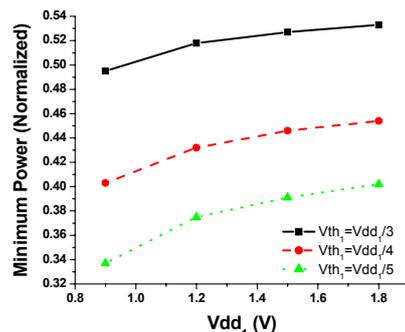


Figure 3. For a fixed K value ($K = 10$ here), a lower Vth₁ allows for more substantial power savings since Vdd can be scaled more aggressively.

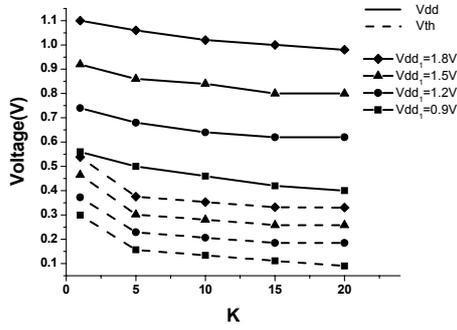


Figure 4. Trends of optimal second Vdd and Vth with varying K values. A sharp upward trend in Vth_2 for $K < 5$ is observed. $Vth_1 = Vdd_1/4$ for all Vdd_1 .

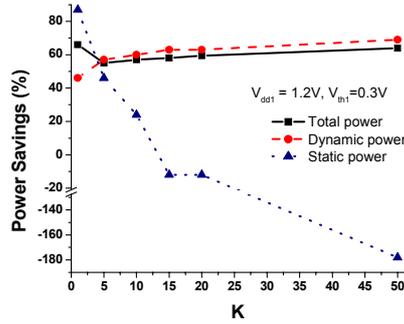


Figure 5. The breakdown of total power savings into static and dynamic components shows a large increase in static power (nearly 3X) at very high K values to achieve a more important dynamic power reduction.

Fig. 6 shows the variation of the minimum achievable power as a function of the second power and threshold voltage. As can be seen the optimal point is not overly sharp, and hence points close to optimal in terms of Vdd_2 and Vth_2 provide near-optimal power savings. This allows for the development of fairly simplified rules of thumb to estimate the optimal second supply and threshold voltages. Previously developed rules of thumb [5] are inapplicable to dual Vdd/Vth designs and to the minimization of total power. Rules of thumb are derived for the second Vdd and Vth as a function of the original voltages as well as the initial breakdown between dynamic and static power, K :

$$Vdd_2 = 0.43 Vdd_1 + 0.82 Vth_1 + \frac{0.72}{K} - \frac{0.55}{K^2} - 0.2 \quad (5)$$

$$Vth_2 = -0.024 Vdd_1 + 1.14 Vth_1 + \frac{0.72}{K} - \frac{0.49}{K^2} - 0.18 \quad (6)$$

These expressions provide excellent fit with an average error of 3.2% and 4.2% for (5) and (6) respectively. Both Vdd_2 and Vth_2 have strong dependencies on the first threshold voltage while Vth_2 only very weakly depends on Vdd_1 . The difference of Vdd_2 and Vth_2 (gate overdrive in $Vdd_2 + Vth_2$ cells, indicating speed) remains nearly unchanged for reasonable variations in K . Also, at higher K the models predict lower values for both Vdd_2 and Vth_2 mirroring the behavior shown in Fig. 4. Equations (5) and (6) were derived for nominal process characteristics ($\alpha = 1.3$, $S = 90\text{mV/dec}$, DIBL coefficient $\eta = 60\text{mV/V}$).

The final optimized static power dissipation varies from 5-35% of the total power for all cases investigated with the majority of likely cases falling in the 10-20% range (Fig. 7). This is smaller than previously estimated in [11] using analytical models and coincides with the highest-performance design points today [3]. There is some negative correlation between the final ratio of static and dynamic power and the initial K value as well as the initial Vdd, implying that more scaled designs with lower supply and threshold voltages will shift towards a more static-power dominated optimal point. The fraction of total optimized power due to leakage is consistent across the three $p(t)$ shapes we examined, indicating that the optimal static power ratio is not a function of the initial path delay distribution.

