Combining Dual-Supply, Dual-Threshold and Transistor Sizing for Power Reduction

Stephanie Augsburger[1], Borivoje Nikolić[2]
[1]Intel Corporation, Enterprise Processors Division, Santa Clara, CA, USA.
[2]Department of EECS, University of California, Berkeley, CA, USA
stephanie.a.augsburger@intel.com

Abstract

Multiple supply voltages, multiple transistor thresholds and transistor sizing could be used to reduce the power dissipation of digital blocks. This paper presents a framework for evaluating the effectiveness of each of these approaches independently and in conjunction with each other. Results show the advantages of multiple supply, transistor sizing, and multiple threshold can be compounded to maximize power reduction. The order of application of these techniques determines the final savings in active and leakage power.

1. Introduction

This work considers the combined effectiveness of multiple threshold voltages, multiple supply voltages and transistor sizing towards both active and leakage power reduction. The analysis includes design directives for achieving the optimal balance of power reduction techniques.

Design techniques for low power consumption in modern VLSI are becoming increasingly important [1], [2]. As technology moves into deep submicron feature sizes, the designs are becoming essentially power limited. Power dissipation due to leakage current is increasing at an exponential rate. Projections show that leakage power will become comparable to dynamic power dissipation in the next few years [3]. Supply voltage has not been scaled aggressively enough to keep power per unit area constant over technology generations [4].

Power consumption in CMOS circuits can be categorized into a number of major components. The dominant source of power consumption is the dynamic switching power needed to drive the capacitive loads on gates. Active power is dissipated when a short circuit current occurs during a switching event. Other components include static power consumption, which is generally zero for most logic families, and leakage power. A general formula for power consumption is shown in (1) [5]. The parameter, α, is the switching probability of a gate.

$$P \sim \alpha \cdot \left(C_L \cdot V_{swing} + \overline{I_{SC}} \cdot \Delta t_{SC}\right) \cdot V_{DD} \cdot f + \left(I_{DC} + I_{Leak}\right) \cdot V_{DD} \quad (1)$$

Leakage power is attributed to leakage current, which is related to threshold voltage as follows [5]:

$$I_{LEAK} = \frac{I_0}{W_0} W \cdot 10^{-\frac{V_T}{S}}. \quad (2)$$

W is the channel width, and S is the subthreshold slope. The typical value of S is 0.1V/decade, which reflects an order of magnitude increase in leakage current with a 0.1V drop in threshold voltage.

In recent years, a number of power reduction techniques have emerged. These focused primarily on the individual effects of multiple-supply, multiple-threshold and gate sizing techniques. In [6], the basics of dual-supply design are introduced. A 10% to 20% power reduction was reported with a clustered voltage scaling (CVS) dual-V_{DD} design. Dual-supply design methodology, including layout issues, is covered in [7]. Reference [8] presents the use of multiple-threshold assignment on a cell-by-cell basis and reports leakage reduction from 75% to 90%. A triple-threshold RISC processor was showcased in [9]. The use of transistor sizing for power reduction is a common technique and has been covered thoroughly in years past [10]. In [11], multiple-supply, multiple-threshold and transistor sizing were looked at individually from a theoretical standpoint. Rules of thumb for optimal supply voltages, threshold voltages and transistor sizes were derived from a series of equations.

Each of these techniques trades off the same timing slack for potential energy savings. Individually, the most effective technique would achieve the largest savings with the same slack. Little work has been done in combining multiple-supply, multiple-threshold and transistor sizing techniques in order to compound power savings. Here, we examine these techniques when used in conjunction.

Proceedings of the 2002 IEEE International Conference on Computer Design: VLSI in Computers and Processors (ICCD'02)
1063-6404/02 $17.00 © 2002 IEEE

This work is restricted to a digital block with a fixed throughput and constant latency. For a given delay, power and energy are minimized through selective adjustments to threshold and supply voltages, as well as transistor sizing. The target application for these results is general ASIC design, which is characterized by a limited number of paths that constitute the critical delay. The observation of cycle slack in the non-critical paths permits the introduction of these power reduction techniques without affecting the overall system throughput.