V. LEVEL CONVERTERS AND CRITICAL PATH DENSITY

One difficulty in implementing multi-Vdd designs is the need for level converters (LCs). Whenever a low Vdd cell fans out to a high Vdd cell, the voltage must be up-converted to avoid excessive leakage due to the

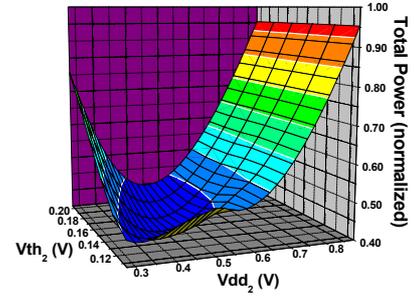


Figure 6. Power reduction as a function of second Vdd and Vth values. This example uses $Vdd_1 = 0.9\text{V}$, $Vth_1 = 0.225\text{V}$, and $K = 10$. The power is minimal at $Vdd_2 = 0.44\text{V}$ and $Vth_2 = 0.145\text{V}$

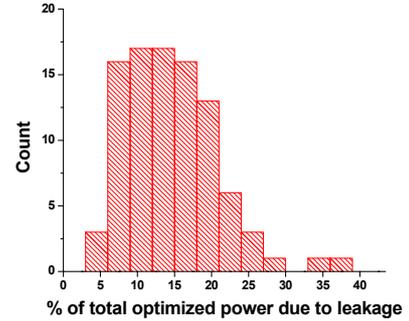


Figure 7. The fraction of power due to leakage in post-optimized designs is typically ~10-20% and is insensitive to initial path delay distribution.

PMOS being unable to fully turn off. There are two basic approaches to incorporating level conversion: 1) Clustered Voltage Scaling (CVS) which only allows level conversion at the flip-flops and 2) Extended-CVS (ECVS) where asynchronous level converters are used to allow any gate along a path to be assigned to low Vdd provided there is sufficient slack. ECVS does not place topological constraints on low Vdd assignment and can thus theoretically achieve larger power reductions. Since our approach assumes that any fraction of the capacitance can be set to low Vdd we inherently assume that ECVS is used.

To model the effects of delay penalties caused by level conversion, we first identify the paths where asynchronous or synchronous level converters are required. Fig. 8 shows the various combinations of supply and threshold voltages to which capacitances along a path are mapped. Fig. 8 shows the mapping of capacitance along each path relative to the initial path-delay distribution for a typical case at moderate K . Regions 1 and 2 are mapped completely to the lower Vdd, hence they only require synchronous up-conversion (and then only if any fan-out paths lie in regions 3 or 4). Paths with initial delay in region 4 (paths that were near critical originally) are all at high Vdd and do not require any up-conversion, even at flip-flops. The additional asynchronous LC penalty is then only associated with paths in region 3. As stated above, depending on the circuit topology some of the low Vdd paths may feed low Vdd paths only and would therefore not even require synchronous up-conversion. Since this effect cannot be considered in our analysis we assume all paths to have synchronous LC, making our analysis conservative. We do not consider the added power consumption of asynchronous LCs, which has been estimated to be 8% [6].

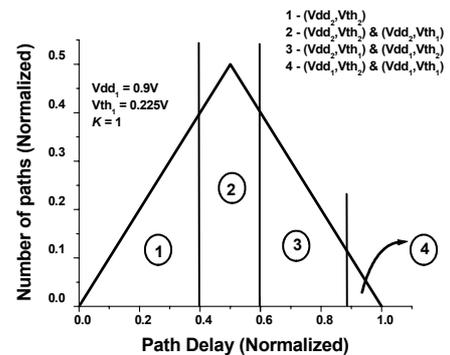


Figure 8. Depending on initial path speed, the capacitance along a path is mapped to either one or two (Vdd, Vth) combinations. Over 60% of paths for this typical case run at all low Vdd (Vdd_2).

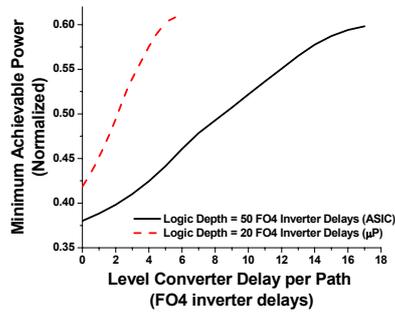


Figure 9. If numerous asynchronous level converters are required along a path, the potential power savings may be significantly reduced.

Fig. 9 shows the dependence of power savings on the delay due to asynchronous level conversion for two cases; first, a short critical path that is typical of high-performance microprocessors and second, a larger logic depth representing high-speed ASICs. Critical path delay is normalized to fan-out of four inverter delays (FO4), a common metric for the speed of a given technology, and is technology-independent [12]. Even when no asynchronous level conversion is required, synchronous level conversion incurs a fixed delay penalty (assumed to be 2 FO4 delays), resulting in a larger relative penalty for shallow logic depths. The LC delay penalty restricts the number of asynchronous conversions that can be performed on a path and hence our numbers in previous sections are upper bounds on the achievable power improvement. Fig. 9 shows the LC penalty to saturate when the conversion delay becomes a large fraction of the total path delay. The saturation is a result of the fact that paths in region 3 (as in Fig. 8) can no longer be mapped to low Vdd due to the level conversion overhead. Assuming that each asynchronous LC has a delay equal to 2.5 FO4 delays [13], the ASIC example in Fig. 9 can employ two level converters per path with a 15% rise in the achievable power compared to the case with no asynchronous LCs considered. In general, the asynchronous LC delay penalties are substantial for very shallow logic depths, but reasonable for high-performance ASICs.

Pushing semiconductor technology and processing equipment to their limits results in considerable uncertainty in key physical parameters and greatly complicates timing analysis. In this context, designs with a large fraction of paths operating at near critical path delay are more likely to produce timing failures after fabrication. As power reduction techniques such as multi-Vdd, dual-Vth, and sizing all result in more critical paths, we analyzed the relationship between the achieved power savings and critical path density for both traditional dual Vdd and dual Vdd/Vth. We define a critical path as one having a delay within 5% of the timing constraint. Fig. 10 shows that the dual Vdd/Vth technique yields 40% fewer critical paths for the same power reduction as dual Vdd, simplifying timing verification. Alternatively, at a given design complexity (quantified by the number of critical paths), dual Vdd/Vth provides an 11% power reduction compared to dual Vdd.

VI. MULTI VDD/VTH

In this section we briefly compare multi Vdd/Vth designs with dual Vdd/Vth designs. Table 1 compares the power savings provided by triple Vdd/dual Vth and dual Vdd/triple Vth to dual Vdd/Vth. The improvements provided by the additional threshold voltage are small. A triple Vdd/dual Vth approach is only 10% better than dual Vdd/Vth. These agree with [5] which showed saturated improvement with an in-

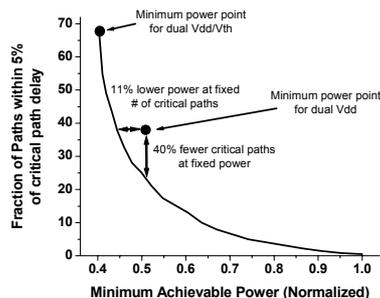


Figure 10. Dual Vdd/Vth provides a better power/criticality tradeoff than dual Vdd – for the same power savings, there are significantly fewer critical paths.

Table 1. Multiple Vdd/Vth techniques show saturating improvements over dual Vdd/Vth. Results are for $V_{dd1} = 0.9V$, $V_{th1} = 0.225V$, and $K = 10$.