Presently, to the best of our knowledge, no gate-level CAD tool support exists for joint evaluation of all of these three techniques. In order to facilitate this exploration, a design framework was constructed using MS Excel software and Spectre simulations. Models of basic gates were derived through simulation and a generic path-delay distribution was generated based on these gates. Inside this environment, a number of combinations of the three power reduction techniques were evaluated.

Section 2 describes the test setup, while Section 3 details the results of the various attempted methods. These results are analyzed in Sections 4, and conclusions are presented in Section 5.

2. Test Setup

2.1. Delay and Energy Models

This work is based on a linear delay model, where the gate delay is expressed as a linear function of the load capacitance. We ignore the delay dependence on the input slope in this early evaluation. A simplified library is based on logic gates, each of them designed with multiple sizes. Using Cadence™ Composer and Spectre, these gates were simulated using a general-purpose 0.13µm CMOS process. The technology provides two threshold voltages, high speed (HS), or low V_T, and low-leakage (LL), or high V_T, with a spread of approximately 100mV. The baseline supply voltage was 1.2V. The value of 0.8V for the second supply, V_{DDL} was chosen in accordance with [11]. Each gate/supply/threshold combination was simulated to determine active energy, leakage power and delay for several different load capacitances. To complete the models, delay and active energy were plotted against the load capacitance and then linear extrapolation was used to determine the slope and y-intercept values for both active energy and delay. An example of the linear model calculation is shown for a 2-input NAND gate in Figure 1.

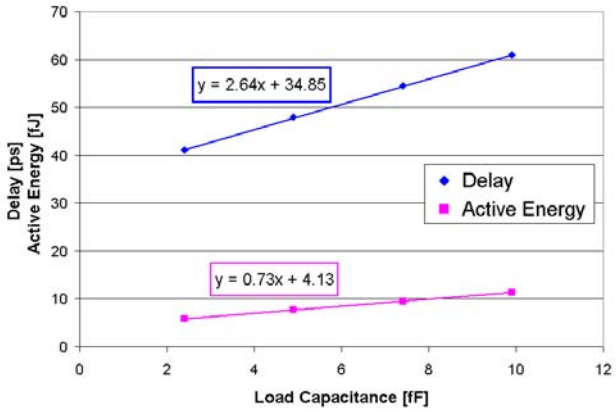

Figure 1. Linear delay and active-energy models for 2-input NAND gate with 0.8V V_{DD}, HS V_T

2.2. Level-converting Flip-Flops

In multiple-V_{DD} designs, level-converters are required to provide proper interfacing when connecting a gate with a lower supply to one with a higher supply. In a CVS approach, the level conversion function is embedded in a flip-flop and all V_{DDL} cells are clustered at the ends of each path [7]. This CVS method of multiple-supply is adopted here.

Conventional level-converting flip-flop (LCFF) [6] are standard master-slave latch pairs that rely on positive feedback of the cross-coupled inverter pair in the slave stage to restore logic swing. As a result an increased clock-to-output delay causes longer delay of the succeeding pipeline stage. In [12], LCFFs were designed that reduce flip-flop delay from the data input to output. The design most suited to this project, a pulsed latch with level conversion performed in a half-latch, was chosen as an alternative to the traditional LCFF. . This design effectively eliminates the delay penalty at $V_{DDL} = 1.0$V and drastically reduces it at $V_{DDL} = 0.8$V. The setup time is slightly increased, while clock-output delay is virtually unchanged. One of the drawbacks of using pulsed latches in general purpose logic is their long hold time. However, in dual supply designs, race conditions can be managed by lowering the supply voltage on short logic paths. The flip-flop timing and energy characteristics are shown in Table 1. The downside of this design is slightly increased power consumption. This power penalty is offset in a dual-supply design by the increase in the number of gates that can be placed in V_{DDL} with this LCFF.