Technique	Optimal Thresholds (V)	Optimal Voltage Supplies (V)	Minimum Power (normalized)
Dual Vdd Dual Vth	0.19	0.46	0.4
Dual Vdd Triple Vth	0.19, 0.15	0.4	0.38
Triple Vdd Dual Vth	0.18	0.52, 0.38	0.36

creasing number of power supplies and threshold voltages in multi Vdd or multi Vth designs. Triple Vth processes will have additional fabrication costs due to the extra implant steps and triple Vdd designs extend the place and routing difficulties of dual Vdd and also have higher level conversion penalties. Since the power savings compared to dual Vdd/Vth designs are limited, the overhead associated with these techniques appears to be prohibitive.

VII. CONCLUSIONS

This work addresses the simultaneous assignment of Vdd and Vth in multi-voltage systems to minimize total power consumption, considering DIBL. Our results show that the total power reduction achievable in modern and future integrated circuits is approximately 60-65% using a dual Vdd/Vth approach. We derive rules of thumb to guide selection of optimal second Vth and Vdd as a function of initial voltages as well as a key weighting factor, K , to drive the optimization towards static or dynamic power reduction. An important finding is that the optimal second Vdd in multi-Vth systems is $\sim 50\%$ of the higher supply voltage which is contrasted with $0.6-0.7 \cdot V_{dd1}$ for single Vth designs as previously found. The total power using dual Vdd/Vth can be reduced by 15-45% compared to dual Vdd. Since most high-performance designs today already use dual-Vth to limit standby power, there is no additional cost for this savings.

The inclusion of level conversion delay penalties demonstrates the tradeoff between allocating available slack to level conversion and achievable power reductions. Typically, 1-2 level conversions per path are tolerable in designs with larger logic depths (30+ FO4 delays) with $< 15\%$ power penalty. Finally, we highlight the relationship between power savings and critical path count – this is important since a rapidly increasing number of critical paths combined with rising process variability increases design times and emphasizes a need for improved statistical timing analysis tools.

REFERENCES

- [1] K. Usami and M. Horowitz, "Clustered voltage scaling technique for low-power design," *ISLPD*, pp. 3-8, 1995.
- [2] N. Rohrer, et al., "A 480MHz RISC microprocessor in a 0.12um Leff CMOS technology with copper interconnects," *ISSCC*, pp. 240-241, 1998.
- [3] S. Tang, et al., "A leakage-tolerant dynamic register file using leakage bypass with stack forcing (LBSF) and source follower NMOS (SFN) techniques," *Symp. on VLSI Circuits*, pp. 320-321, 2002.
- [4] D. Sylvester and H. Kaul, "Future performance challenges in nanometer design," *DAC*, pp. 3-8, 2001.
- [5] M. Hamada, Y. Ootaguro, and T. Kuroda, "Utilizing surplus timing for power reduction," *CICC*, pp. 89-92, 2001.
- [6] K. Usami, et al., "Automated low-power technique exploiting multiple supply voltage applied to a media processor," *IEEE J. Solid-State Circuits*, pp. 463-472, Mar. 1998.
- [7] K. Takeuchi and T. Mogami, "A new multiple transistor parameter design methodology for high speed low power SOCs," *IEDM*, pp. 22.6.1-22.6.4, 2001.
- [8] A. Chandrakasan, W.J. Bowhill, and F. Fox, *Design of High-Performance Microprocessor Circuits*, IEEE Press, 2001.
- [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*, pp. 584-593, Apr. 1990.
- [10] J.D. Warnock, et al., "The circuit and physical design of the POWER4 microprocessor," *IBM J. of R&D*, Vol.46, pp.27-52, Jan. 2002.
- [11] K. Nose and T. Sakurai, "Optimization of Vdd and Vth for low-power and high-speed applications," *ASP-DAC*, pp. 469-474, 2000.
- [12] R. Ho, K.W. Mai, and M. Horowitz, "The future of wires," *Proc. of the IEEE*, Vol.89, pp. 490-504, Apr. 2001.
- [13] Sarvesh Kulkarni, personal communication.