Table 1. LCFF normalized delay and energy

Design	Normalized Delay	Normalized Energy
Baseline (FF)	1.0	1.0
LCFF @ 0.8V	1.25	2.80
LCFF @ 1.0V	0.94	1.29

2.3. Design Framework and Baseline Design

Due to a lack of commercial CAD tool support, a structure was built in a MS Excel workbook in which individual gates can be combined to form paths. The delay, active energy, and leakage power for each path is calculated through the use of the gate models and the sizing of the gates. A baseline design, with lambda-shaped distribution, was formed by randomly stringing together different combinations of gates, up to 12 gates per path, to form 500 paths. Initially, all gates were minimum size. Sizing was then used to bring each path to its maximum speed to approximate a synthesized design. Loading for each path was set to 50x the input capacitance of a flip-flop, or 175fF. This capacitance was chosen to approximate driving a local bus. Figure 2 shows the effect of creating the baseline design through sizing on the path-delay distribution of the initial unsized design. Active energy per transition, leakage power and maximum delay numbers for the two designs are listed in Table 2.

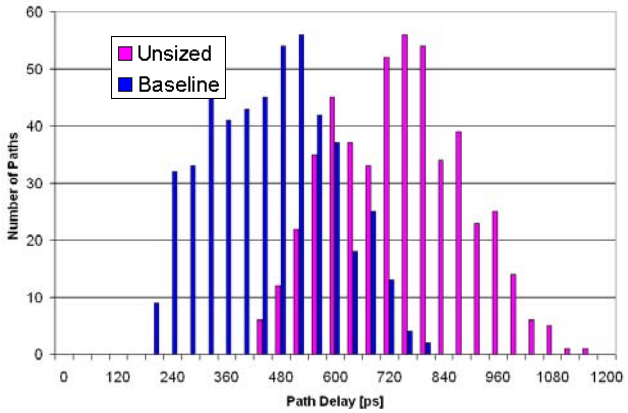

Figure 2. Baseline and initial unsized design path-delay distributions

Table 2. Summary of baseline and unsized designs

Design	Active Energy [pJ]	Leakage Power [µW]	Maximum Delay [ps]
Initial unsized	214.9	45.86	1142.6
Baseline	293.3	76.80	721.2

3. Results

Results are presented for downsizing, dual-threshold and dual-supply applied independently to the initial design. Then, these techniques are combined to explore compounded effectiveness. Each method slows down non-critical paths for energy savings through lowering the supply voltage, increasing the threshold or downsizing the gates.

3.1. Sizing

Gates off the critical paths were downsized where possible. Figure 3 details the effect of sizing on the path-delay distribution, while Table 3 shows the normalized active energy and leakage power for the downsized design against the baseline design.

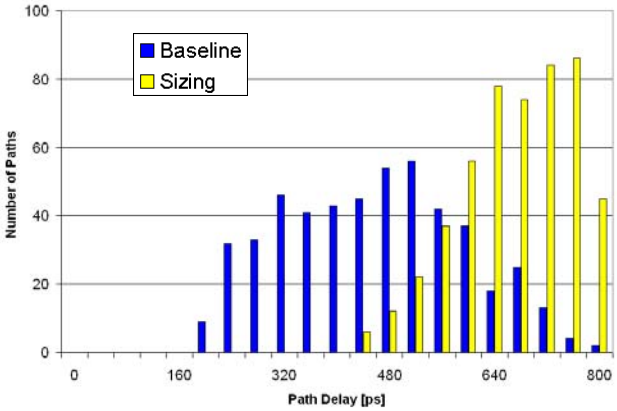

Figure 3. Effect of sizing on path-delay distribution

Table 3. Power/energy with transistor sizing

Design	Normalized Active Energy	Normalized Leakage Power
Baseline	1.00	1.00
Sizing	0.75	0.62

3.2. Dual Threshold Voltages

The dual-threshold technique was applied to both the baseline design and the sized design. Cell assignment (to either LL or HS) was done on a cell-by-cell basis, with the goal being maximum reduction of leakage energy for each path while still meeting timing goals. By replacing high-speed cells with low leakage cells, leakage power was reduced substantially in both the baseline design and the sized design. See Figure 4 for the path-delay distribution, and Table 4 for the normalized power/energy numbers. Leakage power was reduced more substantially

when dual-threshold was applied to the baseline design as opposed to when it was applied to the sized design. The small reduction in active energy with the dual-threshold design is primarily due to reduced gate channel capacitances in the off state and a small reduction in signal swings ($V_{DD} - V_T$) at intermediate nodes. Speed penalty ranges from 25 – 50% for high-V_T cells.

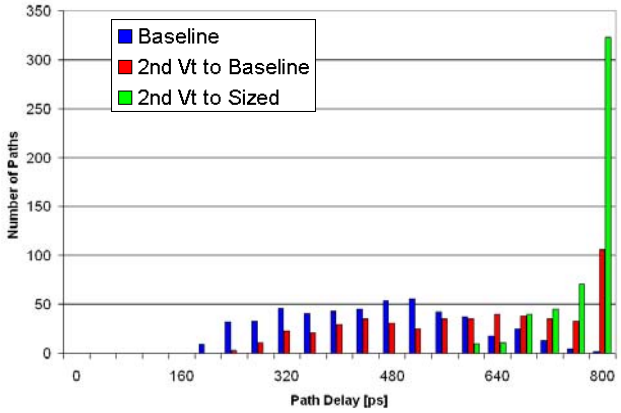

Figure 4. Path-delay distribution with dual-V_T

Table 4. Power/energy with dual-V_T

Design	Normalized Active Energy	Normalized Leakage Power
Baseline	1.00	1.00
Dual-V_T	0.97	0.12
Sizing + dual-V_T	0.74	0.31

3.3. Dual-Supply

The second supply voltage was selected as 0.8V using the rule-of-thumb in [11].

The CVS method [6] was used to determine the cluster of V_{DDL} cells. All V_{DDL} cells were grouped at the end of the path. The dual-supply technique was applied to both the baseline design and the sized design.

The path-delay distributions for the dual-supply design (V_{DDL} = 0.8V) are shown in Figure 5, and the normalized power/energy in Table 5. It can be concluded that using dual-supply is more effective than sizing for both active and leakage power reduction.

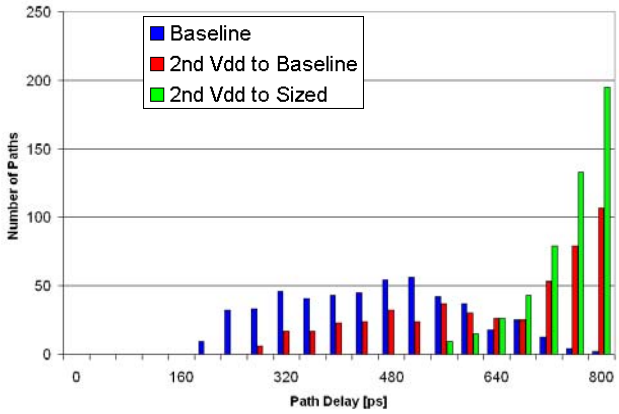

Figure 5. Path-delay distribution with dual-V_{DD}, V_{DDL} = 0.8V

Table 5. Power/energy with dual-V_{DD}

Design	Normalized Active Energy	Normalized Leakage Power
Baseline	1.00	1.00
Dual-V_{DD}	0.66	0.33
Sizing + dual-V_{DD}	0.65	0.52

3.4. Combination of Techniques

Using dual supplies is the most effective technique for active power reduction. However, due to large impact on delay of lowered supply voltage and limitations of CVS method, a part of the timing slack remains unused. It would be beneficial to apply the other two techniques on the paths with remaining slack for added energy savings.

To determine if the benefits of dual-supply, dual-threshold and sizing are cumulative, they were applied in conjunction with each other.

Sizing was added to the dual-V_{DD} design of Section 3.3 (dual-V_{DD} applied to baseline). See Figure 6 and Table 6 for results. Comparing these results with those where the baseline was sized and *then* dual-supply was applied confirms that dual-supply is more effective than sizing for both active energy and leakage power. A second V_T was then applied to this design, which resulted in a small additional decrease in leakage power. The effect of dual-threshold was lessened because many more paths were critical after applying dual-supply and sizing and could not absorb the large delay penalty of high-V_T cells.

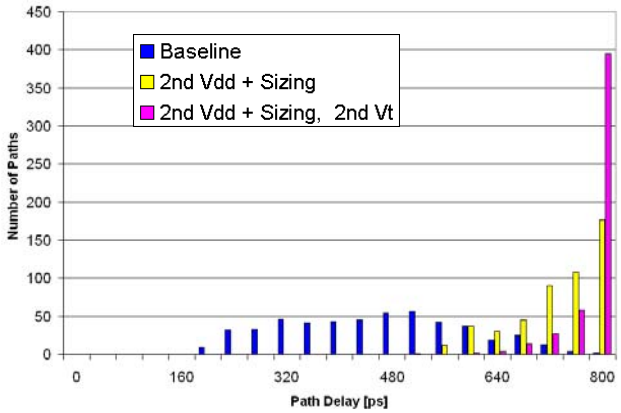

Figure 6. Path-delay distribution, combining dual-V_{DD} with other techniques, V_{DDL} = 0.8V

The second supply voltage was added to the dual-threshold design of Section 3.2 (dual-V_T applied to baseline). Sizing was then used to further reduce power consumption (see Figure 7 and Table 6).

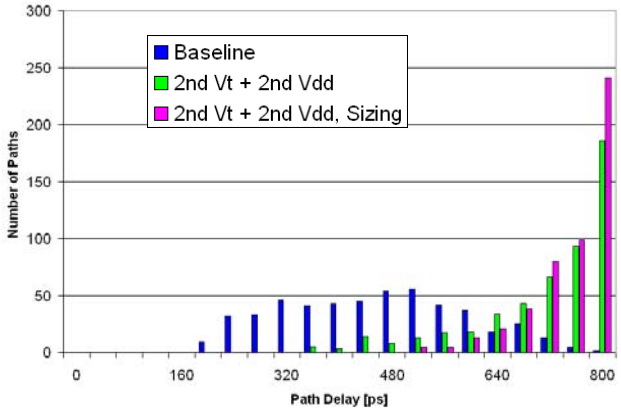

Figure 7. Path-delay distribution, combining dual-V_T with other techniques, V_{DDL} = 0.8V

Table 6. Power/Energy with combination of techniques

Design	Normalized Active Energy	Normalized Leakage Power
Baseline	1.00	1.00
Dual-V_{DD} + sizing	0.55	0.25
Dual-V_{DD} + sizing, dual-V_T	0.55	0.23
Dual-V_T + dual-V_{DD}	0.81	0.12
Dual-V_T + dual-V_{DD}, sizing	0.69	0.10

4. Analysis of Results

There exists an optimal energy for a given block. However, in practice, the techniques of dual-supply, dual-threshold and sizing would typically be applied sequentially. Depending on the order of application, different results are obtained for leakage and active power.

Overall power is dependent on many factors, including switching activity and the process. Depending on the activity of a logic block, a different emphasis should be placed on the techniques used. For high activity, dual-supply should be the first technique applied, followed by transistor sizing and then dual-threshold, only if it does not impact the active power (this will depend on the value of V_{DDL}). However, if leakage power is the chief concern (low activity), dual-threshold takes precedence over the other techniques, followed by dual-supply and then transistor sizing.

In this analysis, the starting point was a logic block with all paths sized for maximum speed with all low-V_T transistors. This presents an over-design, but this is common in today's designs.

Relative power savings are observed to depend on the ratio of the internal capacitance inside the block and the loading capacitance.. Also, absolute values of power savings depend on the type of logic block, as well as its loading. It should be noted that sizing cannot affect energy consumption on the load, while the second supply essentially starts from there. Therefore, in a block without substantial loading, dual-supply may not be superior to downsizing.

5. Conclusions and Future Work

The effects of dual-supply, dual-threshold and transistor sizing were evaluated on a typical logic block in order to gain a consistent design methodology for how and when each of these three common power reduction techniques should be used. The completed experimentation shows that energy savings from these three base techniques can be compounded through proper combination for additional benefit. The leftover slack after the introduction of the second supply, as the most effective active energy reduction technique, can be consumed by downsizing, for additional savings. The second threshold voltage should be used either as a first technique for low activity blocks, or the last technique to consume leftover slack of high-activity blocks.

The large potential power savings shown in this study should motivate EDA support for design environments that combine these techniques, particularly in the area of multiple-supply voltages.

6. Acknowledgment

This work was completed while S. Augsburger was a graduate student at the University of California, Berkeley, where it was supported in part by the MARCO/DARPA Gigascale Silicon Research Center (http://www.gigascale.org). S. Augsburger was supported by an SRC Masters Scholarship. Their support is gratefully acknowledged.

7. References

[1] J. D. Meindl, "Low power Microelectronics: Retrospect and Prosprect," *Proceedings of the IEEE*, Vol. 83, No. 4, pp.619, 1995.

[2] A. Chandrakasan, S. Sheng and R. Brodersen, "Low-Power CMOS Digital Design," *IEEE Journal of Solid-State Circuits*, vol. 27, no. 4, pp 473-484, April 1992.

[3] V. De *et al*., "Techniques for Leakage Power Reduction," in *Design of High-Performance Microprocessor Circuits*, IEEE Press, NJ, 2001, pp46-62.

[4] J. Edmondson, "Impact of Physical Technology on Architecture," in *Design of High-Performance Microprocessor Circuits*, IEEE Press, NJ, 2001, pp3-24.

[5] T. Kuroda, "Low-Power CMOS Circuit Design by Means of Supply-Voltage and Threshold-Voltage Control," Ph.D. Dissertation, University of Tokyo, December 1998.

[6] K. Usami and M. Horowitz, "Clustered Voltage Scaling for Low-Power Design," *International Symposium on Low Power Design*, pp 3-8, April 1995.

[7] K. Usami and M. Igarashi, "Low-Power Design Methodology and Applications Utilizing Dual Supply Voltages," *Proceedings of the Asia and South Pacific Design Automation Conference 2000*, pp 123-128, January 2000.

[8] N. Kato *et al*, "Random Modulation: Multi-Threshold-Voltage Design Methodology in Sub-2V Power Supply CMOS," *IEICE Transactions on Electronics*, vol. E83-C, no.11, pp 1747-1754, November 2000.

[9] T. Yamashita *et al*, "A 450MHz 64b RISC Processor Using Multiple Threshold Voltage CMOS," *2000 IEEE International Solid-State Circuits Conference Digest of Technical Papers*, pp 414-415, February 2000.

[10] J. Rabaey, *Digital Integrated Circuits: A Design Perspective*, Prentice Hall, NJ, 1996.

[11] M. Hamada, Y. Ootaguro and T. Kuroda, "Utilizing Surplus Timing for Power Reduction," *Proceedings of the IEEE 2001 Custom Integrated Circuits Conference*, pp 89-92, May 2001.

[12] F. Ishihara and B. Nikolic, "Level-Converting Flip-Flops for Dual-Supply Systems," to be published, 2002.